

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

Stefan Axelsson

Presented by Kiran Kashalkar

Agenda

1. General Overview of IDS

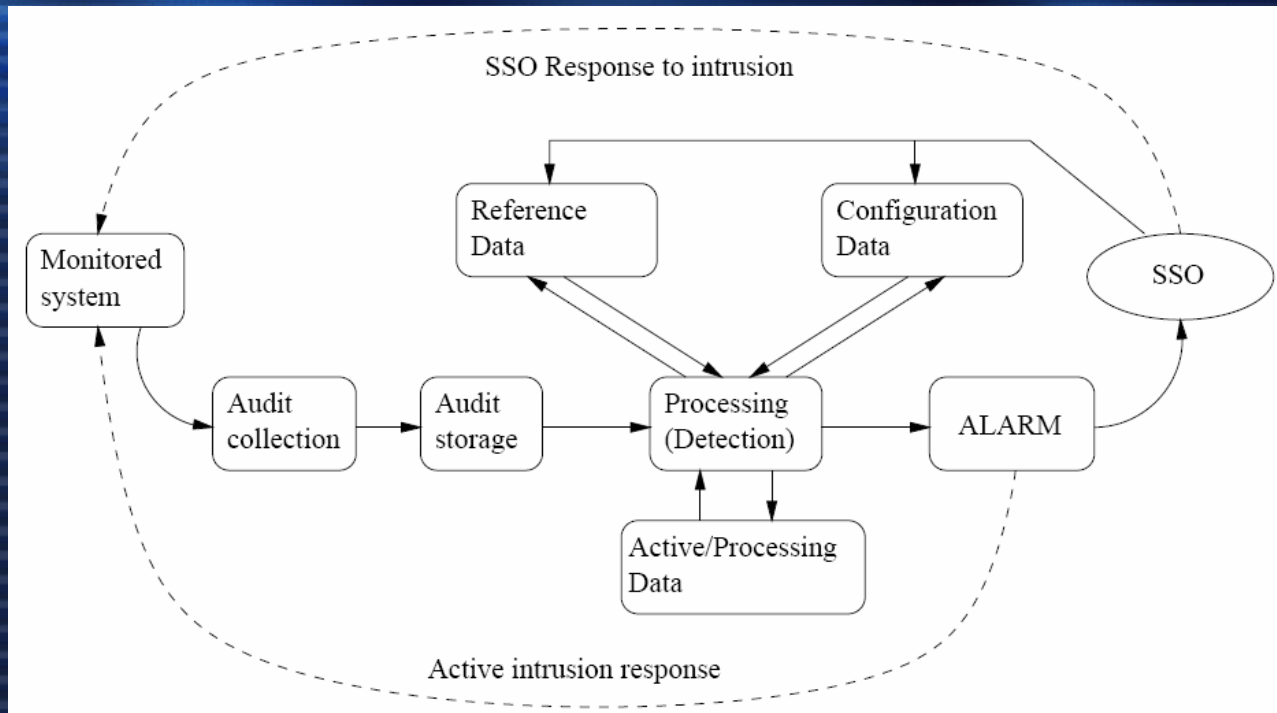
2. Bayes' Theorem and Base-Rate Fallacy

3. Base-Rate Fallacy in Intrusion Detection

4. Impact on Intrusion Detection Systems

5. Conclusion

Intrusion Detection Systems



Organization of a generalized IDS

- Intends to detect security violations from:
 - Outsiders using prepacked exploit scripts
 - Impersonators (outsiders as well as insiders)
 - Insiders abusing legitimate privileges
- Fundamental questions:
 - Effectiveness, Efficiency, Ease of use, Security, Inter-Operability, Transparency
- This paper focuses on “**Effectiveness**”

Intrusion Detection Systems (cont'd)

