On a Difficulty of Intrusion Detection*

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

Aug 09, 1999

Abstract

Research in automated computer security intrusion detection, intrusion detection for short, is maturing. Several difficulties remain to be solved before intrusion detection systems become commonplace as part of real-world security solutions.

One such difficulty regards the subject of *effectiveness*, how successful the intrusion detection system is at actually detecting intrusions with a high degree of certainty. With this as its starting point, this paper discusses the "base-rate fallacy" and how it influences the relative success of an intrusion detection system, under a set of reasonable circumstances. The conclusion is reached that the false-alarm rate quickly becomes a limiting factor.

1 Introduction

Many requirements can be placed on an intrusion detection system (IDS for short) such as *effectiveness*, *efficiency*, *ease of use*, *security*, *interoperability*, *transparency* etc., etc. Although much research has gone into 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 provide a discourse on one serious problem with regard to one of these parameters; *effectiveness*, especially how the base-rate fallacy affects the operational effectiveness of any intrusion detection system.

2 Problems in Intrusion Detection

The field of automated computer security intrusion detection—intrusion detection for short—is currently some nineteen years old. The seminal paper that is most often mentioned is James P. Anderson's technical report [And80], where he states in reference to one class of intruders, the *masquerader*, that:

> Masquerade is interesting in that it is by definition extra use of the system by the unauthorised user. As such it should be possible to detect instances of such use by analysis of audit trail records to determine:
>
> a. Use outside of normal time
> b. Abnormal frequency of use
> c. Abnormal volume of data reference
> d. Abnormal patterns of reference to programs or data

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

As will be discussed in the subsequent section, the operative word is “abnormal” which implies that there is some notion of what “normal” is for a given user.

Later work (See [DN85, Den87, SSHW88]) expanded on these ideas, and 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, and

Policy detection Our term for the strategy of deciding in advance what type of behaviour is undesirable, and through the use of a default permit, or deny, policy, detecting intrusions.

About the same time it was suggested in [HK88, Lun88] that the two main methods ought to be combined to provide a complete intrusion detection system, and that the resulting system should be made autonomous enough to be trusted to respond to detected intrusions unsupervised. The authors recognised that much research remained to be done before that goal could be attained.[1]

Presently, the larger questions regarding intrusion detection remain largely unanswered. Important questions include, but is by no means limited to:

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, aka false alarms?

Efficiency What is the run time efficiency of the intrusion detection system, how much computing resources and 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 the user that is not a security expert, and can that user add new intrusion scenarios to the system? An important part of *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 thought to be able to cope with, under what circumstances is he likely to ignore an alarm[2] etc., etc?

Security When more and more intrusion detection systems are fielded, one would expect more and more attacks directed at the intrusion detection system itself, to circumvent, 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, and how do we make them, interoperate?

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

While there is current interest in some of these issues, they remain largely unaddressed by the research community. This is perhaps not surprising, since, in our minds at least, many of these questions are difficult to formulate and answer. For detailed, and thorough survey of research in intrusion detection to date see [Axe98].

This paper concerns itself with one of the questions above, namely that of *effectiveness*. Especially, how the base-rate fallacy affects the required performance of the intrusion detection system in regard to false alarm rejection etc.

The remainder of this paper is structured as follows: section 3 is a description of the base-rate fallacy, section 4 continues with an application of the base-rate fallacy on 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 remarks on proposed future work, with section 7 concluding the paper. Appendix A diagrams a base-rate fallacy example.

[1] We would like to add that much research *still* remains to be done before this goal can be attained.

[2] It’s been long known in security circles that ordinary electronic alarm systems should be circumvented in daytime, during normal operation, when the supervisory staff is more likely to be lax due to being accustomed to false alarms [MPca].

3 The Base-Rate Fallacy

The base-rate fallacy[3] is one of the cornerstones of Bayesian statistics, as it stems directly from Bayes' famous theorem:

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

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.[4] Suppose that your physician performs a test that is 99% accurate, i.e. when the test was administered to a test population all of which 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 physician to learn of 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 the above information, is the probability of you having the disease?[5]

Let us start by naming the different outcomes. Let S denote sick, and *not* S i.e. $\neg S$ 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 $1/100$, due to the fact that 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, turn to Appendix A). This result often surprises people—we were no exception—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 in probability—that is aptly named "the base-rate fallacy."

