# The Base-Rate Fallacy and the Difficulty of Intrusion Detection

STEFAN AXELSSON
Ericsson Mobile Data Design AB

Many different demands can be made of intrusion detection systems. An important requirement is that an intrusion detection system be *effective*; that is, it should detect a substantial percentage of intrusions into the supervised system, while still keeping the false alarm rate at an acceptable level. This article demonstrates that, for a reasonable set of assumptions, the false alarm rate is the limiting factor for the performance of an intrusion detection system. This is due to the base-rate fallacy phenomenon, that in order to achieve substantial values of the Bayesian detection rate $P(Intrusion|Alarm)$, we have to achieve a (perhaps in some cases unattainably) low false alarm rate. A selection of reports of intrusion detection performance are reviewed, and the conclusion is reached that there are indications that at least some types of intrusion detection have far to go before they can attain such low false alarm rates.

Categories and Subject Descriptors: D.4.6 [**Operating Systems**]: Security and Protection

General Terms: Performance, Security, Theory

Additional Key Words and Phrases: Base-rate fallacy, detection rate, false alarm rate, intrusion detection

## 1. INTRODUCTION

Many demands can be made of an intrusion detection system (IDS for short) such as *effectiveness, efficiency, ease of use, security, interoperability,*

This work was funded in part by the Swedish National Board for Industrial and Technical Development (NUTEK) under project P10435.
An earlier version of this article appeared as "The base-rate fallacy and its implications for the difficulty of intrusion detection" in the *Proceedings of the Sixth ACM Conference on Computer and Communications Security* (Nov. 1–9). ACM Press, New York, 1999, pp. 1–7.
Most of this work was done while the author was at the Department of Computer Engineering at Chalmers University of Technology. He is presently at Ericsson Mobile Data Design AB.
The author's homepage (and all self-referenced papers) can be found at http://www.ce.chalmers.se/staff/sax.
Author's address: Ericsson Mobile Data Design AB, S:t Sigfridsgatan 89, SE-412 66, Göteborg, Sweden; email: Stefan.Axelsson@erv.ericsson.se.

*transparency*, and so on. Although much research has been done in the field in the past 10 years, the theoretical limits of many of these parameters have not been studied to any significant degree. This article discusses one serious problem with regard to the *effectiveness* parameter, especially how the base-rate fallacy may affect the operational effectiveness of an intrusion detection system.

## 2. INTRUSION DETECTION

The field of automated computer intrusion detection (intrusion detection for short) is currently about 20 years old [Anderson 1980], with interest gathering pace during the past 10 years.

Intrusion detection systems are intended to help detect a number of important types of computer security violations, such as:

—attackers using prepacked "exploit scripts"; primarily outsiders;

—attackers operating under the identity of a legitimate user, for example, by having stolen that user's authentication information (password); outsiders and insiders;

—insiders abusing legitimate privileges, and so on.

Early work (see Anderson [1980], Denning and Neumann [1985], Denning [1987], and Sebring et al. [1988]) identified two major types of intrusion detection strategies.

*Anomaly Detection.* The strategy of declaring everything that is unusual for the subject (computer, user, etc.) suspect, and worthy of further investigation. The early anomaly detection systems were all self-learning, that is, they automatically formed an opinion of what the subject's normal behavior was.

Anomaly detection promises to detect abuses of legitimate privileges that cannot easily be codified into security policy, and to detect attacks that are "novel" to the intrusion detection system. Problems include a tendency to take up data processing resources, and the possibility of an attacker teaching the system that his illegitimate activities are nothing out of the ordinary.

*Signature detection* The detection strategy of deciding in advance what type of behavior is undesirable, and through the use of predetermined signatures of such behavior, detecting intrusions.

Signature-based detection systems promise to detect known attacks and violations easily codified into security policies in a timely and efficient manner. Problems include a difficulty in detecting previously unknown intrusions. If a database containing intrusion signatures is employed, it must be updated frequently.

Early in the research it was suggested in Halme and Kahn [1988] and Lunt [1988] that the two main methods ought to be combined to provide a

complete intrusion detection system capable of detecting a wide array of different computer security violations, including the ones listed above.

For a more in-depth review of these and other intrusion detection concepts, the interested reader is referred to a survey of intrusion detection systems [Axelsson 1998] and a taxonomy of intrusion detection systems and principles [Axelsson 2000a], previously written by us.

We wish to at least make the above division between the different principles of detection, since it is easy to conjecture that these fundamentally different modes of detection will exhibit different characteristics with regard to detection and false alarm rates. They probably also show different performance in other characteristics as well, such as run-time efficiency, but a discussion of these parameters falls outside the scope of this article.

## 3. PROBLEMS IN INTRUSION DETECTION

At present, many fundamental questions regarding intrusion detection remain unanswered. They include, but are by no means limited to, the following:

*Effectiveness.* How effective is the intrusion detection? To what degree does it detect intrusions into the target system, and how good is it at rejecting false positives, so-called false alarms?

*Efficiency.* What is the run-time efficiency of the intrusion detection system, how many computing resources and how much storage does it consume, can it make its detections in real-time, and so on?

*Ease of use.* How easy is it to field and operate for a user who is not a security expert, and can such a user add new intrusion scenarios to the system? An important issue in ease of use is the question of what demands can be made of the person responding to the intrusion alarm. How high a false alarm rate can he/she realistically be expected to cope with, and under what circumstances is he/she likely to ignore an alarm? (It has long been known in security circles that, if you are an attacker, you should attempt to circumvent an ordinary electronic alarm system during normal operation of the facility, since if you happened to trigger the alarm, the supervisory staff would more likely be lax because they would be more accustomed to false alarms [Pierce 1948].)

*Security.* Whenever more intrusion detection systems are fielded, one would expect ever more attacks directed at the intrusion detection system itself, to circumvent it or otherwise render the detection ineffective. What is the nature of these attacks, and how resilient is the intrusion detection system to them?

*Interoperability.* As the number of different intrusion detection systems increase, to what degree can they interoperate and how do we ensure this?

*Transparency.* How intrusive is the fielding of the intrusion detection system to the organization employing it? How many resources will it consume in terms of manpower, and the like?

