The Base-Rate Fallacy and its Implications for the Difficulty of Intrusion Detection

Stefan Axelsson

Department of Computer Engineering
Chalmers University of Technology
Göteborg, Sweden
Email: *sax@ce.chalmers.se*

Abstract

Many different demands can be made of intrusion detection systems. An important requirement is that it be *effective* i.e. that it should detect a substantial percentage of intrusions into the supervised system, while still keeping the *false alarm* rate at an acceptable level.

This paper aims to demonstrate 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 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.

1 Introduction

Many demands can be made of an intrusion detection system (IDS for short) such as *effectiveness*, *efficiency*, *ease of use*, *security*, *inter-operability*, *transparency* etc. Although much research has been done in the field in the past ten years, the theoretical limits of many of these parameters have not been studied to any significant degree. The aim of this paper is to discuss 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 Problems in Intrusion Detection

The field of automated computer intrusion detection—intrusion detection for short—is currently some nineteen years old [1], with interest gathering pace in the past ten 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, etc.

Early work (see [1, 4, 5, 18]) 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. We add the requirement that the system be self-learning for it to qualify as an anomaly detection system.

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.

Policy detection Our term for the detection strategy of deciding in advance what type of behaviour is undesirable, and through the use of a default permit or default deny policy, detecting intrusions. The *default permit* case is often referred to as *signature based detection* or *misuse detection*, while we term the few published instances of *default deny* systems *specification-based intrusion detection* after the first such system [8].

Policy-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 [6, 12] 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.

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

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
CCS '99 11/99 Singapore
© 1999 ACM 1-58113-148-8/99/0010 $5.00

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, etc?

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 realistically be expected to cope with, and under what circumstances is he likely to ignore an alarm? (It has long been known in security circles that ordinary electronic alarm systems should be circumvented during normal operation of the facility, when supervisory staff are more likely to be lax because they are accustomed to false alarms [16]).

Security When ever 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?

Inter-Operability As the number of different intrusion detection systems increase, to what degree can they inter-operate and how do we ensure this?

Transparency How intrusive is the fielding of the intrusion detection system to the organisation employing it? How many resources will it consume in terms of manpower, etc?

While interest is being shown in some of these issues, with a few notable exceptions—mainly [7]—they remain largely unaddressed by the research community. This is perhaps not surprising, since many of these questions are difficult to formulate and answer. For a detailed and thorough survey of research into intrusion detection systems to date see [2].

This paper 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 3 gives a description of the base-rate fallacy, section 4 continues with an application of the base-rate fallacy to the intrusion detection problem, given a set of reasonable assumptions, section 5 describes the impact the previous results would have on intrusion detection systems, section 6 considers future work, with section 7 concluding the paper. Appendix A reproduces a base-rate fallacy example in diagram form.

3 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, i.e. 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 equation (2):

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

Combining equations (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, i.e. 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, i.e. only 1 in 10000 people have this ailment. What, given this information, is the probability of you 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 $\neg S$, i.e. *not* S, denote healthy. Likewise, let P denote a positive test result and $\neg P$ denote a negative test result. Restating the information above; given: $P(P|S) = 0.99$, $P(\neg P|\neg S) = 0.99$, and $P(S) = 1/10000$, what is the probability $P(S|P)$?

A direct application of equation (3) above gives:

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

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

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

That is, that 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

[1] The idea behind this approach stems from [13, 14].

[2] This example hinted at in [17].

population with the disease. (For a graphical representation, in the form of a Venn diagram, depicting the different outcomes, turn to Appendix A). 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 solving such problems of probability—that is aptly named "the base-rate fallacy."

4 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.

4.1 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 in the order of 1,000,000 audit records per day with some form of "C2" compliant logging in effect, 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.

The figures above are based on [11], and while the results of that study would seem to indicate that indeed low false alarm rates can be attained, one can raise the objection that since the developers of the tested systems had prior access to "training" data that was very similar to the later evaluation data, the systems' false alarm suppression capability was not sufficiently tested. Another paper that discusses the effectiveness of intrusion detection is [15]. Unfortunately it is not applicable here.

