**IIIA-CSIC Technical Report 2003-??**
**May 2003**

# Alert Triage on the ROC

by

Francisco J. Martin[1]
School of Electrical Engineering and Computer Science
Oregon State University
Corvallis, 97331 OR, USA
Vox: +1-(541)-737-5515, Fax: +1-(541)-737-3014
fmartin@cs.orst.edu

Enric Plaza
IIIA - Artificial Intelligence Research Institute
CSIC - Spanish Council for Scientific Research
Campus UAB, 08193 Bellaterra, Catalonia, Spain
Vox: +34-93-580 95 70, Fax: +34-93-580 96 61
enric@iiia.csic.es

[1]On sabbatical leave from iSOCO-Intelligent Software Components S.A

---

This report is an extended version of the paper published in MMM-ACNS-2003: The Second International Workshop on Mathematical Methods, Models, and Architectures for Computer Networks Security

## ABSTRACT

This work: i) overviews ROC analysis and its use in the evaluation of intrusion detection systems, ii) compares ROC analysis with recent alternatives, and iii) proposes a formal framework based on ROC analysis for the evaluation of alert triage in intrusion detection.

**Keywords: Intrusion Detection, Evaluation of Detection Systems, ROC Analysis, Alert Triage**

# Contents

**1 Introduction** **2**

**2 ROC Analysis** **3**
2.1 Non-parametric Detection Systems . . . . . 3
2.2 Parametric Detection Systems . . . . . 5
2.2.1 ROC Space . . . . . 5
2.2.2 ROC Point . . . . . 6
2.2.3 ROC Curve . . . . . 9
2.2.4 Slope of the Curve . . . . . 9
2.2.5 Area Under the ROC Curve . . . . . 9

**3 ROC Analysis in Intrusion Detection** **10**

**4 ROC Alternatives** **10**
4.1 Detection Error Tradeoff . . . . . 10
4.2 Explicit Representation of Expected Costs . . . . . 11

**5 A Formal Framework for the Evaluation of Alert Triage** **12**
5.1 Ideal Alert Triage Environment . . . . . 14
5.2 Cost-based Alert Triage Evaluation . . . . . 16
5.3 Alert Triage Evaluation in Imprecise Environments . . . . . 17

**6 Conclusions** **19**

# 1 Introduction

The evaluation of intrusion detection systems (IDSes) has gained attention over the last years [DM02, DCW99, GU01, LFG00, McH00]. First, the 1998 and 1999 DARPA evaluations conducted by the Massachusetts Institute of Technology (MIT) Lincoln Laboratory [LFG00, HLF01] and then Mchugh's critique [McH00] of such experiments have shown that this area needs much more research and experimentation before a framework for the evaluation of IDSes can be widely accepted. Moreover, in the industry, IDS benchmarks provide misleading results rather than useful independent evaluations what accentuates this need [DM02, Ran01]. In this work we pursue a less ambitious objective. We intend at providing a framework for the evaluation of *alert triage* —just a cog in the intrusion detection machinery. Alert triage is the process of rapid and approximate priorization for subsequent action of an IDS alert stream [MP01]. Alert triage is the main task to be performed by IDS modules that have been named differently in the literature such as alert correlators [NC02], d-boxes [SS98], analyzers [WE02], ACCs [DW01], assistants [MP03b], etc. Indepently of the name, the correct interpretation of an IDS alert stream constitutes an active area of research in the intrusion detection community [NC02, SS98, WE02, DW01, MP03b, MP03a, CM02, GHH01, MMD02, NCR02, PFV02]. However, to the best of our knowledge nobody has provided a method of analysis for the evaluation of this particular component of IDSes.

Alert triage can be formulated as a detection or classification task[1] as we will see in Sec. 5. ROC analysis is a key evaluation technique for comparing the performance of detection and classification systems [Swe96] and lays the groundwork of this paper. ROC analysis was originally introduced in the field of signal detection theory in the 50's [Ega75]. The term Receiver Operating Characteristic refers to the performance (the operating characteristic) of a human or mechanical observer (the receiver) that has to discriminate between radio signals contaminated by noise (such as radar images) and noise alone [Ega75]. Nowadays, ROC analysis has become a common tool in medical decision making to qualify diagnostic tests specially in radiology [MSL76, HHJ97, HM82, HH02] and in psychology for sensory and cognitive processes [SDM00]. Recently, ROC analysis has also been proposed to compare and improve machine learning classifiers [PFK98, PF01, FFH02, WF02] and information retrieval systems [MP01].

This report proceeds as follows. First, Section 2 overviews ROC analysis and exposes some well-known concepts in the context of detection systems. Section 3 considers the use of ROC analysis in the evaluation of IDSes. Section 4 examines recent alternatives to ROC analysis. Section 5 introduces our formal framework for evaluating alert triage in intrusion detection. Finally, Sec. 6 presents some concluding remarks.

[1] Respectively known is statistics as binary and n-ary hypotheses testing.

# 2 ROC Analysis

Detection is the judgement on the presence of some condition in the present or the occurrence of an event in the future (e.g. an intrusion). A detection system, a person or a device or a combination of both, makes (positive or negative) decisions regarding the presence or absence of some condition [SDM00]. Depending on the output, detection systems can be classified into non-parametric and parametric detection systems.

## 2.1 Non-parametric Detection Systems

When the output of a detection system only considers yes-no alternatives then it has the four possible decision outcomes shown in Table 1.

Table 1: Four possible outcomes of a non-parametric detection system.

| | Condition | |
|---|---|---|
| Decision | Present | Absent |
| Positive | True Positives | False Positives |
| Negative | False Negatives | True Negatives |

On the one hand, a detection system makes two types of correct decisions: true positives (TP) and true negatives (TN). That is, the judgement on the presence and respectively absence of the condition of interest is correct. On the other hand, a detection system can commit two kinds of decision errors: false positives (FP) and false negatives (FN). Respectively, deciding that some condition is present when it is really absent (e.g. an alert is signaled when there is a manifest absence of intrusive behavior), and, conversely, deciding that some condition is absent when it is really present (e.g. an attack is undergoing and no alert is evoked).

The evaluation of a non-parametric detection system can be performed using a *confusion matrix*. A confusion matrix is a two-by-two table that collects the frequency of the four possible decision outcomes for the evaluation of a detection system that made $N$ decisions ($N = TP + TN + FP + FN$). The following fractions and probabilities are of interest.

The *true positive fraction* (TPF) or $P(D+ \mid C+)$ of a detection system is the probability that it makes a positive decision (D+) when the condition is present (C+). It is also called *sensivity* or *recall*.

$$TPF \approx \frac{TP}{TP + FN} \tag{1}$$

The *true negative fraction* (TNF) or $P(D- \mid C-)$ of a detection system is the probability that it makes a negative decision (D-) when the condition is absent (C-). It is also called *specificity*.

$$TNF \approx \frac{TN}{FP + TN} \tag{2}$$

The *false positive fraction* (FPF) or $P(D+ \mid C-)$ of a detection system is the probability that it makes a positive decision (D+) when the condition is absent (C-).

$$FPF = 1 - TNF \tag{3}$$

The *false negative fraction* (FNF) or $P(D- \mid C+)$ of a detection system is the probability that it makes a negative decision (D-) when the condition is present (C+).

$$FNF = 1 - TPF \tag{4}$$

Notice that the TPF answers the question: if the condition is present, how likely will the detection system make a positive decision? The TNF answers the question: if the condition is absent, how likely will the detection system make a negative decision? Nevertheless, often the questions of interest are: if the decision was positive, how likely is the condition to be present? and if the decision was negative, how likely is the condition to be absent?. These questions are answered using the following concepts.