*Collaboration.* The best effect is often achieved when several security measures are brought to bear together. How should intrusion detection collaborate with other security mechanisms to achieve this synergy effect? How do we ensure that the combination of security measures provides at least the same level of security as each applied singly would provide, or that the combination does not in fact *lower* the overall security of the protected system?

Although interest is being shown in some of these issues, with a few notable exceptions (mainly Helman and Liepins [1993]), they remain largely unaddressed by the research community. This is perhaps not surprising, since many of these questions are difficult to formulate and answer.

This article is concerned with one aspect of one of the questions above, that of *effectiveness*. More specifically, it addresses the way in which the base-rate fallacy affects the required performance of the intrusion detection system with regard to false alarm rejection.

In what follows, Section 4 gives a description of the base-rate fallacy. Section 5 then continues with an application of the base-rate fallacy to the intrusion detection problem, given a set of reasonable assumptions. Section 6 describes the impact the results presented in the previous section would have on intrusion detection systems. Section 7 considers future work, with Section 8 concluding the article. Appendix A reproduces a base-rate fallacy example in diagram form.

## 4. THE BASE-RATE FALLACY

The base-rate fallacy[1] is one of the cornerstones of Bayesian statistics, stemming as it does directly from Bayes' famous theorem that states the relationship between a conditional probability and its opposite, that is, with the condition transposed:

$$P(A|B) = \frac{P(A) \cdot P(B|A)}{P(B)}. \tag{1}$$

Expanding the probability $P(B)$ for the set of all $n$ possible, mutually exclusive outcomes $A$, we arrive at Eq. (2),

$$P(B) = \sum_{i=1}^{n} P(A_i) \cdot P(B|A_i). \tag{2}$$

[1]The idea behind this approach stems from Matthews [1996; 1997].

Combining Eqs. (1) and (2), we arrive at a generally more useful statement of Bayes' theorem:

$$P(A)B = \frac{P(A) \cdot P(B|A)}{\sum_{i=1}^{n} P(A_i) \cdot P(B|A_i)} \tag{3}$$

The base-rate fallacy is best described through example.[2] Suppose that your doctor performs a test that is 99% accurate; that is, when the test was administered to a test population all of whom had the disease, 99% of the tests indicated disease, and likewise, when the test population was known to be 100% free of the disease, 99% of the test results were negative. Upon visiting your doctor to learn the results, he tells you he has good news and bad news. The bad news is that indeed you tested positive for the disease. The good news however, is that, out of the entire population, the rate of incidence is only $1/10000$; that is, only 1 in 10000 people have this ailment. What, given this information, is the probability of your having the disease? The reader is encouraged to make a quick "guesstimate" of the answer at this point.

Let us start by naming the different outcomes. Let $S$ denote sick, and $\neg S$, that is, *not* $S$, denote healthy. Likewise, let $R$ denote a positive test result and $\neg R$ denote a negative test result. Restating the information above: given: $P(R|S) = 0.99$, $P(\neg R|\neg S) = 0.99$, and $P(S) = 1/10000$, what is the probability $P(S|R)$?

A direct application of Eq. (3) gives:

$$P(S|R) = \frac{P(S) \cdot P(R|S)}{P(S) \cdot P(R|S) + P(\neg S) \cdot P(R|\neg S)}. \tag{4}$$

The only probability above that we do not immediately know is $P(R|\neg S)$. This is easily found though, since it is merely $1 - P(\neg R|\neg S) = 1\%$ (likewise, $P(\neg S) = 1 - P(S)$). Substituting the stated values for the different quantities in Eq. (4) gives:

$$P(S|R) = \frac{1/10000 \cdot 0.99}{1/10000 \cdot 0.99 + (1 - 1/10000) \cdot 0.01} = 0.00980\ldots \approx 1\%. \tag{5}$$

That is, even though the test is 99% certain, your chance of actually having the disease is only $1/100$, because the population of healthy people is much larger than the population with the disease. (For a graphical representation, in the form of a Venn diagram, depicting the different outcomes, see the Appendix). This result often surprises people, ourselves included, and it is this phenomenon—that humans in general do not take the basic rate of incidence, the base-rate, into account when intuitively

[2]This example is hinted at in Russel and Norvig [1995].

solving such problems of probability—that is aptly named "the base-rate fallacy."

## 5. THE BASE-RATE FALLACY IN INTRUSION DETECTION

In order to apply this reasoning in computer intrusion detection, we must first find the different probabilities, or if such probabilities cannot be found, make a set of reasonable assumptions regarding them.

### 5.1 Basic Frequency Assumptions

Let us, for the sake of further argument, hypothesize a figurative computer installation with a few tens of workstations, a few servers (all running UNIX), and a couple of dozen users. Such an installation could produce on the order of 1,000,000 audit records per day with some form of "C2" compliant logging in effect [US Department of Defense 1985], in itself a testimony to the need for automated intrusion detection.

Suppose further that, in such a small installation, we would not experience more than a few, say one or two, actual attempted intrusions per day. Even though it is difficult to get any figures for real incidences of attempted computer security intrusions, this does not seem to be an unreasonable number.

Furthermore, assume that, at this installation, we do not have the manpower to have more than one site security officer (SSO for short), who probably has other duties, and that the SSO, being only human, can only react to a relatively low number of alarms, especially if the false alarm rate is high (50% or so).

Even though an intrusion could possibly affect only one audit record, it is likely, on average, that it will affect a few more than that. Furthermore, a clustering factor actually makes our estimates more conservative, so it was deemed prudent to include one. Using data from a previous study of the trails that SunOS intrusions leave in the system logs [Axelsson et al. 1998], we can estimate that 10 audit records would be affected in the average intrusion.

### 5.2 Human-Machine Interaction in Intrusion Detection

The previous assumptions are "technical" in nature; that is, anyone well versed in the field of computer security can make similar predictions, or adjust the ones above to suit their liking. It is a simple matter to verify or predict similar measures. However, the factor of the performance of the human operator does not lend itself to the same technological estimates. Thus, a crucial question in this discussion is the capacity of the human operator to correctly respond to the output of the system, especially his/her capacity to tolerate false alarms.