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.

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 [3], we can estimate that ten audit records would be affected in the average intrusion.

4.2 Calculation of Bayesian detection rates

Let I and $\neg I$ denote *intrusive*, and *non-intrusive* behaviour 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. The probability $P(A|I)$, i.e. that quantity that we can obtain when testing our detector against a set of scenarios we know represent intrusive behaviour.

False alarm rate 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:

$$P(\neg A|I) = 1 - P(A|I); P(\neg A|\neg I) = 1 - P(A|\neg I) \quad (6)$$

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:

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

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

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

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:

$$\begin{aligned} P(I) &= 1 \Big/ \frac{1 \cdot 10^6}{2 \cdot 10} = 2 \cdot 10^{-5}; \\ P(\neg I) &= 1 - P(I) = 0.99998 \end{aligned} \quad (9)$$

Inserting equation (9) into equation (7):

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

Studying equation (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 makes it clear that we have an overwhelming number of non-events (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 $P(A|I)$ (including the "best" case $P(A|I) = 1$), as a function of $P(A|\neg I)$ (see figure 1 on the following page). It should be noted that both axes are logarithmic.

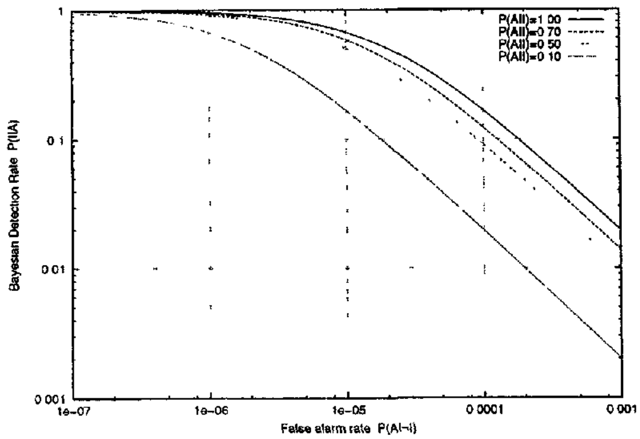

Figure 1: Plot of Bayesian detection rate versus false alarm rate

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%, i.e. 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 fifty-fifty. 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 in total of 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 5.

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 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 equation (8) gives:

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

A quick glance at the resulting equation (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 $P(A|\neg I)$ under discussion here, regardless of the value of $P(A|I)$.

[3] Another way of calculating that differs from equation (10) is of course to realise that 100 false alarms and only a maximum of 2 possible valid alarms gives: $\frac{2}{2+100} \approx 2\%$.

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 non-intrusions, combined with the fact that we do not have many intrusions to begin with. Truly a question of finding a needle in a haystack.

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!

5 Impact on Intrusion Detection Systems

As stated in the introduction, approaches to intrusion detection can be divided into two major groups, *policy*-based, and *anomaly*-based. 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. This section will compare these requirements with reported results on the effectiveness of intrusion detection systems.

It can be argued that this reasoning applies mainly to policy-based intrusion detection. In some cases anomaly-based detection tries not to detect intrusions per se, but rather to differentiate between two different subjects, flagging anomalous behaviour in the hopes that it is indicative of a stolen user identity for instance, see for example [9], which even though it reports performance figures, is not directly applicable here. However, we think the previous 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.

5.1 ROC curve analysis

There are general results in detection and estimation theory that state that the *detection* and *false alarm* rates are linked [20], though the extent to which they are applicable here is still an open question. Obviously, if the *detection* rate is 1, saying that all events are intrusions, we will have a *false alarm* rate of 1 as well, and conversely the same can be said for the case where the rates are 0.[4] Intuitively, 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.

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 [20].) We have already stated that the points (0; 0) and (1; 1) are members of the ROC curve for any intrusion detector. Furthermore, the curve between these points is convex; were it concave, we would do better to reverse our decision.