The *positive predictive value* (PPV) or $P(C+ \mid D+)$ of a detection system is the probability that the condition is present (C+) given that the decision was positive (D+). It is also called *precision*.

$$PPV \approx \frac{TP}{TP + FP} \tag{5}$$

The *negative predictive value* (NPV) or $P(C- \mid D-)$ of a detection system is the probability that the condition is absent (C-) given that the decision was negative (D-).

$$NPV \approx \frac{TN}{FN + TN} \tag{6}$$

The *prevalence* $P(C+)$ of a detection system is the probability that the condition is present. See [Axe00] for a discussion on base-rates.

$$prevalence \approx \frac{TP + FN}{N} \tag{7}$$

The *accuracy* of a detection system is the percentage of correct decisions that it makes, i.e the percentage of true positives plus the percentage of true negatives.

$$accuracy \approx \frac{TP + TN}{N} \tag{8}$$

Frequently, accuracy has been used as the most basic form of measuring the performance of a detection system when misdetection costs are not contemplated. However, even so it is not a good measure of the performance of a detection system. Imagine that we evaluate a detection system using 997 innocuous alerts and only three that correspond to malicious

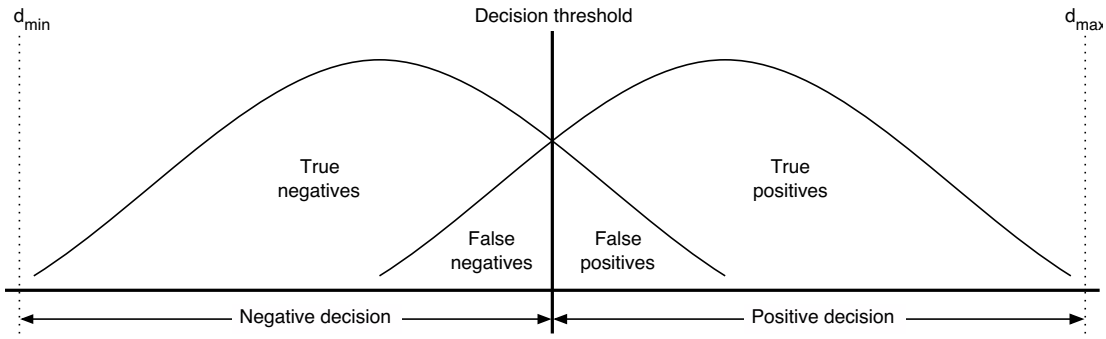



Figure 1: A decision threshold establishes a balance between both types of error.

attacks. Then a simple detection system that always makes negative decisions will have 0.997 accuracy even when it has missed all the malicious alerts. Moreover, that level of accuracy will surely be hard to achieve by a more sophisticated detection system [PF01]. We will show later that the performance of a detection system should be measured in terms of not only the number of decisions that it makes correctly but also in terms of the cost of the errors that it commits.

## 2.2 Parametric Detection Systems

Often, some detection systems are designed to return a value on a continuous measure — i.e. indicating the likelihood of the presence of the condition being inspected— instead of a yes or not. In these detection systems, called parametric detection systems, an arbitrary *decision threshold* (cut-off point) has to be chosen such that above the decision threshold the decision about the condition is positive and below it the decision is negative. Figure 1 uses two Gaussian curves to illustrate the workings of a decision threshold. The Gaussian on the left represents the population where the condition is absent (e.g. alerts corresponding to innocuous attacks) whereas the Gaussian on the right represents the population where the condition is present (e.g. alerts corresponding to malicious attacks). The decision threshold establishes a tradeoff between the two types of errors. In other words, if we move the decision threshold to the left on the abcissa axis then the number of false negatives will be decreased but the number of false positives will be increased. On the contrary, if we move the decision threshold on the opposite direction both errors will be altered conversely. So, how can we choose an optimal decision threshold?

ROC analysis allows one to select the optimal decision threshold of a parametric detection system. ROC analysis is best explained in terms of its components.

### 2.2.1 ROC Space

ROC space is a two-dimensional Cartesian space that lies in a unit square where the ordinate axis designates the true positive fraction and the abcissa axis designates the false positive

fraction (see Fig. 2). It has four singular points:

1. $(\mathbf{0},\mathbf{0})$ This point symbolizes a detection system that never makes positive decisions and corresponds to a parametric detection system whose decision threshold has been set to the maximum value $d_{max}$.

2. $(\mathbf{0},\mathbf{1})$ This point represents the perfect detection system. It always makes correct decisions and therefore the FPF is 0 and the TPF is 1.

3. $(\mathbf{1},\mathbf{0})$ This point denotes the worst detection system. FPF is 1 and the TPF is 0 since it always makes wrong decisions.

4. $(\mathbf{1},\mathbf{1})$ This points symbolizes a detection system that always makes positive decisions and corresponds to a parametric detection system whose decision threshold has been set to the minimum value $d_{min}$.

A diagonal line from the point $(0,0)$ to the point $(1,1)$ (the dotted line from the lower left hand corner to the upper right hand corner in Fig. 2) represents a detection system with no discriminating power (i.e a detection system that works no better than chance).

### 2.2.2 ROC Point

Given a decision threshold $d$ a ROC point (operating characteristic) is a pair of values $(FPF_d, TPF_d)$ such that $FPF_d$ represents the false positive fraction and $TPF_d$ represents the true positive fraction for threshold $d$. We say that one point in a ROC curve dominates another if its above and to the left. That is, it has a higher or equal TPF and lower or equal FPF.

$$(FPF_{d_1}, TPF_{d_1}) \gg (FPF_{d_2}, TPF_{d_2}) \text{ if } TPF_{d_1} \geq TPF_{d_2} \text{ and } FPF_{d_1} \leq FPF_{d_2} \qquad (9)$$