Unfortunately, there have been no experiments concerning these factors in the setting of computer security intrusion detection. There is, however, some research in the context of process automation and plant control, such as would be the case in a (nuclear) power station, paper mill, steel mill,

large ship, and so on [Rasmussen 1986; Wickens 1992; Nygren 1994; Deatherage 1972]. These studies seem to indicate that our required level of false alarms, 50%, is a *very* conservative estimate. Most human operators will have completely lost faith in the device at that point, opting to treat every alarm with extreme skepticism, if one would be able to speak of a "treatment" at all. The intrusion detection system would most likely be completely ignored in a "civilian" setting. More research into this issue is clearly needed.

### 5.3 Calculation of Bayesian Detection Rates

Let $I$ and $\neg I$ denote *intrusive* and *nonintrusive* behavior, respectively, and $A$ and $\neg A$ denote the presence or absence of an intrusion alarm. We start by naming the four possible cases (false and true positives and negatives) that arise by working backwards from the above set of assumptions:

*Detection rate* (or *true positive* rate) is the probability $P(A|I)$; that is, that quantity that we can obtain when testing our detector against a set of scenarios we know represent intrusive behavior;

*False alarm rate* is the probability $P(A|\neg I)$, the *false positive* rate, obtained in an analogous manner.

The other two parameters, $P(\neg A|I)$, the *False Negative* rate, and $P(\neg A|\neg I)$, the *True Negative* rate, are easily obtained since they are merely

$$P(\neg A|I) = 1 - P(A|I); P(\neg A|\neg I) = 1 - P(A|\neg I). \tag{6}$$

Of course, our ultimate interest is that both:

—$P(I|A)$, that an alarm really indicates an intrusion (henceforth called the *Bayesian detection rate*), and

—$P(\neg I|\neg A)$, that the absence of an alarm signifies that we have nothing to worry about,

remain as large as possible.

Applying Bayes' theorem to calculate $P(I|A)$ results in:

$$P(I|A) = \frac{P(I) \cdot P(A|I)}{P(I) \cdot P(A|I) + P(\neg I) \cdot P(A|\neg I)}. \tag{7}$$

Likewise, for $P(\neg I|\neg A)$:

$$P(\neg I|\neg A) = \frac{P(\neg I) \cdot P(\neg A|\neg I)}{P(\neg I) \cdot P(\neg A|\neg I) + P(I) \cdot P(\neg A|I)}. \tag{8}$$

These assumptions give us a value for the rate of incidence of the actual number of intrusions in our system, and its dual (10 audit records per

intrusion, 2 intrusions per day, and 1,000,000 audit records per day). Interpreting these as probabilities:

$$P(I) = 1 \Big/ \frac{1 \cdot 10^6}{2 \cdot 10} = 2 \cdot 10^{-5};$$

$$P(\neg I) = 1 - P(I) = 0.99998. \tag{9}$$

Inserting Eq. (9) into Eq. (7),

$$P(I|A) = \frac{2 \cdot 10^{-5} \cdot P(A|I)}{2 \cdot 10^{-5} \cdot P(A|I) + 0.99998 \cdot P(A|\neg I)}. \tag{10}$$

Studying Eq. (10), we see the base-rate fallacy clearly. By now it should come as no surprise to the reader, since the assumptions made about our system make it clear that we have an overwhelming number of nonevents (benign activity) in our audit trail, and only a few events (intrusions) of any interest. Thus, the factor governing the *detection* rate ($2 \cdot 10^{-5}$) is completely dominated by the factor (0.99998) governing the *false alarm* rate. Furthermore, since $0 \leq P(A|I) \leq 1$, the equation will have its desired maximum for $P(A|I) = 1$ and $P(A|\neg I) = 0$, which results in the most beneficial outcome as far as the *false alarm* rate is concerned. While reaching these values would be an accomplishment indeed, they are hardly attainable in practice. Let us instead plot the value of $P(I|A)$ for a few fixed values of $P(I|A)$ (including the "best" case $P(A|I) = 1$), as a function of $P(A|\neg I)$ (see Figure 1). It should be noted that both axes are logarithmic.

It becomes clear from studying the plot in Figure 1 that, even for the unrealistically high *detection* rate 1.0, we have to have a very low *false alarm* rate (on the order of $1 \cdot 10^{-5}$) for the Bayesian detection rate to have a value of 66%, that is, about two-thirds of all alarms will be a true indication of intrusive activity. With a more realistic *detection* rate of, say, 0.7, for the same *false alarm* rate, the value of the Bayesian detection rate is about 58%, nearing 50-50. Even though the number of events (intrusions/alarms) is still low, it is our belief that a low Bayesian detection rate would quickly "teach" the SSO to (un)safely ignore *all* alarms, even though their absolute numbers would theoretically have allowed a complete investigation of all alarms. This becomes especially true as the system grows; a 50% false alarm rate of a total 100 alarms would clearly not be tolerable. Note that even quite a large difference in the *detection* rate does not substantially alter the Bayesian detection rate, which instead is dominated by the *false alarm* rate. Whether such a low rate of false alarms is at all attainable is discussed in Section 6.

It becomes clear that, for example, a requirement of only 100 false alarms per day is met by a large margin with a *false alarm* rate of $1 \cdot 10^{-5}$. With