[3] The idea behind this approach stems from [Mat96, Mat97].

[4] This example hinted at in [RN95].

[5] The reader is encouraged to make a quick "guesstimate" of the answer, at this point.

4 The Base-Rate Fallacy in Intrusion Detection

In order to apply the above reasoning to the computer intrusion detection case 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's for the sake of further argument hypothesize about 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 1000,000 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 of real incidences of attempted computer security intrusions, this does not seem to be an unreasonable number.

The figures above are based on [LGG+98], and while the results of that study would seem to indicate that indeed low false alarm rates can be attained, one could 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 false alarm suppression capability of the systems was not sufficiently tested. Another paper that discusses the effectiveness of intrusion detection is [Max98]. Unfortunately it is not applicable here.

Furthermore assume, that at this installation we do not have the man power to have more than one site security officer—SSO for short—which probably has other duties also, 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 that it affects a few more than that, on average. We have previously made a study that concerns itself with the trails that SunOS intrusions leave in the audit trail [ALGJ98], and from that data we can make an estimate; say ten audit records affected for the average intrusion.

4.2 Calculation of Bayesian detection rate

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. Working backwards from the above set of assumptions we can obtain the required values of the:

True Positive rate Or *detection* 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 Positive rate The probability $P(A|\neg I)$, i.e. the *false alarm* rate, obtained in the same manner as above.

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 indicate an intrusion (Henceforth called the *Bayesian detection rate*), and
- $P(\neg I|\neg A)$—that the absence of an alarm signify 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 $P(\neg I|\neg A)$:

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

The assumptions above gives us a value for the rate of incidence of the actual number of intrusion in our system, and its dual (10 audit records per intrusion, two intrusions per day, and 1000,000 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 equation (9) into equation (7):

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

Studying equation (10) we see the base-rate fallacy clearly. This should by now come as no surprise to the reader, since the assumptions made about our system makes it clear that we have an overwhelming amount on non-events (benign activity) in our audit trail, and only a few events (intrusions) of interest. Thus, the factor governing the *detection* rate ($2 \cdot 10^{-5}$) is completely overwhelmed 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 gives the most beneficial outcome as far as the *false alarm* rate is concerned. While these values would be desirable accomplishments 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. Please, note that both axes are logarithmic.

It becomes clear from studying the plot in figure 1 that indeed, even for the unrealistically high *detection* rate 1.0 , we have to have a very low (on the order of $1 \cdot 10^{-5}$) *false alarm* rate, 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's the author's belief that a low Bayesian detection rate would quickly "teach" the SSO to safely ignore *all* alarms, even though their absolute numbers would theoretically have allowed complete investigation of all alarms. This becomes especially true as the system scales; a 50% false alarm rate, with a 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 the following section (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 $1 \cdot 10^{-5}$; resulting in $1 \cdot 10^6/10 = 1 \cdot 10^5$ "events" per day, which in turn results in on average $1 \cdot 10^{-5} \cdot 1 \cdot 10^5 = 1$ *false alarm* per day. 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%,[6] by which time no-one will bother to care when the alarm goes off.

Substituting (6) and (9) in equation (8):

[6] Another way of calculating that than from equation (10) is of course by realising that 100 false alarms and only a maximum of two possible valid alarms gives: $\frac{2}{2+100} \approx 2\%$.