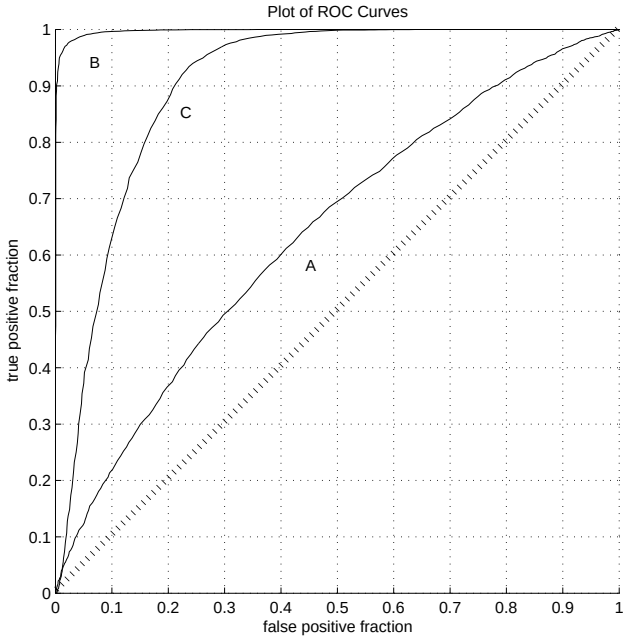
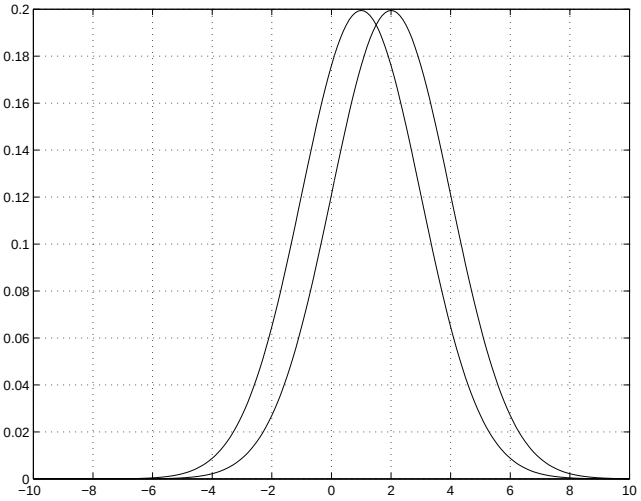
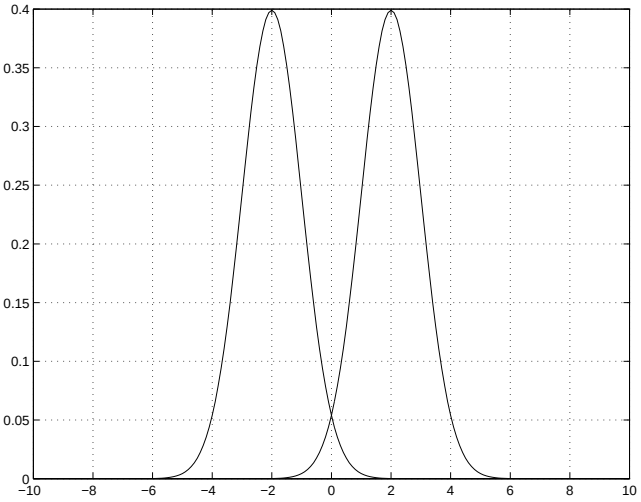


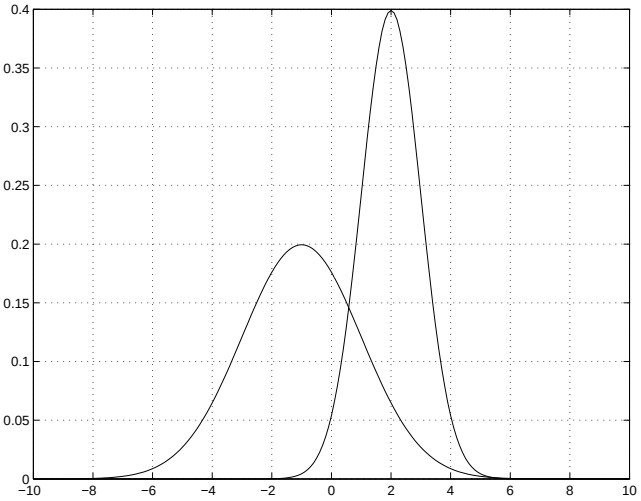

Figure 2: Three illustrative Receiver Operating Characteristic curves corresponding to distributions on Fig. 3.




(a) Non-distinguishable distributions: N(2,2) and N(1,2)




(b) Distinguishable distributions: N(2,1) and N(-2,1)




(c) Distributions N(2,1) and N(-1, 2)

Figure 3: Three illustrative pairs of distribution to generate ROC curves in Fig 2.

### 2.2.3 ROC Curve

A ROC curve is constructed calculating the true positive fraction and false positive fraction of a detection system for each possible decision threshold from strict to lenient. Figure 2 depicts three hypothetical ROC Curves. Notice that the shape of the ROC curve depends on how distinguishable are the underlying distributions (see Fig. 2 and Fig. 3). When the decision threshold is set to its strictest value $d_{max}$ then all decisions are negative and therefore the FPF is 0 as well as the TPF. This corresponds to the point $(0, 0)$. On the other hand, the point $(1, 1)$ corresponds to the the most lenient decision threshold $d_{min}$ when all decisions are positive and therefore both FPF and TPF are 1. A ROC curve encapsulates all the information provided by a confusion matrix for all possible values of a decision threshold. As a matter of fact, it is said that A ROC curve makes explicit the inherent tradeoff between sensivity and specificity for a number of different decision thresholds.

### 2.2.4 Slope of the Curve

At any point along the curve, the slope of the ROC curve $S$ measures the decision threshold used to generate that point [SDM00]. $S$ can be used to determine the optimal decision threshold as we will see in Sec. 5.

### 2.2.5 Area Under the ROC Curve

The area under the curve (AUC) gives an estimation of the accuracy of a detection system [HM82]. The higher the curve, the greater the accuracy. A perfect detection system will have an area of 1. A detection system with no discriminating power will have an area of 0.5. There exist two types of methods to compute the area under the ROC curve[2]: parametric and non-parametric methods [HHJ97]. Parametric methods use a maximum likelihood estimator of the AUC [MSL76] whereas non-parametric methods approximate the AUC by means of trapezoids. This last method has be shown to be equal to Mann-Whitney-Wilcoxon test statistic [HM82, HT01, HH02]. See [HT01] for a generalization of the area under the ROC curve to multiple class classification problems. As we will see later, these methods allows one to compare one or several detection systems in different populations[3].

Non-parametric detection systems can be represented in a ROC space by a single ROC point. Sometimes in the literature [DCW99, LFG00] it has been suggested to complete a ROC curve for non-parametric detection systems connecting that single point with straight lines to points $(0, 0)$ and $(1, 1)$. However, such as it has been shown elsewhere, it is pointless [GU01].

In this section we have introduced some basic concepts that we will use in our framework to evaluate alert triage in three scenarios: when the costs of misdetection are obviated, when

[2]Notice that the area under the ROC curve is equivalent to Gini index in Lorenz diagrams [Gas72].

[3]The Department of Radiology of the University of Chigaco provides a number of computer programs to compute the area under the ROC curve. They are available at http://www-radiology.uchicago.edu/krl/toppage11.htm.

the costs of misdetection are known a priori, and when we cope with imprecise environments and the costs are unknown a priori and can change dynamically.

# 3 ROC Analysis in Intrusion Detection

The beginnings of ROC analysis in intrusion detection were misleading. ROC analysis was originally introduced in the evaluation of IDSes in the 1998 DARPA experiments carried out by MIT Lincoln Laboratory [LFG00, McH00]. MIT researchers performed a cost insensitive evaluation of DARPA funded IDSes. However their investigations were based on some misconceptions of ROC analysis such as signaled in [McH00]:

1. They did not employ a comprehensive unit measure to properly construct confusion matrices.

2. The evaluation of parametric IDSes were performed without previously selecting an operating point for those IDSes.

3. The interpretation of the results did not use a standard ROC space.

In [SFL00] an method based on cost metrics was proposed as an alternative to ROC analysis to study the performance of IDSes. However, in their evaluations they did not use all the valuable information provided by ROC analysis such as it was demonstrated later in [GU01]. To the best of our knowledge, the work in [GU01] has been the soundest proposal of ROC analysis for the intrusion detection community. They provide a method grounded in decision, cost and ROC analysis to compare the performance of intrusion detectors. However, they did not provide any method for the construction of the ROC curves that they analyzed given that as mentioned above it is hard to find a comprehensive unit of measure when a complete intrusion detection system is evaluated. Other works, such as [THC02], have proposed complete models for the assessment of risk in intrusion detection. However, they use simple performance metrics instead of the more powerful analysis capabilities of ROC curves.

# 4 ROC Alternatives

Next, we briefly overview two ROC alternatives. See [GTD01] for an additional comparison between ROC analysis and the Taguchi method for quality engineering.