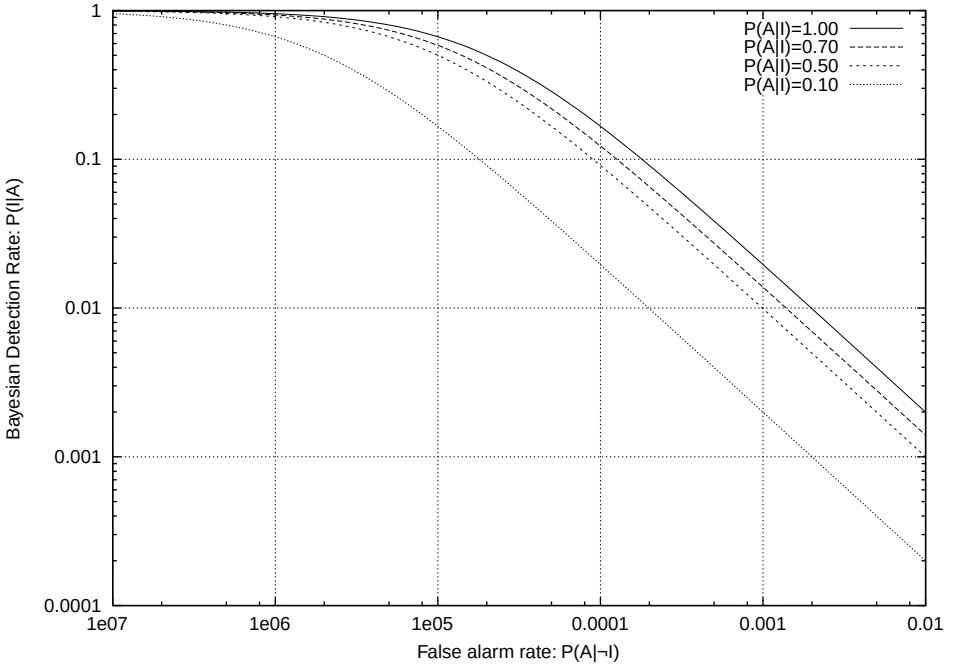


Fig. 1. Plot of Bayesian detection rate versus false alarm rate.

$10^5$ "events" per day, we will see only 1 *false alarm* per day, on average. By the time our ceiling of 100 false alarms per day is met, at a rate of $1 \cdot 10^{-3}$ *false alarms*, even in the best-case scenario, our Bayesian detection rate is down to around 2%,[3] by which time no one will care less when the alarm goes off.

Substituting (6) and (9) in Eq. (8) gives

$$P(\neg I \mid \neg A) = \frac{0.99998 \cdot (1 - P(A \mid \neg I))}{0.99998 \cdot (1 - P(A \mid \neg I)) + 2 \cdot 10^{-5} \cdot (1 - P(A \mid I))}. \quad (11)$$

A quick glance at the resulting Eq. (11) raises no cause for concern. The large $P(\neg I)$ factor (0.99998) will completely dominate the equation, giving it values near 1.0 for the values of $P(A \mid \neg I)$ under discussion here, regardless of the value of $P(A \mid I)$.

This is the base-rate fallacy in reverse, if you will, since we have already demonstrated that the problem is that we will set off the alarm too many times in response to nonintrusions, combined with the fact that, to begin with, we do not have many intrusions: truly a question of finding a needle in a haystack.

[3]Another way of calculating that differs from Eq. (10) is of course to realize that 100 false alarms and only a *maximum* of 2 possible valid alarms gives: $2/(2 + 100) \approx 2\%$.

The author does not see how the situation underlying the base-rate fallacy problem will change for the better in years to come. On the contrary, as computers get faster, they will produce more audit data, while it is doubtful that intrusive activity will increase at the same rate. In fact, it would have to increase at a substantially higher rate for it to have any effect on the previous calculations, and were it ever to reach levels sufficient to have such an effect, say 30% or more, the installation would no doubt have a serious problem on its hands, to say the least!

## 6. IMPACT ON INTRUSION DETECTION SYSTEMS

The previous section developed requirements regarding false alarm rates and detection rates in intrusion detection systems in order to make them useful in the stated scenario, where we would have 100,000 "events" (each consisting of 10 audit records), and only 2 intrusions per day, affecting 1 event each. This section compares these requirements with reported results on the effectiveness of intrusion detection systems.

As stated in the introduction, approaches to intrusion detection can be divided into two major groups, *signature*-based and *anomaly*-based. It can be argued that our scenario does not apply to anomaly-based intrusion detection as it, in some cases, tries not to detect intrusions per se, but rather to differentiate between two different subjects, flagging anomalous behavior in the hopes that it is indicative of a stolen user identity. From that perspective, our assumption that an "attack" only affects one event (10 audit records) in the audit logs would be less well founded, since it is possible that a masquerader would affect considerably more audit records than that. Lane and Brodley [1999] study the problem of how to differentiate between different users based on the traces their actions leave in audit logs. However, we still think our scenario is useful as a description of a wide range of more "immediate," often network-based, attacks, where we will not have had the opportunity to observe the intruder for an extended period of time "prior" to the attack. Since anomaly-based intrusion detection systems promise other advantages, the ability to detect "novel" intrusions, or the ability to operate without a well-defined security policy, they would of course be most valuable if they were applicable to the situation in our more direct scenario as well.

### 6.1 ROC Curve Analysis

Plotting the *detection* rate as a function of the *false alarm* rate, we end up with what is called a ROC (receiver operating characteristic) curve. (For a general introduction to ROC curves, and detection and estimation theory, see Trees [1968]. A shorter introduction that attempts to tie detection and estimation theory to intrusion detection can be found in Axelsson [2000b].)

A few points about ROC curve analysis are worth mentioning here, however. First, the points (0;0) and (1;1) are members of the ROC curve for any intrusion detector. Obviously, if we say that that all events are intrusions, the *detection* rate is 1, but in doing so we will incorrectly

classify all benign activity as intrusive, and consequently we will have a *false alarm* rate of 1 as well.[4] Conversely, the same can be said for the case where the rates are 0. (Classifying all activity as benign will not give us *any* false alarms, but also no detections.) There are general results in detection and estimation theory that state that the *detection* and *false alarm* rates are linked [Trees 1968], although the extent to which these results are applicable in the intrusion detection case is still an open question. Intuitively however, we see that by classifying more and more events as intrusive—in effect relaxing our requirements on what constitutes an intrusion—we will increase our *detection* rate, but also misclassify more of the benign activity, and hence increase our *false alarm* rate.

Note also that we can easily construct a detector with the performance equal to any point along the straight line between (0;0) and (1;1) by making a randomized decision. If we wanted a detector with a 50% false alarm, and detection rate, we would simply say *detection* in half the cases (randomly) and *no detection* in the other. Thus all operational points of sensible detectors should lie strictly above the diagonal. This argument is valid for any two points on the ROC-curve. A randomized detector would then choose between randomly applying the detector represented by the rightmost operating point and the leftmost operating point, the average of the random decisions biased for how close we want to be to one or the other operating points. Because of this the curve between the endpoints should be convex; the ROC-curve cannot contain dips between any two operating points, as that would in effect indicate a *faulty*, nonoptimal detector, since a randomized test would then be better.