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

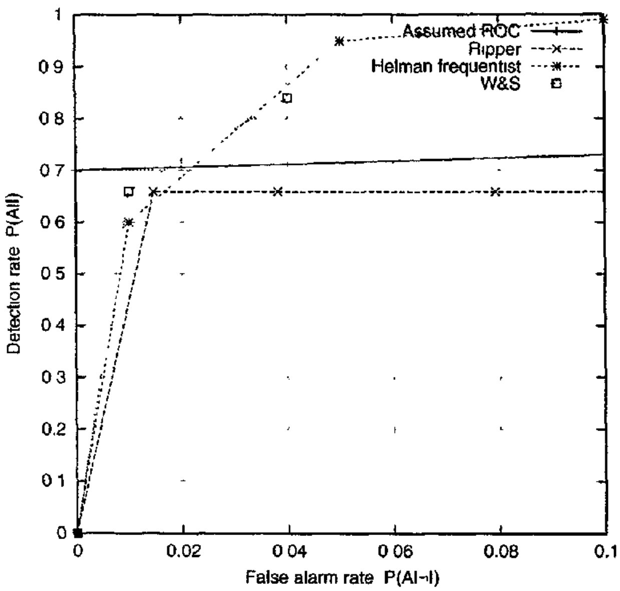

Figure 2: ROC-curves for the second and third studies

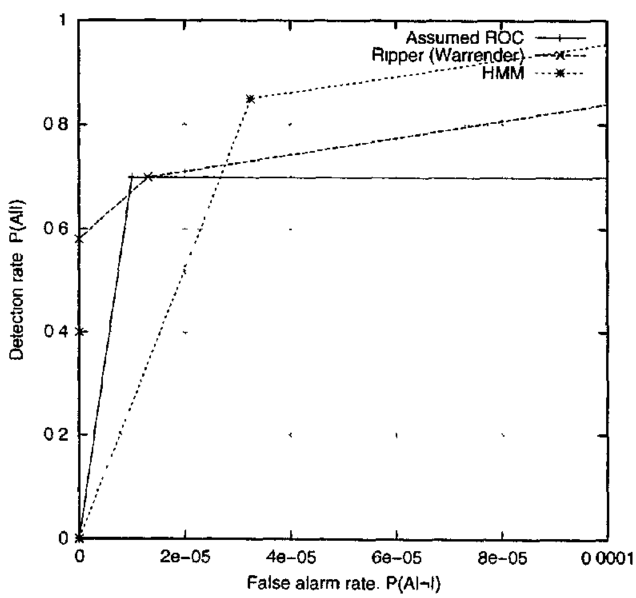

Figure 3: ROC-curve for the first study

Nor can it contain any dips, as that would in effect indicate a faulty, non-optimal detector, since a randomised test would then be better. See "Assumed ROC" curve in figures 2 and 3 for the ROC curve that depicts our previous example.

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.

5.2 Previous experimental intrusion detection evaluations

As previously mentioned, the literature is not overladen with experimental results from tests of intrusion detection systems. One recent evaluation performed by DARPA exists [11], but no comprehensive results have been published, and the data is unavailable for independent evaluation because of U.S. export restrictions. We have chosen two recent publications [10, 21] on the effectiveness of several policy-based methods, and one theoretically advanced treatise on anomaly-based methods [7], on which to base our evaluation.

The first study [21] 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, i.e. 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 3 and 4.

The second study [10], reports results from one of the tools entered into the DARPA evaluation. The DARPA data is supposedly modelling a realistic situation, having been synthesized from several months' long measurements on two large computer sites. The author claims that this tool faired well in competition with the other systems so evaluated[5]. Interestingly the same tool has been applied (in a different manner) to the data generated by the first study above, which makes for an interesting comparison. Surprisingly, the independent evaluation reports better results—by as much as several orders of magnitude—than the author of the tool himself reports.

The third study [7] is a treatise on the fundamental limits of the effectiveness of intrusion detection. The authors constructs a model of the intrusive and normal process 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 & Sense [19].

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

The results from the three 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" if you will—value was chosen. Furthermore, since not all the work reffered to provided actual numerical data, some points are based on our interpretation of the presented values. We feel

[5] In the words of the author "We can see from the figure that our detection model has the best overall performance..."

that these are accurate enough for the purpose of giving the reader an idea of the performance of the systems.