## 4.1 Detection Error Tradeoff

DET (Detection Error Tradeoff) curves were introduced in [MDK97] as a variant of ROC curves. The main difference with respect to ROC curves is that a DET curve shows the tradeoff between the two types of error involved in a detection task. A DET curve plots a performance curve showing the range of possible operating characteristics. As ROC curves

the abcissa axis shows the false positive fraction however the ordinate axis shows the false negative fraction instead of the true positive fraction. Figure 4 depicts three ROC curves and their alternative representation using DET curves. A DET curve gives equal treatment to false negatives and false positives what produces quasi-linear plots. When the curves are straight lines means that the underlying likelihood distributions are normal. DET curves have been used in large vocabulary speech recognition tasks.

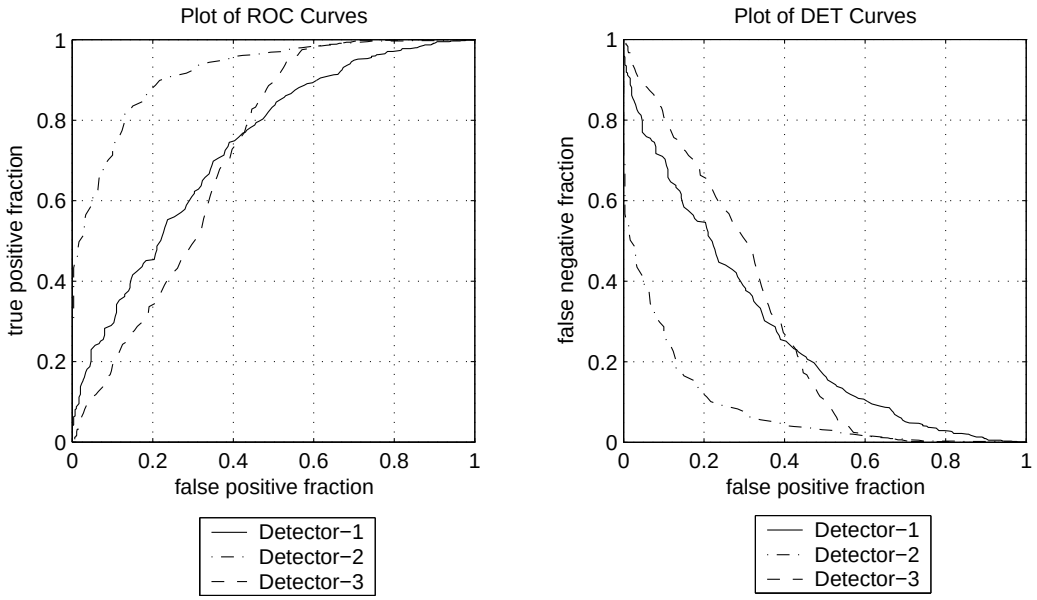



Figure 4: ROC Curves and DET Curves Comparison. DET curves plot error fractions on both axes abcissa and ordinate.

## 4.2 Explicit Representation of Expected Costs

Other alternative to ROC representation is to explicitly represent the cost of a detection system such as was proposed in [DH00]. An explicit representation allows one to directly seize the range of frequencies and the range of costs where a detection system surpass the others and to quantify how much better it is. Explicitly representing expected cost sets a cost space where the abcissa axis designates the probability-cost function for positive examples $PFC(C+)$ and the ordinate axis indicates the expected cost normalized ($NEC$) with respect to the cost of the worst detection system.

$$PFC(C+) = \frac{P(C+) \cdot C(D- \mid C+)}{P(C+) \cdot C(D- \mid C+) + P(C-) \cdot C(D+ \mid C-)} \tag{10}$$

$$NEC = (1 - TPF - FPF) \cdot PCF(C+) + FPF \tag{11}$$

An important feature of explicitly representing expected cost is its duality with respect to ROC space. That is, a line in the ROC space can be directly transformed to a point in the cost space whereas a point in the ROC space is converted into a line in the cost space.

**Bidirectional point/line duality**

$$\begin{aligned} NEC &= (1 - TP_o) \times PCF(C+) \\ PFC(C+) &= \frac{1}{1+S} \end{aligned} \tag{12}$$

Where $S$ is the slope of the line and $TP_o$ the intersection with the abcissa axis. Fig. 5 depicts 3 ROC points and their dual representation whose error type weighting is 1:10 (i.e. a cost of 1 for each false positive and a cost of 10 for each false negative). We will be back to Fig. 5 later on. The main advantage of explicitly representing expected cost is its comprehensible visual interpretation.

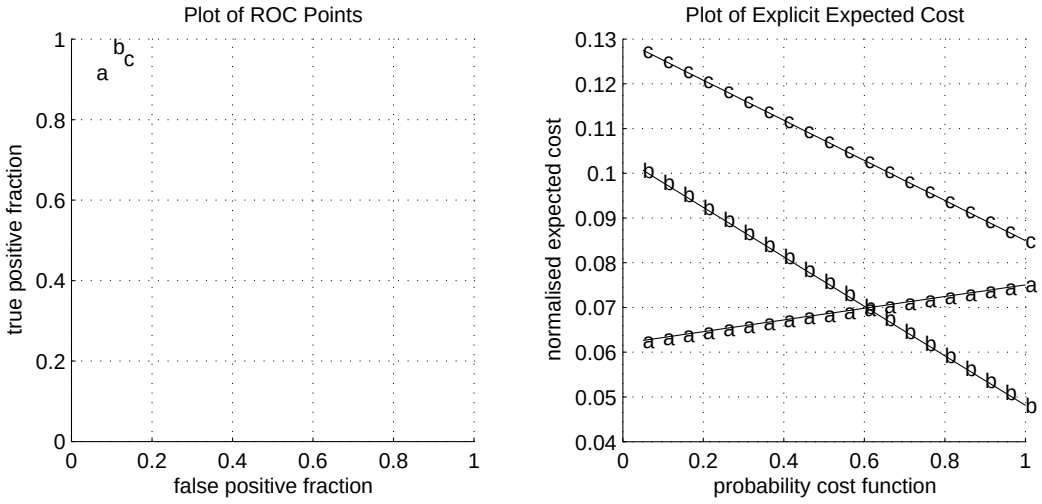



Figure 5: Three ROC points and their dual representation explicitly representing expected cost.

# 5 A Formal Framework for the Evaluation of Alert Triage

Alert triage is the process of rapid and approximate priorization for subsequent action of an IDS alert stream [MP01]. Alert triage is a classification task that takes as input the alerts generated by a number of (distributed) intrusion detection sensors (e.g. Snort [Roe99]) and classifies them producing a tag indicating its malignancy (degree of threat) for each alert. Such tags priorize alerts according to their malignancy. Depending on its priorization each alert will require an automatic response, notification, or manual response from the site security officer (SSO). The tag assigned to alerts can be a (fuzzy) label (e.g. `malicious`,

`innocuous`, `falsepositive`, etc) or it can be a continuous value. Alert triage can be explained in terms of the following priority formulation [PFV02] where *mission* is defined in terms of the critical and sensitive assets of the target network.