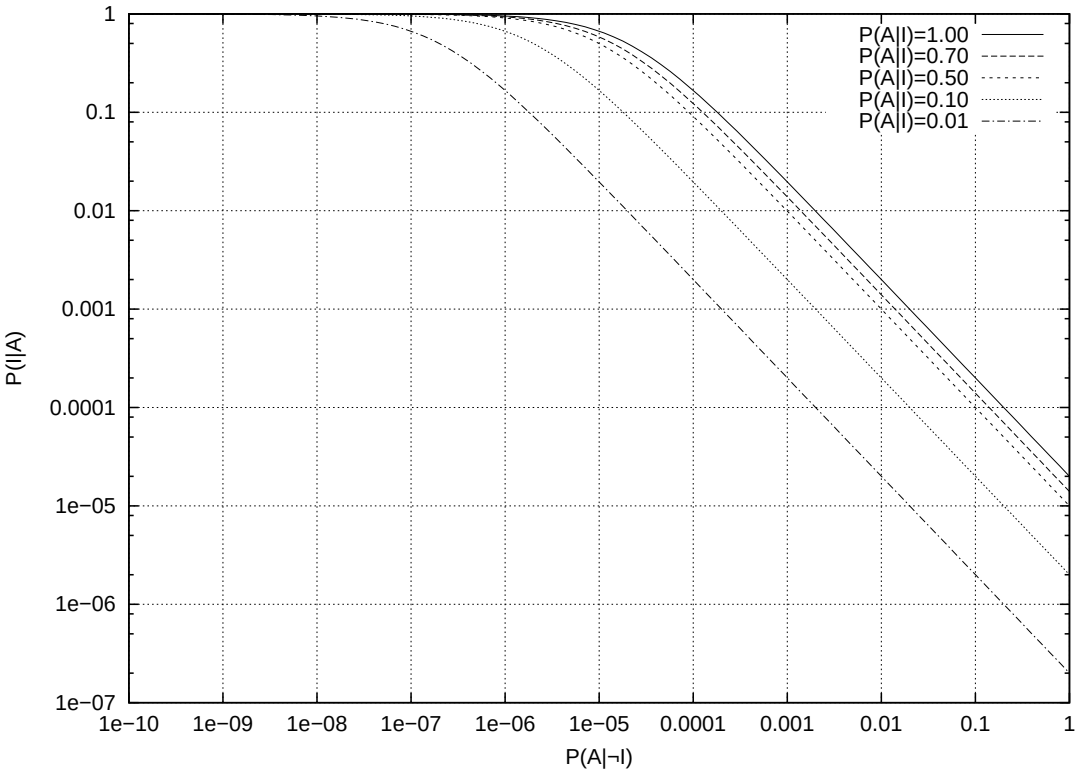

Figure 1: Plot of $P(I|A)$

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

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 $P(A|\neg I)$ we are talking about here, regardless of the value of $P(A|I)$.

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 non-intrusions, combined with the fact that we don't have many intrusions to begin with. Truly a problem of finding a needle in a haystack.

The author does not see how the situation behind the base-rate fallacy problem would change for the better in the 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.[7]

5 Impact on the Different Types of 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 to place on intrusion detection systems in order to make them useful in the stated scenario.

[7] 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 level enough 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…

It could be argued that the above reasoning applies mainly to policy based intrusion detection. In some cases Anomaly based detection tries not to detection intrusions per se, but rather to differentiate between two different subjects, flagging anomalous behaviour, in the hope that it would be indicative of e.g. a stolen user identity. 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.

In order to pass sound judgment on the effectiveness of an anomaly based intrusion detection system, we also have to have a very well founded hypotheses about what constitutes "normal" behaviour for the observed system. We know of only one attempt at such an evaluation in conjunction with the presentation of an anomaly based intrusion detection method: [LB98].

There are general results in detection and estimation theory that state that the *detection* and *false alarm* rate are linked [VT68]. Obviously, if the *detection* rate is 1 i.e. 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.[8] 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. Unfortunately, to apply these results to the current situation we need to have a firm grasp—in the form of a statistical model—of, what constitutes "normal" or "background" traffic.

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, detection and estimation theory, see [VT68]). We have already stated that the points $(0;0)$ and $(1;1)$ are members of the ROC curve for any intrusion detector. Furthermore between these points the curve is convex, were it concave, we would be better off to reverse our decision, and it cannot contain any dips—that would in effect indicate a faulty, non-optimal detector, since a randomised test would then be better. See figure 2 for the ROC curve of our previous example.
We see that our ROC curve has a very sharp rise from $(0;0)$ since we quickly will have to reach acceptable *detection* rate values (0.7) while still keeping the *false alarm* rate at bay. It is doubtful if even policy detection, the type of detection often thought to be the most resilient to *false alarms* can reach as low values as $1 \cdot 10^{-5}$ i.e. $1/100,000$ while still keeping the *detection* rate as high as 0.5–0.7 or above.