For reference, the ROC curve that depicts our scenario laid out in Section 5 (i.e., a required detection rate of 0.7 at a false alarm rate of $1/100{,}000$) is plotted in Figures 2 and 3 as "Assumed ROC." For reasons of clarity, the ROC diagrams do not display the results for larger values of the false alarm rates (i.e., the horizontal axis is truncated), since this would make the scale much too small to discern the regions of interest in the diagrams. In all cases, the plot of the curves continues uneventfully along the straight lines to the (1;1) point.

From the diagrams, we see that the required ROC curve has a very sharp rise from (0;0) since we quickly have to reach acceptable *detection* rate values (0.7) while still keeping the *false alarm* rate under control. Note that we have indicated the possible randomized detectors by plotting the interpolated lines from (0;0) and (1;1) to our required operational point. We have also plotted similar interpolation lines for all other detectors, the results of which we report. Even so, it should be pointed out that we do not seriously advocate the construction of a randomized detector as outlined above, instead the interpolated lines serve only as a sanity check when comparing against a new detector, or when we have varied the parameters

[4]If you call everything with a large red nose a clown, you will spot all the clowns, but also Santa's reindeer, Rudolph, and vice versa.

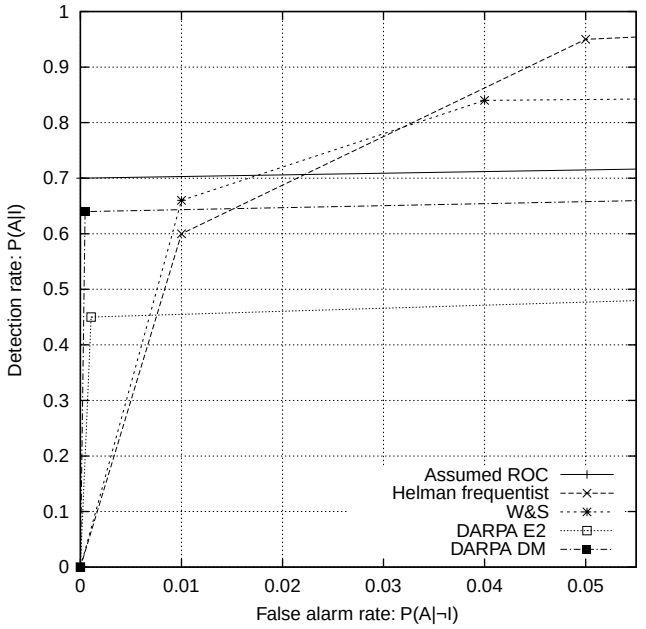


Fig. 2. ROC-curves for the "low performers".

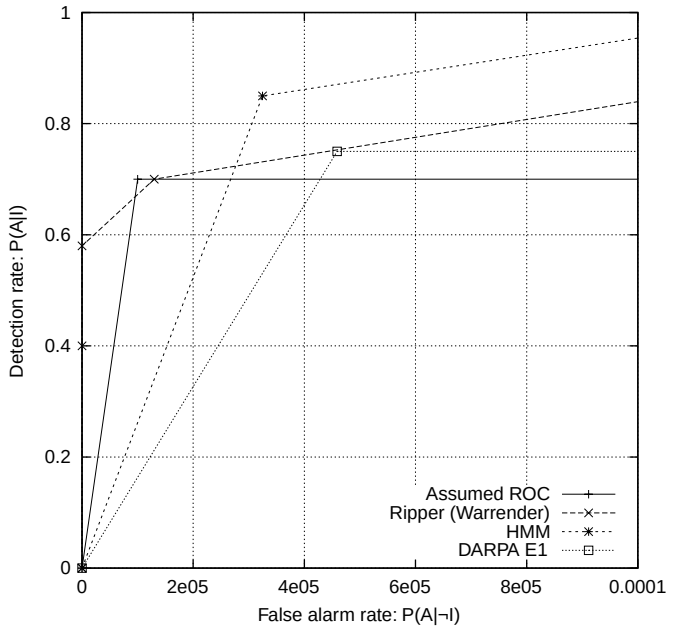


Fig. 3. ROC-curve for the "high performers".

for our detector, resulting in a new operating point. The new operating point *must* lie above the interpolated lines; otherwise, we have not improved on our detector, since a naive randomized detector would outperform it.

### 6.2 Previous Experimental Intrusion Detection Evaluations

As previously mentioned, the literature is not overladen with experimental results from tests of intrusion detection systems. Ideally, we would like several different results from the different classes of intrusion detection systems. Unfortunately, there only exists one report of anomaly detection performance in this regard (with a strong theoretical foundation), [Helman and Liepins 1993]. However, several signature-based detectors have been tested for DARPA by Lincoln Labs [Lippmann et al. 2000] in the by far most ambitious evaluation of intrusion detection systems to date.

Unfortunately, we are not able to evaluate the suitability of this study for our purposes since the data are unavailable to us for independent evaluation because of US export restrictions.

What has been made known about the latter study indicates that it was conducted using a simulated network of workstations, transmitting simulated traffic. This traffic was generated based on real traffic observed on a large US Air Force base, and a large research institute. This lends some credibility to an argument about the generality of the background traffic, but no such argument is made by the authors. Of course, the degree to which the background traffic is representative of the background traffic in the field is a crucial question when it comes to the question of the value of the test as an indicator of false alarm rates during normal usage.

In the test, a number of different attacks were then inserted into the simulated network, including denial of service attacks against the network, and "root" exploits against individual workstations. The experimenters invited several different intrusion detectors to participate in the study. These were all signature-based detectors operating on either network or host data. Even though there is considerable variation in the study (the detection rate varies between approximately 20% to 90% for the best scoring detector for all attacks) we limit the presentation to the best overall scores for the best of the participating detectors; we take "best" here to mean the highest detection rates, coupled with the lowest false alarm rates.

Also not all detectors performed equally well when dealing with all intrusions, and it is a general criticism that in the case of signature-based detection, the designer of the signature can easily trade off detection rate for false alarm rate by varying the generality of the signature. The more general, abstract if you will, it is, the more variations of the same intrusive behaviour it will detect, but at the cost of a higher false alarm rate. It is not known to what extent the DARPA evaluation used variations of the attacks presented to the designers of the intrusion detection systems for training purposes, in the final evaluation. This is an important point in that when such systems are commercialized, it will be impossible to keep the detection signatures secret from the would-be intruders, and the more savvy among them will of course attempt to vary their techniques to evade the intrusion detection system.[5]

[5]Compare with a so-called polymorphic computer virus, that will undergo random semantic preserving code transformations, in order to avoid detection by virus scanning tools.