**Priority Formulation** Let an alert stream $\mathcal{A} = \langle a_1, a_2, \ldots, a_{t_n} \rangle$ find:
$\mathbf{HighImpact} = \{a_\alpha, a_\beta, \ldots, a_\psi\} \subseteq \mathcal{A}$ :

$$\forall_{a_i \in \mathbf{HighImpact}} \mathbf{ThreatRank}(e_i, \mathbf{Mission}) > T_{\mathbf{acceptable}} \tag{13}$$

For the sake of brevity, we suppose that once the alerts have been priorized there will be only two possible responses {**notification**, ¬**notification**}. In addition, looking at Eq. 13 it can be said that the decision threshold $T_{\mathbf{acceptable}}$ will discern between two (apparently) disjoint classes of alerts (i.e those alerts that require notification to the SSO and those alerts that do not). We, therefore, without loosing generality, proceed in our evaluation as if alert triage was a detection task[4]. We consider the evaluation of alert triage as a tournament process where a raking is established among the competing systems. We proceed as follows:

1. Participants are provided with a detailed model of the target networks and their missions using a common ontology [MP03b].
2. A window of $N$ consecutive alerts is selected from the alert stream (e.g. extracted from a database) and sent to the participants.
3. Participants emit their judgement on each alert.
4. The outcomes of each participant are represented by means of a ROC point (non-parametric systems) or ROC curve (parametric systems). The decision threshold (operating point) of parametric systems will be varied in the evaluations so that all possible values of the decision threshold can be evaluated.
5. The system or group of systems that outperforms the rest of participants is selected as the winner.

The performance of the participants will depend on the environment where the evaluation is carried out. We contemplate three possible environments (scenarios) in increased order of uncertainty and therefor of complexity. It is worthy to notice here that the first environment is a particular case of the second environment that in turn is subsumed by the third environment.

[4]Notice that as we mentioned above and we will see below the mathematical methods used in this work (AUC, ROC Convex Hull) can be extended to multiple class classification [HT01, Sri99].

## 5.1 Ideal Alert Triage Environment

In this environment, we do not contemplate misdetection or misclassification costs at all, neither a utility function over the decisions that are made correctly. This scenario is valid at design time when we want to check new triage techniques and for example fix some kind of requirement such as the maximum number of false positives allowable [Axe00]. The goal of alert triage in this scenario is either to maximize the overall percentage of correct decisions or minimize the overall percentage of incorrect decisions. The performance of parametric systems can be computed using the area under the ROC curve as we saw in Sec. 2 whereas the performance of non-parametric systems can be measured using Eq. 8 but as we saw in Sec. 2 accuracy is not always a good measure, particularly, if true negatives abound [KHM97]. There exist other performance measures in the literature such as:

**e-distance** E-distance is the Euclidean distance from the perfect classifier (point $(0,1)$) and the ROC point of interest [HGF02].

$$\text{e-distance} = 1 - \sqrt{W \cdot (1 - TPF)^2 + (1 - W) \cdot FPF^2} \tag{14}$$

Where $W$ is a parameter that ranges between 0 and 1 and establishes the relative importance between false positives and false negatives.

**f-measure** F-measure is a combination of recall (TPF) and precision (PPV) [LG94].

$$\text{f-measure} = \frac{(\beta^2 + 1) \cdot TPF \cdot PPV}{\beta^2 \cdot PPV + TPF} \tag{15}$$

Where $\beta$ is a parameter ranging from 0 to infinity that weights recall and precision.

**g-mean** Geometric mean is high when both TPF and TNF are high and when the difference between both is small [KHM97].

$$\text{g-mean} = \sqrt{TPF \times TNF} \tag{16}$$

**t-area** T-area is the area of the quadrilateral formed by the segments connecting the ROC point of interest and all the singular points of the ROC space except the perfect detection system. We compute T-area using Heron's formula:

$$\text{t-area} = \begin{cases} 1/2 + \sqrt{s \cdot (s-a) \cdot (s-b) \cdot (s-c)} & \text{if } TPF > FPF \\ 1/2 & \text{if } TPF = FPF \\ 1/2 - \sqrt{s \cdot (s-a) \cdot (s-b) \cdot (s-c)} & \text{if } TPF < FPF \end{cases} \tag{17}$$

Where $s = \frac{a+b+c}{2}$ i.e. half the perimeter, $a = \sqrt{2}$, $b = \sqrt{TPF^2 + FPF^2}$, and $c = \sqrt{(1 - TPF)^2 + (1 - FPF)^2}$. An advantage of this measure is that it makes parametric and non-parametric systems comparable.

Table 2: Confusion matrices for alert triage systems $a$, $b$, and $c$.

| a | | b | | c | |
|---|---|---|---|---|---|
| **TP** | **FP** | **TP** | **FP** | **TP** | **FP** |
| 133 | 53 | 156 | 87 | 101 | 116 |
| **FN** | **TN** | **FN** | **TN** | **FN** | **TN** |
| 12 | 802 | 3 | 754 | 5 | 778 |
| **TPF** | **FPF** | **TPF** | **FPF** | **TPF** | **FPF** |
| 0.9712 | 0.0620 | 0.9811 | 0.1034 | 0.9528 | 0.1298 |

Table 3: Accuracy measure results for alert triage systems $a$, $b$, and $c$.

| | a | b | c |
|---|---|---|---|
| **accuracy** | 0.9350 | 0.9100 | 0.8790 |
| **e – distance** | 0.9269 | 0.9256 | 0.9024 |
| **f – measure** | 0.8036 | 0.7761 | 0.6254 |
| **g – mean** | 0.9276 | 0.9379 | 0.9106 |
| **t – area** | 0.9276 | 0.9388 | 0.9115 |

Table 2 shows the confusion matrices of three alert triage systems $a$, $b$, and $c$ that have taken part in an evaluation using a window of 1000 alerts. The corresponding ROC points ($FPF_a = 0.0620, TPF_a = 0.9712$), ($FPF_b = 0.1034, TPF_b = 0.9811$), ($FPF_c = 0.1298, TPF_c = 0.9528$) are depicted in Fig. 6. We have used the above measures to rank them. The results are shown in Table 3 and Fig. 6. Clearly, the ROC points of $a$ and $b$ dominate the ROC point of $c$ (according to Eq. 9). All measures discern between $a$ & $b$ and $c$. However, there is no consensus among the different accuracy measures to signal $a$ or $b$ as a winner. We advocate for the use of t-area given that it has an intuitive explanation (ROC AUC), serves to compare parametric and non parametric systems, and does not depend on parameters that establish an artificial weight between the errors. On the other hand, to compare parametric and non-parametric participants we have to check whether the ROC point of a non-parametric participant is or not beneath the area of the parametric detection system under comparison.

In intrusion detection misdetection costs are asymmetric. That is, the cost of notifying a SSO when an alert corresponds to an innocuous attack (or false positive) is really lower compared with the cost of not adverting the presence of an intruder. Thus, once we have designed an alert triage system we should examine that it will contemplate such difference. Next scenarios allow one to evaluate an alert triage system in a cost sensitive way.