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. In the figure, "Ripper" denotes the original author's overall DARPA results, "Helman frequentist," and "W&S" denote the anomaly detection results. It is interesting, especially in the light of the strong claims made by the authors of these evaluations, to note that all of the presented false alarm rates are several orders of magnitude larger than the requirements put forth in section 4.

The second class of detectors, depicted in figure 3, consists of the average results of Ripper, and a high performance Hidden Markov Model detector (labeled "HMM" in the figure) tested by Warrander et. al. Here the picture is less clear. In these experiments the specific application of Ripper performs admirably. The authors 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.

6 Future Work

One sticking point is the basic probabilities that the previous calculations are based on. 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, i.e. that of deciding whether there has been an "intrusion" or not. The work presented does not differentiate between the different kinds of intrusions that can take place, and nor does it recognise 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.

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 for example?

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

7 Conclusions

This paper 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 behaviour correctly as intrusive, but rather *its ability to suppress false alarms.* 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 misuse-based detection methods the picture is less clear. One detector performs well in one study—and meets our expectations—but is much less convincing in another, where it performs on a par with the anomaly-based methods studied. Whether some of the more difficult demands, such as the detection 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.

8 Acknowledgements

I would like to thank my colleague Ulf Lindqvist and my supervisor Erland Jonsson for valuable insights. I would also like to thank the anonymous reviewers for their comments and suggestions.

This work was funded by the Swedish National Board for Industrial and Technical Development (NUTEK) under project P10435.

References

[1] J. P. Anderson. Computer security threat monitoring and surveillance. Technical Report Contract 79F26400, James P. Anderson Co., Box 42, Fort Washington, PA, 19034, USA, Feb. 26, revised Apr. 15, 1980.

[2] S. Axelsson. Research in Intrusion-Detection systems: A Survey. Technical Report 98–17, Department of Computer Engineering Chalmers University of Technology, SE-412 96 Göteborg, Sweden, Dec. 1998. URL: http://www.ce.chalmers.se/staff/sax.

[3] S. Axelsson, U. Lindqvist, U. Gustafson, and E. Jonsson. An approach to UNIX security logging. In *Proceedings of the 21st National Information Systems Security Conference*, pages 62–75, Crystal City, Arlington, VA, USA, Oct. 5–8, 1998. NIST.

[4] D. E. Denning. An intrusion-detection model. *IEEE Transactions on Software Engineering*, Vol. SE-13(No. 2):222–232, Feb. 1987.

[5] D. E. Denning and P. G. Neumann. Requirements and model for IDES—A real-time intrusion detection system. Technical report, Computer Science Laboratory, SRI International, Menlo Park, CA, USA, 1985.

[6] L. Halme and B. Kahn. Building a security monitor with adaptive user work profiles. In *Proceedings of the 11th National Computer Security Conference*, Washington DC, Oct. 1988.

[7] P. Helman and G. Liepins. Statistical foundations of audit trail analysis for the detection of computer misuse. *IEEE Transactions on Software Engineering*, 19(9):886–901, Sept. 1993.

[8] C. Ko, M. Ruschitzka, and K. Levitt. Execution monitoring of security-critical programs in distributed systems: A specification-based approach. In *Proceedings of the 1997 IEEE Symposium on Security and Privacy*, pages 175–187, Oakland, CA, USA, May 1997.

[9] T. Lane and C. E. Brodie. Temporal sequence learning and data reduction for anomaly detection. In *5th ACM Conference on Computer & Communications Security*, pages 150–158, San Francisco, California, USA, Nov. 3–5, 1998.

[10] W. Lee. A data mining framework for building intrusion detection models. In *IEEE Symposium on Security and Privacy*, pages 120–132, Berkeley, California, May 1999.