Furthermore, when the detectors were subjected to previously unknown attacks, their detection rates fell sharply. Their false alarm rates did not see a corresponding increase, but we conjecture that this is because while the attacks in this case were varied between the training data and test data, the background traffic was not. This in turn will favor intrusion detection systems with an overly specific view of what the background traffic consists of; it will not be stressed sufficiently to expose lower false alarm rejection capabilities in a novel, but benign, situation. We would have liked to confirm or reject such a hypothesis, but as mentioned before, the evaluation data are not available to us.

Much more can be said about this evaluation. For an independent and detailed critique of the DARPA evaluation, the reader is directed to McHugh [2000], which raises some of the above questions and many others, in detail.

The second study [Warrender et al. 1999] lists test results for six different intrusion detection methods that have been applied to traces of system calls made into the operating system kernel by nine different privileged applications in a UNIX environment. Most of these traces were obtained from "live" data sources; that is, the systems from which they were collected were production systems. The authors' hypothesis is that short sequences of system calls exhibit patterns that describe normal benign activity, and that different intrusion detection mechanisms can be trained to detect abnormal patterns, and flag these as intrusive. The researchers thus trained the intrusion detection systems using part of the "normal" traffic, and tested their false alarm rate on the remaining "normal" traffic. They then trained the systems on intrusive scenarios, and inserted such intrusions into normal traffic to ascertain the detection rate. The experimental method is thus close to the one described in Sections 4 and 5. This study evaluated as one of the systems the unconventional self-learning detector, RIPPER, described by Lee [1999].

The third study [Helman and Liepins 1993] is a treatise on the fundamental limits of the effectiveness of intrusion detection. The authors construct a model of the intrusive and normal processes and investigate the properties of this model from an anomaly intrusion detection perspective under certain assumptions. Their approach differs from ours in that they do not provide any estimates of the parameters in their model, opting instead to explore the limits of effectiveness when such information is unavailable. Of greatest interest here is their conclusion in which the authors plot experimental data for two implementations, one a frequentist detector that (it is claimed) is close to optimal under the given circumstances, and an earlier tool designed by the authors, Wisdom and Sense [Vaccaro and Liepins 1989]. These tools are interesting in that their outputs are continuous, increasing with decreasing observed frequency of the measured phenomenon. The operator decides when he wants to flag a particular behavior as intrusive by applying a threshold, such that the alarm will be raised when the output signal exceeds that threshold. By varying the threshold the performance point of the detector can be tuned to

meet the requirements of the operating environment. Thus, by raising the threshold we will lower our false alarm rate, but also lower our detection rate, and vice versa. The same general argument is also valid for "Ripper" although it is not an "anomaly" system per se, and the particulars of the implementation are different. Hence these systems begin to trace out the convex ROC curve that is familiar to those accustomed to studying ROC curves of, for example, digital radio communications detectors.

Unfortunately, only one type of anomaly detection system, one that operates with descriptive statistics of the behavior of the subject, is covered. More "sophisticated" detectors, such as neural network-based detectors (such as Debar et al. [1992]), that take time series behavior of the subject into account, are unfortunately not covered.

Lack of space precludes a more detailed presentation of these experiments, and the interested reader is referred to the cited papers where available.

### 6.3 Interpretation of Results

The results from the three cited studies above have been plotted in figures 2 and 3. Where a range of values were given in the original presentation, the best, most "flattering" value was chosen. Furthermore, since not all the work cited provided actual numerical data, some points are based on our interpretation of the presented values. In the case of the DARPA study the results were rescaled to conform to our requirements. (The original DARPA test assumes 66,000 events per day instead of our 100,000 events per day.) Even though it is difficult to express with certainty how many audit records these events consist of, there is some indication that they are variable in size, and perhaps larger than ours. We feel that these values are accurate enough for the purpose of giving the reader an idea of the performance of the systems, in relation to our stated scenario.

The cited work can be roughly divided into two classes depending on the minimum false alarm rate values that are presented, and hence, for clarity, the presentation has been divided into figures, where the first (Figure 2) presents the first class, with larger values for the false alarm rate. These consist of all the anomaly detection results in this study, and the DARPA results "E2" and "DM." In the figure, "Helman frequentist" and "W&S" denote the detection results from Helman and Liepins [1993]. It is interesting, especially in the light of the strong claims made by the authors of this evaluation, to note that all of the presented false alarm rates are at least an order of magnitude larger than the requirements put forth in Section 5. We also put the two DARPA results here, since they are at least an order of magnitude from the top performer (E1) in the DARPA evaluation, and hence would fall to the right of Figure 3.

The second class of detectors, depicted in Figure 3, consists of the average results of Ripper [Lee 1999], a high performance hidden Markov model detector (labeled "HMM" in the figure) tested by Warrander et al. [1999], and the top performer from the DARPA results, listed as E1. Here

the picture is less clear. Warrander et al. report false alarm results close to zero for lower detection rates, with one performance point nearly overlapping our required performance point. The HMM detector is also close to what we would require. It is more difficult to generalize these results, since they are based on one method of data selection, and the authors do not make as strong a claim as those made for the previous set of detectors. The DARPA data from Lippmann et al. [2000], show up as "E1" in Figure 3. It too is close to our required performance. It is unfortunately impossible to give a better name to the systems participating in the DARPA evaluations, or to compare these results with other reported results, since the names of the participating systems have been intentionally withheld in the cited study.

As we can see in the figures above, several systems are between one and three orders of magnitude larger than our false alarm requirement, and some of them not even reaching our 70% target detection rate, at this high false alarm rate. As is evident from Figure 1, this would result in Bayesian detection rates on the order of 0.15 to 0.0015; that is, 1 in 10 alarms to 1 in 1,000 alarms would be correctly indicating an intrusion. Sifting through that many false alarms, especially on the higher end, would of course be anything from discouraging to completely infeasible for the human operator.