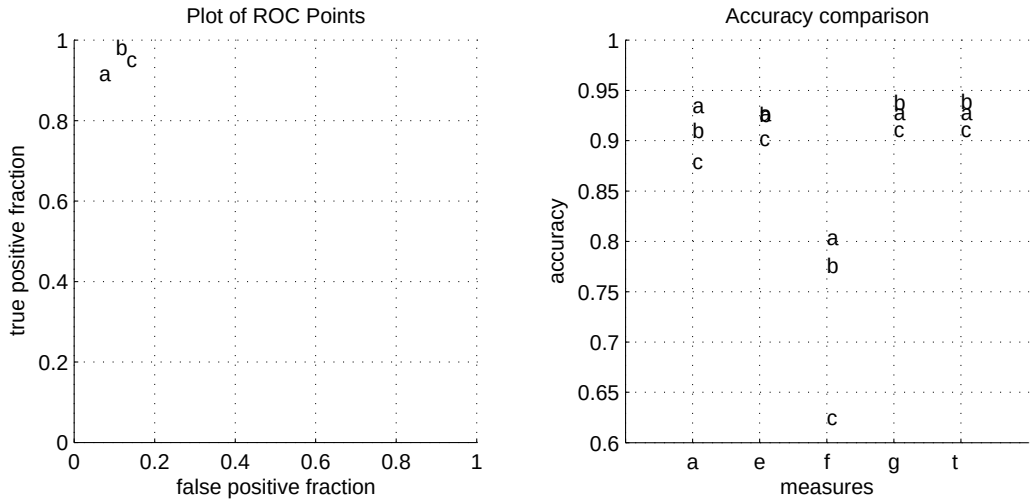



Figure 6: ROC points and ranking corresponding to three alert triage systems.

## 5.2 Cost-based Alert Triage Evaluation

We consider now scenarios where correct decision outcomes have associated a benefit $B$ and incorrect decision outcomes have associated a cost $C$. $B(A \mid A)$ represents the benefit obtained for correctly classifying an alert of type $A$ and $C(A \mid B)$ is the cost incurred if an alert of type B was misclassified as being an alert of class A. These scenarios are valid to test alert triage systems in simulated environments and to determine their optimal decision threshold. In this case, an alert triage detection system outperforms another if it has a lower expected cost. Thus, the goal here is to maximize the expected value of each decision. The expected cost of a non-parametric detection system or a parametric detection system operating at a given decision threshold is given by:

$$\begin{array}{c} \textbf{Expected Cost} \\ P(C+) \cdot P(D+ \mid C+) \cdot B(D+ \mid C+) + (1 - P(C+)) \cdot P(D- \mid C-) \cdot B(D- \mid C-) \quad + \\ P(C+) \cdot P(D- \mid C+) \cdot C(D- \mid C+) + (1 - P(C+)) \cdot P(D+ \mid C-) \cdot C(D+ \mid C-) \end{array} \tag{18}$$

Two consequences can be inferred directly from 18:

1. Two detection system will have the same expected value if:

$$\frac{TPF_2 - TPF_1}{FPF_2 - FPF_1} = \frac{1 - P(C+)}{P(C+)} \times \frac{B(D- \mid C-) + C(D+ \mid C-)}{B(D+ \mid C+) + C(D- \mid C+)} \tag{19}$$

   Equation 19 defines the slope of an *iso-performance* line [PF01].

2. The slope $S$ that corresponds to the optimal decision threshold can be computed as follows:

$$S_{\mathbf{optimal}} = \frac{1 - P(C+)}{P(C+)} \times \frac{B(D- \mid C-) + C(D+ \mid C-)}{B(D+ \mid C+) + C(D- \mid C+)} \tag{20}$$

Fig. 5 shows the explicit representation of costs using the alternative introduced in Sec. 4.2 for the alert triage systems analyzed in Sec. 5.1. It is appreciable that the alert triage system $b$ will be better under certain cost conditions than system $a$ and viceversa. Thus, if the conditions are known a priori we can clearly choose $a$ or $b$.

Once an intrusion detection system has been developed, its operational costs will depend on the the importance of the target's mission (system under surveillance) and the nature of the possible future attacks (e.g. to be attacked by a CodeRed propagation is quite different from being wounded by a malefactor or experiencing a DoS attack) and the level of hostility [GU01]. The ultimate objective of intrusion detection is to develop robust systems able to face imprecise environments, thus we also have to address this type of scenarios.

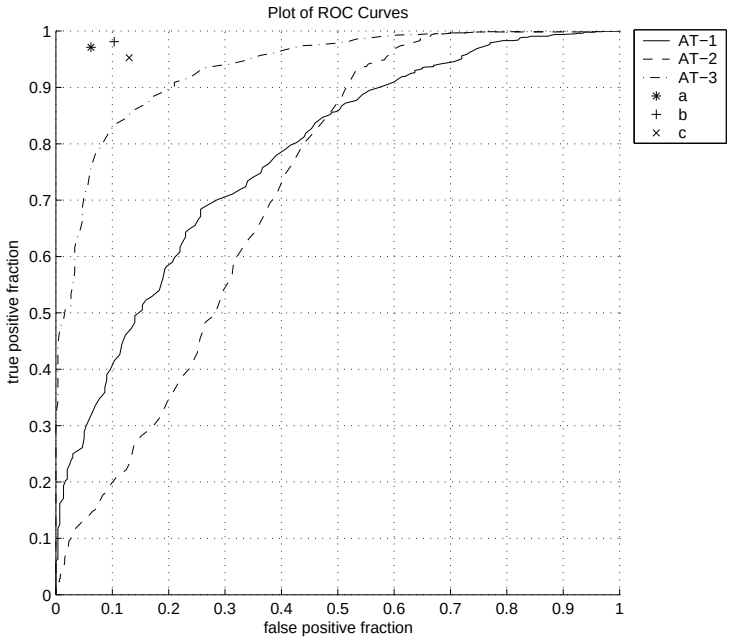



Figure 7: Three ROC points and three ROC curves representing six alert triage systems.

## 5.3 Alert Triage Evaluation in Imprecise Environments

In these scenarios, misdetection costs not only will be unknown a priori but also will vary over time. These scenarios are useful for the evaluation of systems for real-world deployment. To decide whether an alert triage systems outperforms others in this type of environment we will use a robust and incremental method for the comparison of multiple detection systems in imprecise and dynamic environments that has been proposed in [PF01]. This method, named

ROCCH (ROC Convex Hull) is a combination of ROC analysis, computational geometry and decision analysis. Thus, once the ROC points or ROC curves have been computed for the different participants we proceed as follows to select the best alert triage systems:

1. Firstly, we compute the convex hull[5] of the set of ROC points that symbolizes a composite alert triage system.

2. Secondly, we compute the set of iso-performance lines corresponding to all possible cost distributions.

3. Finally, for each iso-performance line we select the point of the ROC convex hull with the largest TPF that intersects it.

Figure 7 depicts the alert triage systems analyzed above. To select which systems will be of interest for a serie of unknown conditions we compute the convex hull such as it is shown in Fig. 8. In that figure, the circled points denote the points that forms part of the convex hull and therefore can be optimal for a number of conditions. Thus they form part of the composite alert triage system chosen as a winner. Therefore, we can continue our analysis without the non-parametric system $c$ and the parametric systems 2 and 3. Finally, Fig. 9 shows two illustrative iso-performance lines corresponding to the slopes $S_1$ and $S_2$. Thus, for each of the different sets of conditions that determine the values of slopes $S_1$ and $S_2$ the alert triage systems $a$ and $b$ will be the optimal respectively.