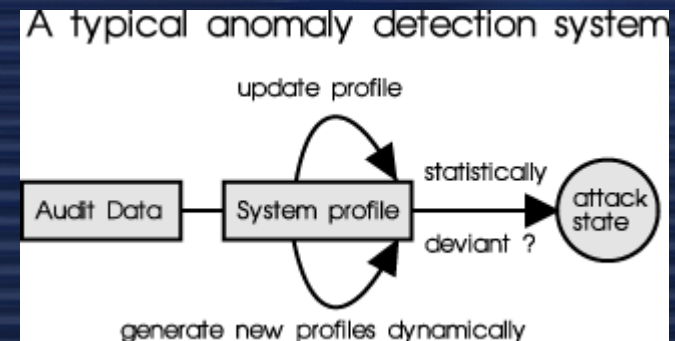
- **Few anti-intrusion techniques:**
 - **Prevention** – e.g. don't connect to the internet
 - **Preemption** – strike against threat before attack is mounted
 - **Deterrence** – increase perceived risk of negative consequences for the attacker
 - **Deflection** – e.g. use of honeypots
 - **Detection** – detect anomalies and notify authority to initiate proper response
 - **Countermeasures** – actively and autonomously counter intrusions as they are attempted

Types of Intrusion Detection Strategies

- Broadly classified into:

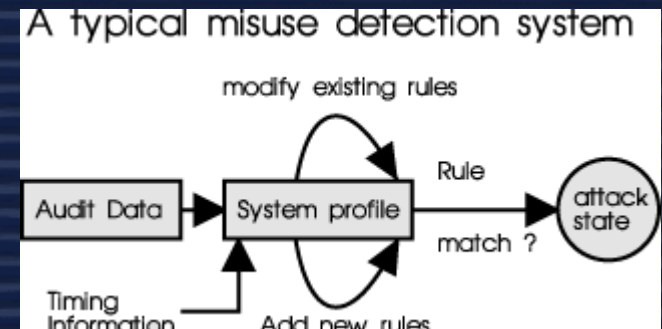
- **Anomaly detection**

- System reacts to deviations of subject behavior from normal behavior
- “Normal” subject behavior is updated as new knowledge about subject behavior becomes known



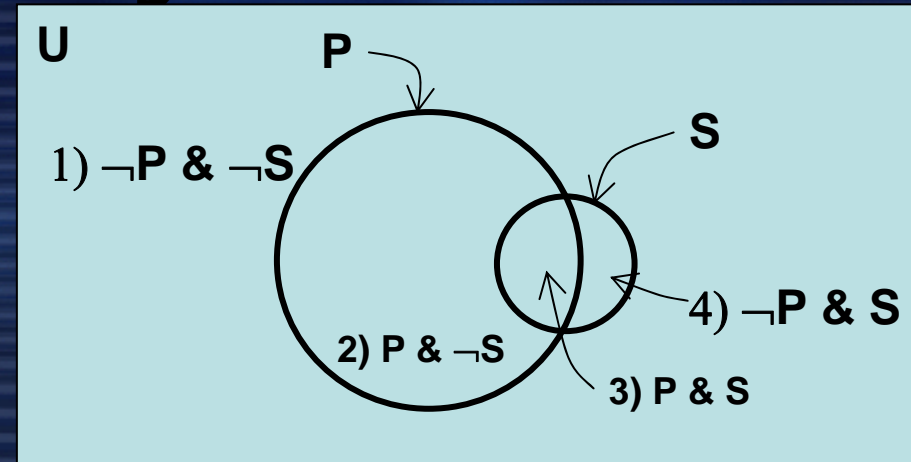
- **Policy detection (Misuse detection)**

- System tries to find evidence that matches known signatures of intrusive or suspect behavior (signature-based detection)
- System flags every action deviating from known signatures of benign behavior (specification-based detection)



Bayes' Theorem and Base-Rate Fallacy

- Given:
 - P: Person tests positive for a disease
 - S: The person is sick
 - $P(P|S)=0.99$ and $P(\neg P|\neg S)=0.99$
 - A person tested positive
 - Only 1 in 10000 people have this ailment ... Rate of incidence – $P(S)$
- Find the probability that the person is infected with that disease $P(S|P)$... works out to be only about 1%
- Fallacy: Humans don't consider base rate when intuitively solving such problems of probability



$$\begin{aligned} P(S|P) &= \frac{P(P|S)P(S)}{P(P|S)P(S) + P(P|\neg S)P(\neg S)} \\ &= \frac{1/10000 \cdot 0.99}{1/10000 \cdot 0.99 + (1-1/10000) \cdot 0.01} \\ &= 0.00980 \approx 1\% \end{aligned}$$

Base-Rate Fallacy in Intrusion Detection

- Assumptions in the hypothesized system:
 - Few tens of workstations running UNIX
 - Few servers running UNIX
 - Couple of dozen users
 - Capable of generating 1,000,000 audit records per day (with C2 compliant logging)
 - Single site security officer (SSO)
 - 10 audit records affected in the average intrusion
 - 2 intrusions per day => 20 records per 1,000,000 account to actual intrusions