We feel a more detailed discussion would be of little additional value, since our model is really quite simple. It only deals with one kind of intrusion, with a fixed unit of measurement. The cited work somewhat departs from such a simple model, since the systems were all tested in an environment with at least two different types of intrusions.

## 7. FUTURE WORK

One sticking point is the basic probabilities on which the previous calculations are based. These probabilities are subjective at present, but future work should include measurement either to attempt to calculate these probabilities from observed frequencies, the *frequentist* approach, or to deduce these probabilities from some model of the intrusive process and the intrusion detection system, the *objectivist* approach. The latter would in turn require real-world observation to formulate realistic parameters for the models.

Furthermore, this discourse treats the intrusion detection problem as a binary decision problem, that is, deciding whether there has been an "intrusion." The work presented does not differentiate between the different kinds of intrusions that can take place, nor does it recognize that different types of intrusions are not equally difficult or easy to detect. Thus on a more detailed level, the intrusion detection problem is not a binary but rather an $n$-valued problem, where in reality we would make binary decisions between $n$ different types of intrusions.

Closely related is the *unit of analysis* problem; that is, how many data does the individual intrusion detection system need to examine before it

can detect the intrusion, or perhaps more important from our perspective, before it can be said to have *missed* the detection of an intrusion. Here we have somewhat skirted the issue, by declaring the unit length to be 10 audit records. Even though we are not alone in treating the problem in this way [Warrender et al. 1999], we believe a more detailed study would define different units of measurement for both different intrusion detection mechanisms, and different types of intrusions.

Another area that needs attention is that of the SSO's capabilities. How does the human-computer interaction take place, and precisely which Bayesian detection rates would an SSO tolerate under what circumstances?

The other parameters discussed in the introduction (*efficiency*, etc.) also need further attention.

## 8. CONCLUSIONS

This article aims to demonstrate that intrusion detection in a realistic setting is perhaps harder than previously thought. This is due to the base-rate fallacy problem, because of which the factor limiting the performance of an intrusion detection system is not the ability to identify behavior correctly as intrusive, but rather *its ability to suppress false alarms*. That is, one should measure the false alarm rate in relation to how many *intrusions* one would expect to detect, not in relation to the maximum number of *possible false alarms*. Thus, a very high standard, less than $1/100{,}000$ per "event" given the stated set of circumstances, will have to be reached for the intrusion detection system to live up to these expectations as far as effectiveness is concerned.

The cited studies of intrusion detector performance that were plotted and compared indicate that anomaly-based methods may have a long way to go before they can reach these standards, since their false alarm rates are several orders of magnitude larger than what we demand. When we come to the case of signature-based detection methods the picture is less clear. Even though the cited work seems to indicate that current signature intrusion detectors can operate close to the required performance point, how well these results generalize in the field is still an open question.

Of course, whether some of the more difficult demands, such as the detection of masqueraders or the detection of novel intrusions, can be met without the use of anomaly-based intrusion detection is still an open question.

Much work still remains before it can be demonstrated that current IDS approaches will be able to live up to real-world expectations of effectiveness. However, we would like to stress that, the present results notwithstanding, an equal amount of work remains before it can be proven that they *cannot* live up to such high standards.

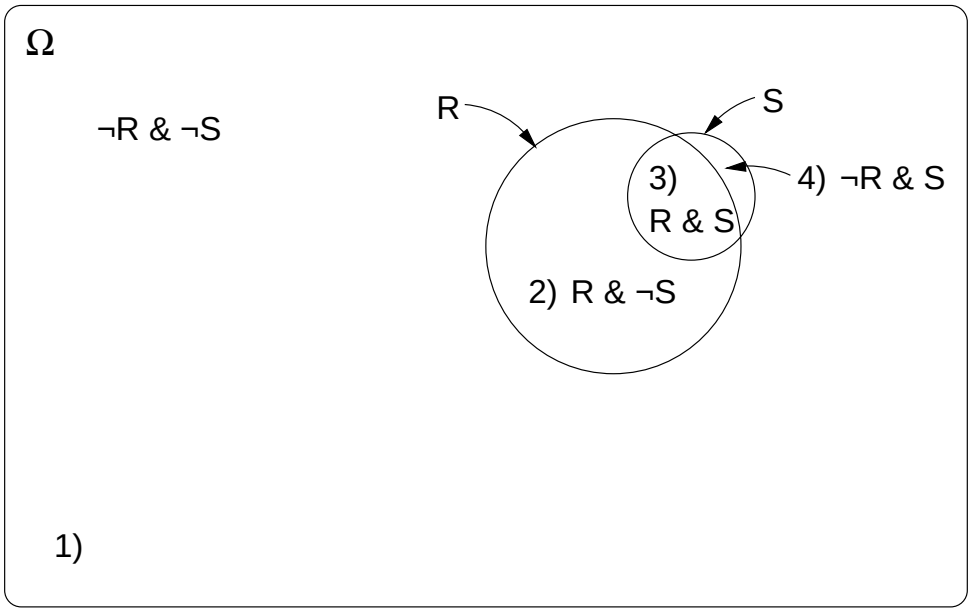


Fig. 4. Venn diagram of medical diagnostic example.

## APPENDIX

### A. VENN DIAGRAM OF THE BASE-RATE FALLACY EXAMPLE

The Venn diagram in Figure 4 depicts the situation in the medical diagnostic example of the base-rate fallacy given earlier.

Although for reasons of clarity the Venn diagram is not to scale, it clearly demonstrates the basis of the base-rate fallacy, that is, that the population in the outcome $S$ is much smaller than that in $\neg S$, and hence, even though $P(R \mid S) = 99\%$ and $P(\neg R \mid \neg S) = 99\%$, the relative sizes of the missing 1% in each case—areas 2) and 4) in the diagram—are very different.

Thus, when we compare the relative sizes of the four numbered areas in the diagram, and interpret them as probability measures, we can state the desired probability, $P(S \mid R)$, that is, "What is the probability that we are in area 3) given that we are inside the $R$-area?" It may be seen that, area 3) is small relative to the entire $R$-area, and hence, the fact that the test is positive does not say much, in absolute terms, about our state of health.