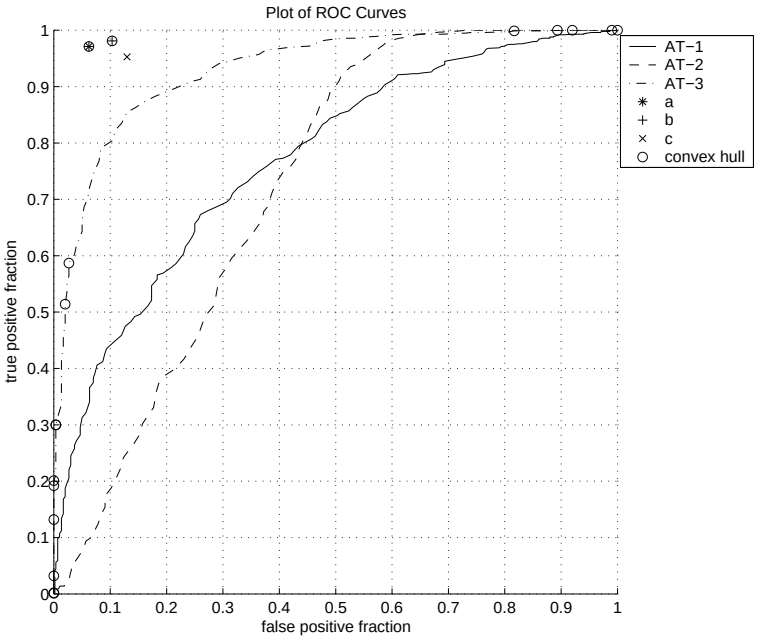



Figure 8: ROC Convex Hull of ROC points and curves of Fig. 7.

[5]The convex hull of a set of points is the smallest convex set that includes the points.

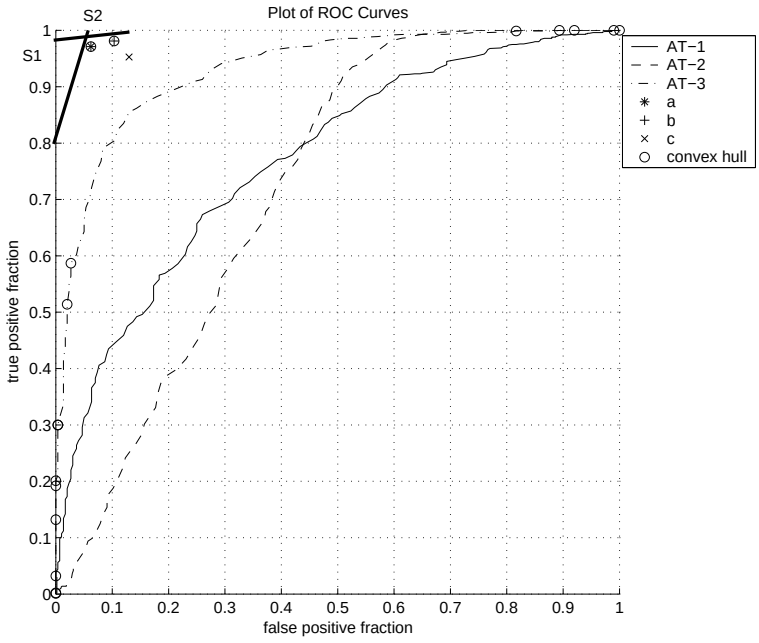



Figure 9: Iso-performance lines for different sets of conditions.

# 6 Conclusions

The ultimate goal of intrusion detection is to develop systems able to perform their task confronting imprecise environments. That requires to evaluate the systems under a different range of conditions in order to properly design and develop them to cope with different operational scenarios. However, the number of methodologies for the evaluation of intrusion detection systems or their different components is really scarce [DM02, DCW99, GU01, LFG00, McH00]. Moreover, the first experiences on the evaluation of intrusion detection systems have not been satisfactory [LFG00, McH00].

This work takes the first step towards the construction of a formal framework for the evaluation of a specific component of intrusion detection systems —*alert triage*. This framework will not only allow one to select the best alert triage system but also to make practical choices when assessing different components of alert triage. We have characterized alert triage as a detection task such that for each alert in the alert stream an action is taken {**notification**, ¬**notification**} in function of a previous judgement on its malignancy. Notice here that our framework also contemplates the evaluation of systems that need to evaluate a complete window of alerts in the alert stream before they are able to emit a judgement on the malignancy of each individual alert [MP03a].

ROC analysis constitutes the basis of our framework. Thus, we have tried to provide a compact explanation of ROC elements and some alternatives. ROC analysis allows one to establish when a detection system has the best performance indepently of distribution of the condition of interest and the cost of misdetections. We have seen how in these circumstances

is difficult to choose between certain alternatives (i.e. system $a$ and system $b$) because different measures to estimate the accuracy do not agree. We do recommend the use of the t-area for this purpose for the reasons mentioned above. Then, we have also shown that when misdetection costs are known a proper decision can be made using alternatives such as explicitly representing expected cost. Finally, when we deal with imprecise environments then the optimal alert triage detection system lie on the edge of the convex hull of ROC points that represent the detection system in a ROC space. The convex hull that dominates all the ROC points of a set of alert triage systems determines the group of best alert triage systems to confront imprecise environments.

This work forms part of a more ambitious effort where we are involved in developing case-based reasoning techniques [AP94] for the assessment of dynamic processes in imprecise and adversarial environments [MP03b, MP03a]. Currently, we aim at constructing some datasets composed of thousands of alerts compiled in real world environments during months of mission critical activities in order to put in practice our framework.

**Acknowledgments.**

Part of this work has been performed in the context of the MCYT-FEDER project SAMAP (TIC2002-04146-C05-01) and the SWWS project funded by the EC under contract number IST-2001-37134.

# References

[AP94] Agnar Aamodt and Enric Plaza. "Case-Based Reasoning: Foundational Issues, methodological variations, and system approaches." *Artificial Intelligence Communications*, **7**(1):39–59, 1994.

[Axe00] Stefan Axelsson. "The Base-Rate Fallacy and the Difficulty of Intrusion Detection." *ACM Transactions on Information and System Security*, **3**(3):186–205, August 2000.

[CM02] F. Cuppens and A. Miege. "Alert correlation in a cooperative intrusion detection framework." In *IEEE Symposium on Research in Security and Privacy*, 2002.

[DCW99] Robert Durst, Terrence Champion, Brian Witten, Eric Miller, and Luigi Spagnuolo. "Testing and evaluating computer intrusion detection systems." *Communications of the ACM*, **42**(7):53–61, 1999.

[DH00] Chris Drummong and Robert C. Holte. "Explicity Representing Expected Cost: An Alternative to ROC Representation." In *Sixth ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2000.

[DM02] Hervé Debar and Benjamin Morin. "Evaluation of the Diagnostic Capabilities of Commercial Intrusion Detection Systems." In *Proc. of the RAID 2002*, number 2516 in Lecture Notes in Computer Science, pp. 177–198. Springer, 2002.

[DW01] Hervé Debar and Andreas Wespi. "Aggregation and Correlation of Intrusion Detection Alerts." In *Proceedings of the 4th symposium on Recent Advances in Intrusion Detection (RAID 2001)*, 2001.

[Ega75] J. P. Egan. *Signal Detection Theory and ROC Analysis.* Academic Press, 1975.

[FFH02] Cèsar Ferri, Peter Flach, and José Hernández-Orallo. "Learning Decision Trees Using the Area Under the ROC Curve." In Claude Sammut and Achim Hoffmann, editors, *Proceedings of the 19th International Conference on Machine Learning*, pp. 139–146. Morgan Kaufmann, July 2002.

[Gas72] Joseph L. Gastwirth. "The Estimation of the Lorenz Curve and Gini Index." *The Review of Economics & Statistics*, **54**(3):306–316, 1972.

[GHH01] Robert P. Goldman, Walter Heimerdinger, Steven A. Harp, Christopher W. Geib, Vicraj Thomas, and Robert L. Carter. "Information Modeling for Intrusion Report Aggregation." In *DICEX*. IEEE Computer Society, 2001.