Base-Rate Fallacy in Intrusion Detection (cont'd)

- Calculation of Bayesian detection rates
 - I: Intrusive behavior
 - A: Presence of an intrusion alarm
 - With the assumptions, we have:
 - $P(I) = 2 \cdot 10^{-5}$; $P(\neg I) = 1 - P(I) = 0.99998$
 - Detection rate or True positive rate: $P(A|I)$
 - False alarm rate: $P(A|\neg I)$
 - False negative rate: $P(\neg A|I) = 1 - P(A|I)$
 - True negative rate: $P(\neg A|\neg I) = 1 - P(A|\neg I)$
 - Maximize
 - $P(I|A)$: Bayesian detection rate
 - $P(\neg I|\neg A)$

Base-Rate Fallacy in Intrusion Detection (cont'd)

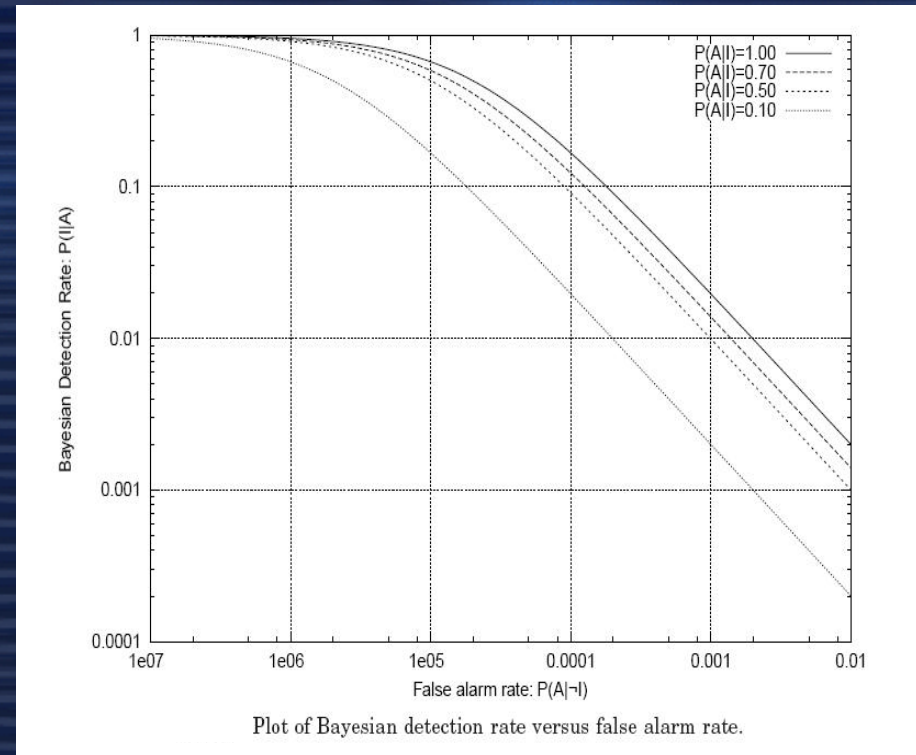
- $$P(I|A) = \frac{P(I)P(A|I)}{P(I) \cdot P(A|I) + P(\neg I) \cdot P(A|\neg I)}$$
$$= \frac{2 \cdot 10^{-5} P(A|I)}{2 \cdot 10^{-5} \cdot P(A|I) + 0.999998 \cdot P(A|\neg I)}$$
- Thus, factor governing detection rate is completely dominated by factor governing false alarm rate
- Desired maximum at **$P(A|I) = 1$ and $P(A|\neg I) = 0$**
 - Is this achievable in practice? ... NOT REALLY

Base-Rate Fallacy in Intrusion Detection (cont'd)

- For $P(A|I)=1$, $P(A|\neg I)=1\cdot 10^{-5}$, we get $P(I|A)$ as 0.66
- For $P(A|I)=0.7$, $P(A|\neg I)=1\cdot 10^{-5}$, we get $P(I|A)$ as 0.58