[11] R. P. Lippmann, I. Graf, S. L. Garfinkel, A. S. Gorton, K. R. Kendall, D. J. McClung, D. J. Weber, S. E. Webster, D. Wyschogrod, and M. A. Zissman. The 1998 DARPA/AFRL off-line intrusion detection evaluation. Presented to The First Intl. Workshop on Recent Advances in Intrusion Detection (RAID-98), Lovain-la-Neuve, Belgium, *No printed proceedings*, Sept. 14–16, 1998.

[12] T. F. Lunt. Automated audit trail analysis and intrusion detection: A survey. In *Proceedings of the 11th National Computer Security Conference*, pages 65–73, Baltimore, Maryland, Oct. 17–20 1988. NIST.

[13] R. Matthews. Base-rate errors and rain forecasts. *Nature*, 382(6594):766, Aug. 29 1996.

[14] R. Matthews. Decision-theoretic limits on earthquake prediction. *Geophys. J. Int.*, 131(3):526–529, Dec. 1997.

[15] R. A. Maxion. Measuring intrusion-detection systems. Presented to The First Intl. Workshop on Recent Advances in Intrusion Detection (RAID-98), Lovain-la-Neuve, Belgium, *No printed proceedings*, Sept. 14–16, 1998.

[16] G. McGuire Pierce. Destruction by demolition, incendiaries and sabotage. Field training manual, Fleet Marine Force, US Marine Corps, 1943–1948. Reprinted: Paladin Press, PO 1307, Boulder CO, USA.

[17] S. J. Russel and P. Norvig. *Artificial Intelligence—A Modern Approach*, chapter 14, pages 426–435. Prentice Hall Series in Artificial Intelligence. Prentice Hall International, Inc., London, UK, first edition, 1995. Exercise 14.3.

[18] M. M. Sebring, E. Shellhouse, M. E. Hanna, and R. A. Whitehurst. Expert systems in intrusion detection: A case study. In *Proceedings of the 11th National Computer Security Conference*, pages 74–81, Baltimore, Maryland, Oct. 17–20, 1988. NIST.

[19] H. S. Vaccaro and G. E. Liepins. Detection of anomalous computer session activity. In *Proceedings of the 1989 IEEE Symposium on Security and Privacy*, pages 280–289, Oakland, California, May 1–3, 1989.

[20] H. L. Van Trees. *Detection, Estimation, and Modulation Theory, Part I, Detection, Estimation, and Linear Modulation Theory.* John Wiley and Sons, Inc., 1968.

[21] C. Warrender, S. Forrest, and B. Perlmutter. Detecting intrusions using system calls: Alternative data models. In *IEEE Symposium on Security and Privacy*, pages 133–145, Berkeley, California, May 1999.

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.

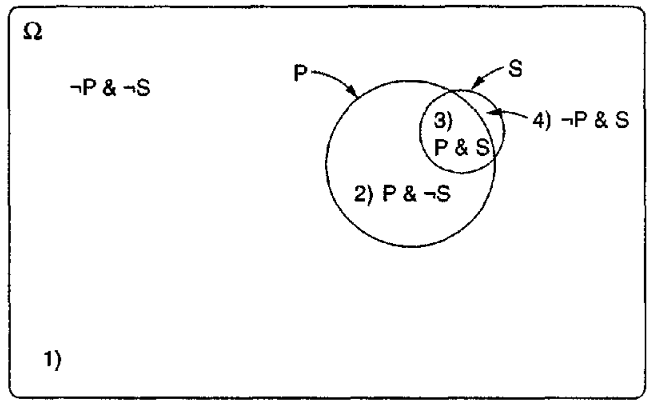

Figure 4: Venn diagram of medical diagnostic example

Although for reasons of clarity the Venn diagram is not to scale it clearly demonstrates the basis of the base-rate fallacy, i.e. that the population in the outcome S is much smaller than that in $\neg S$ and hence, even though $P(P|S) = 99\%$ and $P(\neg P|\neg S) = 99\%$, 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, $P(S|P)$—i.e. "What is the probability that we are in area 3) given that we are inside the P-area?" It may be seen that, area 3) is small relative to the entire P-area, and hence, the fact that the test is positive does not say much, in absolute terms, about our state of health.