[GTD01] Albert Güveni, Tolga Taner, and Irini Dimitriyadis. "Evaluation of Taguchi and ROC Techniques for the Quality Assessment of Binary Decision Models in Health Care and Industrial Aplications." http://www.bme.boun.edu.tr/guvenis/taguchi.htm, 2001.

[GU01] John E. Gaffney and Jacob W. Ulvila. "Evaluation of Intrusion Detectors: A Decision Theory Approach." In *The IEEE Symposium on Security and Privacy*, 2001.

[HGF02] Howard Hamilton, Ergun Gurak, Leah Findlater, and Wayne Olive. "Computer Science 831: Knowledge Discovery in Databases." http://www2.cs.uregina.ca/~hamilton/courses/831/index.html, 2002.

[HH02] Karim O. Hajian-Tilaki and J. A. Hanley. "Comparisson of Three Methods for Estimating the Standard Error of the Area under the Curve in ROC Analysis of Quantitative Data." *Acad Radiol*, **9**:1278–1285, 2002.

[HHJ97] Karim O. Hajian-Tilaki, J. A. Hanley, L. Joseph, and J. P. Collet. "A Comparison of Parametric and nonparametric approaches to ROC analysis of Quantitative Diagnostic Tests." *Medical Decision Making*, pp. 94–102, 1997.

[HLF01] Joshua W. Haines, Richard P. Lippmann, David J. Fried, Eushiuan Tran, Steve Boswell, and Marc A. Zissman. "1999 DARPA Intrusion Detection System Evaluation: Design and Procedures." Technical report, MIT Lincoln Laboratory, 2001.

[HM82] J. A. Hanley and B. J. McNeil. "The meaning and use of the area under a receiver operating characteristic ROC curve." *Radiology*, **143**:29–36, 1982.

[HT01] David J. Hand and Robert J. Till. "A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems." *Machine Learning*, **45**:171–186, 2001.

[KHM97] Miroslav Kubat, Robert Holte, and Stan Matwin. "Learning when Negative Examples Abound." In *European Conference on Machine Learning*, 1997.

[LFG00] Richard Lippmann, David Fried, Isaac Graf, Joshua Haines, Kristopher Kendall, David McClung, Dan Weber, Seth Webster, Dan Wyschogrod, Robert Cunningham, and Marc Zissman. "Evaluating Intrusion Detection Systems: The 1998 DARPA Off-line Intrusion Detection Evaluation." In *Proceedings of the DARPA Information Survivability Conference and Exposition*, Los Alamitos, CA, 2000. IEEE Computer Society Press.

[LG94] David D. Lewis and William A. Gale. "A sequential algorithm for training text classifiers." In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 3–12. Springer-Verlag New York, Inc., 1994.

[McH00] John McHugh. "Testing Intrusion Detection Systems: A Critique of the 1998 and 1999 DARPA Intrusion Detection System Evaluations as Performed by Liconln Laboratory." *ACM Transactions on Information and System Security*, **3**(4):262–294, November 2000.

[MDK97] Alvin Martin, George Doddington, Terri Kamm, Mark Ordowski, and Mark Przybocki. "The DET Curve in Assessment of Detection Task Performance." In *Proc. Eurospeech '97*, pp. 1895–1898, Rhodes, Greece, 1997.

[MMD02] Benjamin Morin, Ludovic Mé, Herveé Debar, and Mireille Ducassé. "M2D2: A Formal Data Model for IDS Alert Correlation." In *Proc. of the RAID 2002*, 2002.

[MP01] Sofus A. Macskassy and Foster Provost. "Intelligent Information Triage." In *The 24th Annual International Conference on Research and Development in Information Retrieval*, 2001.

[MP03a] Francisco J Martin and Enric Plaza. "Case-based Sequence Analysis for Intrusion Detection." In *ICCBR 2003 Workshop on Applying CBR to Time Series Prediction. Trondheim, Norway*, 2003.

[MP03b] Francisco J Martin and Enric Plaza. "SOID: an Ontology for Agent-Aided Intrusion Detection." In *7th International Conference on Knowledge-based Intelligent Information & Engineering Systems*, 2003.

[MSL76] C. Metz, S. Starr, and L. Lusted. "Observer performance in detecting multiple radiographic signals: Prediction and analysis using a generalized ROC approach." *Radiology*, pp. 337–347, 1976.

[NC02] Peng Ning and Yun Cui. "An Intrusion Alert Correlator Based on Prerequisites of Intrusions." Technical Report TR-2002-01, Department of Computer Science, North Carolina State University, 2002.

[NCR02] Peng Ning, Yun Cui, and Douglas S. Reeves. "Analyzing Intesive Intrusion Alerts via Correlation." In Andreas Wespi, Giovanni Vigna, and Luca Deri, editors, *Proceedings of the RAID 2002*, number 2516 in Lecture Notes in Computer Science, pp. 74–94. Springer, 2002.

[PF01] Foster Provost and Tom Fawcett. "Robust Classification for Imprecise Environments." *Machine Learning Journal*, **42**(3), 2001.

[PFK98] Foster Provost, Tom Fawcett, and Ron Kohavi. "The Case Against Accuracy Estimation for Comparing Induction Algorithms." In *Fifteenth International Conference on Machine Learning*, 1998.

[PFV02] Phillip A. Porras, Martin W. Fong, and Alfonso Valdes. "A Mission-Impact-Based Approach to INFOSEC Alarm Correlation." In *Proc. of the RAID 2002*, number 2516 in Lecture Notes in Computer Science, pp. 95–114. Springer, 2002.

[Ran01] Marcus J. Ranum. "Experiences Benchmarking Intrusion Detection Systems." Technical report, NFR Security, 2001.

[Roe99] Martin Roesch. "Snort - Lightweight Intrusion Detection for Networks." In *Proceedings of LISA '99: 13th Systems Administration Conference Seattle, Washington, USA*, November 1999.

[SDM00] John A. Swets, Robyn M. Dawes, and John Monahan. "Psychological Science can Improve Diagnostic Decisions." *Psychological Science in the Public Interest*, **1**(1), 2000.

[SFL00] Sal Stolfo, Wei Fan, Wenke Lee, Andreas Prodromidis, and Philip K. Chan. "Cost-based Modeling for Fraud and Intrusion Dectection: Results from the JAM Project." In *Proceedings of DISCEX*, 2000.

[Sri99] Ashwin Srinivasan. "Note on the Location of Optimal Classifiers in N-Dimensional ROC Space." Technical Report PG-TR-2-99, Oxford University Computing Laboratory, 1999.

[SS98] S. Staniford-Chen and D. Schnackenberg. "The Common Intrusion Detection Framework (CIDF)." In *Information Survivability Workshop*, October 1998.

[Swe96] John A. Swets. *Signal Detection Theory and ROC Analysis in Psychology and Diagnostics. Collected Papers.* Lawrence Erlbaum Associates, 1996.

[THC02] Sapon Tanachaiwiwat, Kai Hwang, and Yue Chen. "Adaptive Intrusion Response to Minimize Risk over Multiple Network Attacks." In *Submitted*, 2002.

[WE02] M. Wood and M. Erlinger. "Intrusion Detection Message Exchange Requeriments." Internet draft (Work in Progress), 2002.

[WF02] S. Wu and P. A. Flach. "Model selection for dynamic processes." In M. Bohanec, B. Kašek, N. Lavrač, and D. Mladenic, editors, *ECML/PKDD'02 workshop on Integration and Collaboration Aspects of Data Mining, Decision Support and Meta-Learning*, pp. 168–173. University of Helsinki, August 2002.