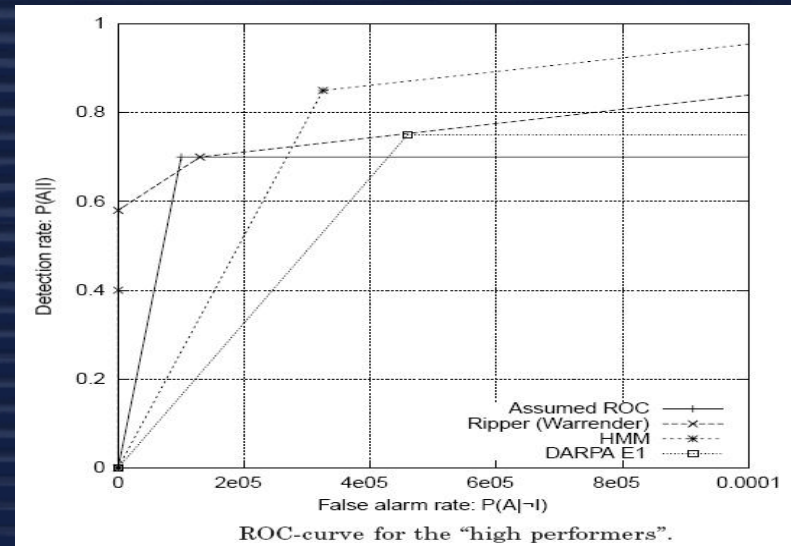
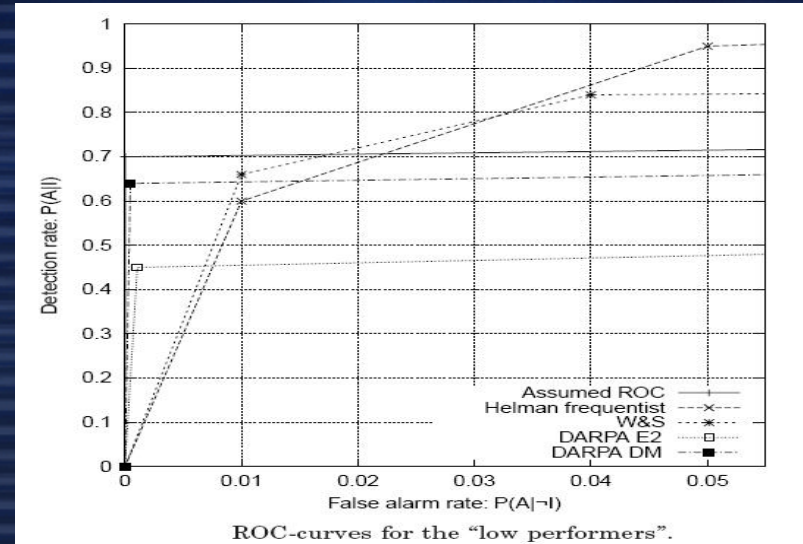
=> Even for large detection rate, viz. $P(A|I)$, Bayesian detection rate is dominated by the factor of false alarm rate, viz. factor of $P(A|\neg I)$

=> Even if $n(I|A)$ were low, $P(I|A)$ close to 50% will induce SSO to ignore all (or most) of the alarms generated



Impact on Intrusion Detection Systems

- Comparison of reported results with the established requirements on the effectiveness of intrusion detection systems
- Plot of **detection rate vs. false alarm rate**
- ROC curve analysis of the results re-establishes that detection and false alarm rates are linked
- Studies roughly divided into 2 classes:
 - With **larger false rate values** (anomaly-based detection methods)
 - With **smaller false rate values** closer to the requirements (misuse-based detection methods)



Future Work and Conclusions

- **Limitations of this paper:**
 - Use of subjective probabilities in calculations
 - Treats intrusion detection as a binary problem
 - Does not consider SSO behavior in entire correctness
- **Conclusions:**
 - Intrusion detection is difficult in real world
 - The “effectiveness” of an intrusion detection system depends not on its ability to detect intrusive behavior but on its ability to suppress false alarms
 - Comparison shows anomaly-based detection methods have larger false alarm rates than misuse-based detection, but misuse-based detection methods cannot provide protection against novel intrusions

References

- S. Axelsson. *Research in Intrusion Detection Systems: A Survey*
- <http://www.acm.org/crossroads/xrds2-4/intrus.html>
- Lawrence R. Halme and R. Kenneth Bauer. *AINT Misbehaving: A Taxonomy of Anti-Intrusion Techniques*

THANK YOU