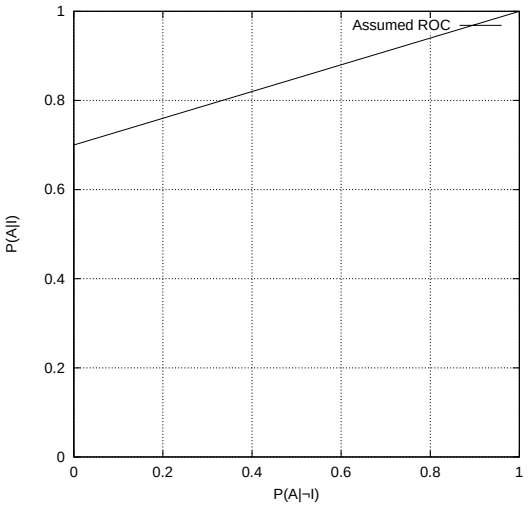

Figure 2: Plot of $P(A|I)$ as a function of $P(A|\neg I)$

To reach such levels it is imperative that the designer of intrusion detection systems do not introduce some policy element that has even a remote chance of triggering in the face of benign activity—perhaps not even known at the time of making the policy—lest the system will fall prey to too low a Bayesian detection rate. Note that this also includes changes over time, that a policy based system would be incapable of adapting to. This situation could well arise in some of the commercial systems we have seen, which contain ad-hoc patterns of intrusion signatures that would not be appropriate in the face of new quite probable traffic. If the designer was to make any such "pattern" more general to catch more variations that would immediately, following the discussion above, result in an increased risk of false alarms.

[8] 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.

6 Future Work

A difficult point is the basic probabilities that the previous calculations are based on. These probabilities are presently subjective, but future work must include measurement to either attempt to calculate these probabilities from observed frequencies—the *frequentist* approach—or the deduction of these probabilities from some model of the intrusive process, and the intrusion detection system—taking the *objectivist* approach.

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 could take place, and that the detection mechanism could very well cross-respond to any one of them in an undesired fashion. Thus on a more detailed level, the intrusion detection problem is not a binary but an n-valued problem.

With observed or otherwise soundly founded probabilities one would calculate the Bayesian and other properties, and even construct an optimal detector for the intrusion detection problem. It would then be possible to state under which circumstances the intrusion detection problem would not only be difficult, but even impossible.

Another problem is that in order to employ soundly founded results in information, and communication theory, we need knowledge of the distributions of the studied features, both when the system is operating undisturbed, and when the system is under attack. Armed with such information, one could make—more or less—absolute predictions of the effectiveness of the detection, and one could even decide optimally the level of detection thresholds etc, under the given assumptions of the distributions.[9]

Another area that needs attention from the perspective of this paper, is that of the capabilities of the SSO. How does the human-computer interaction take place, and precisely what Bayesian detection rates would an SSO tolerate under what circumstances etc.

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

7 Conclusions

This paper has aimed to demonstrate that intrusion detection in a realistic setting is perhaps harder than previously thought. This is due to the base-rate fallacy problem, and because of it, the factor limiting the performance of an intrusion detection system is not the ability to correctly identify behaviour 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, from an *effectiveness* standpoint. 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.

[9] See [VT68] for an introduction to the field.

References

[ALGJ98] Stefan Axelsson, Ulf Lindqvist, Ulf Gustafson, and Erland 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, October 5–8 1998. NIST, National Institute of Standards and Technology/National Computer Security Center.

[And80] James P. Anderson. Computer security threat monitoring and surveillance. Technical Report Contract 79F26400, James P. Anderson Co., Box 42, Fort Washington, PA, 19034, USA, Februrary 26, revised April 15 1980.

[Axe98] Stefan Axelsson. Research in Intrusion-Detection systems: A Survey. Technical Report 98–17, Dept. of Computer Eng. Chalmers Univ. of Tech, SE-412 96 Göteborg, Sweden, December 1998. URL: http://www.ce.chalmers.se/staff/sax.

[Den87] Dorothy E. Denning. An intrusion-detection model. *IEEE Transactions on Software Engineering*, Vol. SE-13(No. 2):222–232, February 1987.

[DN85] Dorothy E. Denning and Peter 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.

[HK88] 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, October 1988.

[LB98] Terran Lane and Carla E. Brodie. Temporal sequence learning and data reduction for anomaly detection. In *5th ACM Conference on Computer & Communications Security*, pages 150–158, San Fransico, California, USA, November 3–5 1998.

[LGG+98] Richard P. Lippmann, Isaac 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*, 14–16 September 1998.

[Lun88] Teresa 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, 17–20 October 1988. National Institute of Standards and Technology/National Computer Security Center.

[Mat96] Robert Matthews. Base-rate errors and rain forecasts. *Nature*, 382(6594):766, 29 August 1996.