### ACKNOWLEDGMENTS

I would like to thank my colleague, Ulf Lindqvist, for valuable discussion and comments on early drafts of this article. John McHugh and Roy Maxion also provided valuable discussion, comments, and support, especially during the later stages of this work. Ericsson Mobile Data Design AB kindly let me spend time finishing this manuscript. I would also like to thank the anonymous reviewers for their suggestions.

REFERENCES

ANDERSON, J. P. 1980. Computer security threat monitoring and surveillance. 79F26400 26 Feb revised April 15.

AXELSSON, S. 1998. Research in intrusion-detection systems: A survey. 98--17.

AXELSSON, S. 2000. Intrusion-detection systems: A taxonomy and survey. 99-15 (March).

AXELSSON, S. 2000. A preliminary attempt to apply detection and estimation theory to intrusion detection. 00--4 (March).

AXELSSON, S., LINDQVIST, U., GUSTAFSON, U., AND JONSSON, E. 1998. An approach to UNIX security logging. In *Proceedings of the 21st NIST-NCSC National Conference on Information Systems Security* (Crystal City, Arlington, VA, Oct. 5-8). National Institute of Standards and Technology, Gaithersburg, MD, 62–75.

DEATHERAGE, B. H. 1972. Auditory and other sensory forms of information. In *Human Engineering Guide to Equipment Design: Army, Navy, Air Force*, H. Van Cott and R. Kinkade, Eds.

DEBAR, H., BECKER, M., AND SIBONI, D. 1992. A neural network component for an intrusion detection system. In *Proceedings of the ACM/IEEE Symposium on Research in Security and Privacy* (Oakland, CA, May). IEEE Computer Society Press, Los Alamitos, CA, 240–250.

DENNING, D. E. 1987. An intrusion-detection model. *IEEE Trans. Softw. Eng. SE-13*, 2 (Feb.), 222–232.

DENNING, D. E. AND NEUMANN, P. G. 1985. Requirements and model for IDES: A real-time intrusion detection system.

HALME, L. AND KAHN, B. 1988. Building a security monitor with adaptive user work profiles. In *Proceedings of the 11th National Computer Security Conference* (NIST-NCSC, Baltimore, Maryland, Oct.17-20). National Institute of Standards and Technology, Gaithersburg, MD, 000–000.

HELMAN, P. AND LIEPINS, G. 1993. Statistical foundations of audit trail analysis for the detection of computer misuse. *IEEE Trans. Softw. Eng. 19*, 9 (Sept.), 886–901.

LANE, T. AND BRODLEY, C. E. 1999. Temporal sequence learning and data reduction for anomaly detection. *ACM Trans. Inf. Syst. Secur. 2*, 3, 295–331.

LEE, W. 1999. A data mining framework for building intrusion detection models. In *Proceedings of the 1999 IEEE Computer Society Symposium on Research in Security and Privacy* (Berkeley, CA, May). IEEE Computer Society Press, Los Alamitos, CA, 120–132.

LIPPMANN, R. P., FRIED, D., GRAF, I., ET AL. 2000. Evaluating intrusion detection systems: The 1998 DARPA off-line intrusion detection evaluation. In *Proceedings of the DARPA Information Survivability Conference and Exposition* (DISCEX '00, Hilton Head, South Carolina, Jan. 25-27). IEEE Computer Society Press, Los Alamitos, CA, 12–26.

LUNT, T. F. 1988. Automated audit trail analysis and intrusion detection. In *Proceedings of the on 11th National Computer Security Conference* (NIST-NCSC, Baltimore, Maryland, Oct.17-20). National Institute of Standards and Technology, Gaithersburg, MD, 65–73.

MATTHEWS, R. 1996. Base-rate errors and rain forecasts. *Nature 382*, 6594, 766.

MATTHEWS, R. 1997. Decision-theoretic limits on earthquake prediction. *Geophys. J. Int. 131*, 3 (Dec.), 526–529.

MCHUGH, J. 2000. Testing intrusion detection systems: A critique of the 1998 and 1999 Lincoln Laboratory evaluations. *ACM Trans. Inf. Syst. Secur. 3*.

NYGREN, E. 1994. Moderna tider: teknikutveckling inom medicinsk service.

PIERCE, G. M. 1943. Destruction by demolition, incendiaries and sabotage: Field training manual, Fleet Marine Force, US Marine Corps.

RASMUSSEN, J. 1986. *Information Processing and Human-Machine Interaction: An Approach to Cognitive Engineering*. North-Holland Publishing Co., Amsterdam, The Netherlands.

RUSSELL, S. J. AND NORVIG, P. 1995. *Artificial intelligence: a modern approach*. Prentice-Hall series in artificial intelligence. Prentice-Hall, Inc., Upper Saddle River, NJ.

SEBRING, M. M., SHELLHOUSE, E., HANNA, M. E., AND WHITEHURST, R. A. 1988. Expert systems in intrusion detection: A case study. In *Proceedings of the 11th National Computer Security Conference* (NIST-NCSC, Baltimore, Maryland, Oct.17-20). National Institute of Standards and Technology, Gaithersburg, MD, 74–81.

TREES, H. L. V. 1968. *Detection, Estimation, and Modulation Theory, Part I: Detection, Estimation, and Linear Modulation Theory*. John Wiley and Sons, Inc., New York, NY.

U.S. DEPARTMENT OF DEFENSE. 1985. Trusted Computer System Evaluation Criteria. DoD 5200.28-STD.

VACCARO, H. S. AND LIEPINS, G. E. 1989. Detection of anomalous computer session activity. In *Proceedings of the IEEE Symposium on Research in Security and Privacy* (Oakland, CA, May 1-3). IEEE Computer Society Press, Los Alamitos, CA, 280–289.

WARRENDER, C., FORREST, S., AND PERLMUTTER, B. 1999. Detecting intrusions using system calls: Alternative data models. In *Proceedings of the 1999 IEEE Computer Society Symposium on Research in Security and Privacy* (Berkeley, CA, May). IEEE Computer Society Press, Los Alamitos, CA, 133–145.

WICKENS, C. 1992. *Engineering Psychology and Human Performance*. 2nd ed. HarperCollins Publishers, New York, NY.

Received: January 2000; revised: May 2000; accepted: May 2000