[Mat97] Robert Matthews. Decision-theoretic limits on earthquake prediction. *Geophys. J. Int.*, 131(3):526–529, December 1997.

[Max98] Roy 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*, 14–16 September 1998.

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

[RN95] Stuart J. Russel and Peter Norvig. *Artificial Intelligence—A Modern Approach*, chapter 14, pages 426–435. Prentice Hall Series in Artificial Intelligence. Prentice Hall Internationall, Inc., London, UK, first edition, 1995. Excersice 14.3.

[SSHW88] Michael M. Sebring, Eric Shellhouse, Mary E. Hanna, and R. Alan Whitehurst. Expert systems in intrusion detection: A case study. In *Proceedings of the 11th National Computer Security Conference*, pages 74–81, Baltimore, Maryland, October 17–20, 1988. National Institute of Standards and Technology/National Computer Security Center.

[VT68] Harry L. Van Trees. *Detection, Estimation, and Modulation Theory Part I, Detection, Estimation, and Linear Modulation Theory*. John Wiley and Sons, Inc., New York, London, Sydney, 1968.

Appendix A Venn Diagram of the Base-Rate Fallacy Example

The Venn diagram in figure 3 graphically depicts the situation in the medical diagnostic example of the base-rate fallacy given earlier.

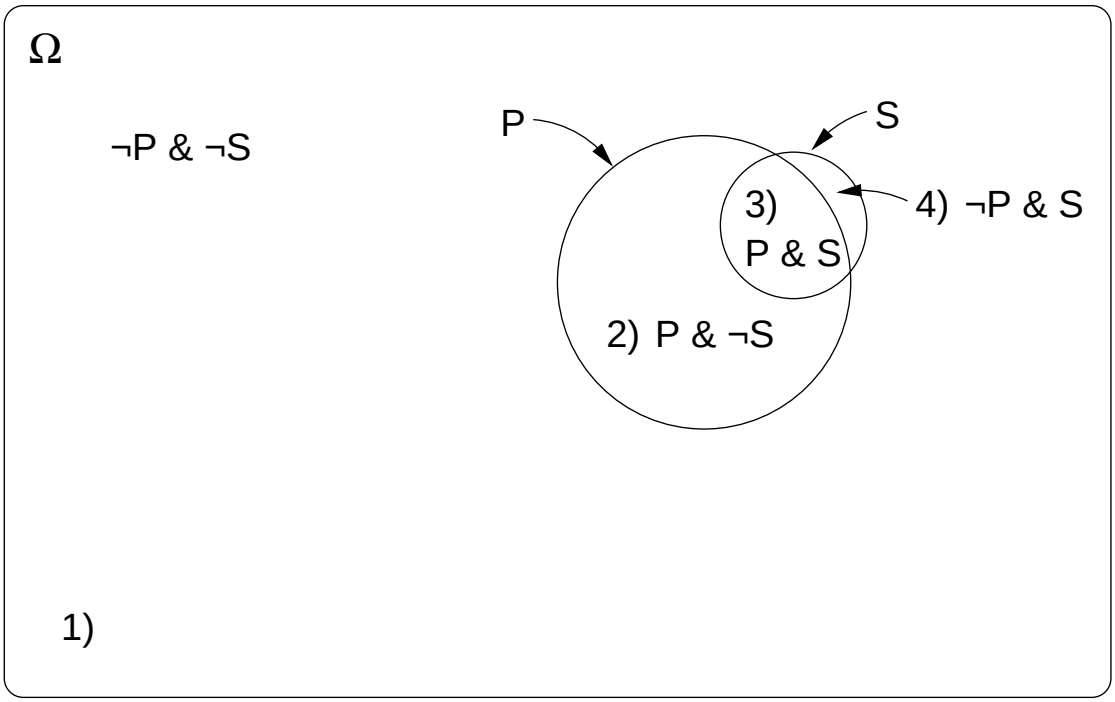

Figure 3: Venn diagram of medical diagnostic example

Though the Venn diagram is not to scale—the interesting parts would not be discernible if it were—it clearly demonstrates the basis behind the base-rate fallacy, i.e. that the population in the outcome S is much smaller than that in $\neg S$ and hence that even though $P(P|S) = 99\%$, and $P(\neg P|\neg S) = 99\%$, the relative sizes of the missing 1% in each case—area 2) and 4) in the picture—are vastly 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?" As is clear from the graph, 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.