Intrusion Detection and the Role of the System Administrator

T. Sommestad and A. Hunstad

Swedish Defence Research Agency (FOI), Sweden
e-mail: Teodor.Sommestad@foi.se; Amund.Hunstad@foi.se

Abstract

The expertise of a system administrator is believed to be important for effective use of intrusion detection systems (IDS). This paper examines two hypotheses concerning the system administrators' ability to filter alarms produced by an IDS by comparing the performance of an IDS to the performance of a system administrator using the IDS. The comparison was made through an experiment where five computer networks are attacked during four days. The experiment shows that the system administrator significantly improves the system's Bayesian detection rate, without significantly decreasing the probability that an attack is detected. Also, an analysis is made of the types of expertise that is used when output from the intrusion detection system is processed by the administrator.

Keywords

Intrusion detection systems, intrusion detection, system administrator, system administration

1. Introduction

When system administrators monitor the security of computer network(s) they often use an intrusion detection system (IDS). The IDS examines events (in network traffic, operating systems etc.) and raises an alarm if the events are believed to be symptoms of an intrusion. A number of studies investigate IDSs from a technical perspective. These studies typically investigate the technical quality of different solutions in terms of variables such as the probability of attack detection, probability of false alarm, performance constraints, and attack coverage (Mell et al. 2003; Biermann 2001). In practice, however, the IDS is not a standalone entity which makes decisions, but a tool for system administrators. The IDS administrator monitors the IDS output to filter out false alarms and attempts to verify if compromise has occurred, for example by investigating the affected system directly (Werlinger et al. 2010). Hence, in operational environments the output of an IDS is processed by an administrator who tries to detect and respond to attacks.

IDSs used in practice are tools to support system administrators. However, there are few research efforts that investigate intrusion detection that include system administrators in the investigation. Branlat (2011) studied system administrators under attack during a security exercise and identified a number of issues concerning attack detection in this context, e.g., to identify or guess the intentions of the

attackers. Werlinger et al. (2008, 2009, 2010) have acknowledged the important role of administrators and have through interviews explored both usability issues and the need of expertise and human interaction qualitatively. Similar research has been performed by Goodall et al (2004, 2009). The findings in Goodall et al. (2009) are that the work of administrators of IDSs require expertise in computer networks and security as well as a high degree of situated expertise and problem solving ability.

For an intrusion detection solution to be effective, it is of course important that it detects attacks that are made, i.e. it should have a high *P(Alarm=True | Attack=True)*. However, it is also important that it has a usable *Bayesian detection rate*. The Bayesian detection rate is the probability that an alarm which is raised corresponds to an actual attack (Axelsson 2000), i.e. *P(Attack=True | Alarm=True)*. False alarms will introduce costs (e.g., due to unnecessary actions taken by IDS administrators) and reduce the faith in that the alarms raised are worth further investigation (which might lead to real attacks being overlooked). Hence, the Bayesian detection rate cannot be too low for an effective detection processes.

The high ratio of false positives is a problem with IDSs available today and several ideas that are aimed at reducing the false alarm rate have been presented, e.g., (Sourour et al. 2009) and (Spathoulas & Katsikas 2010). However, the effectiveness of such solutions in operational scenarios remains unclear. Werlinger et al. (2008, 2009, 2010) and Goodall et al. (2004, 2009) have through qualitative studies found that the expertise of system administrators play an important role for the effectiveness of an IDS when they are used in operational settings. In other words, it is believed that a substantial amount of human involvement and expertise is required to produce a high detection rate and at the same time keep the Bayesian detection rate at acceptable levels. Thompson et al. (2007) has assessed how different types of visualisation support a system administrator. However, the authors of this paper are not aware of any quantitative results that test if a system administrator who use and interpret the output of an IDS will produce a higher Bayesian detection rate. Additionally, no previous work has been found concerning the influence a system administrator's filtering and analysis have on the probability of detection. This paper examines two hypotheses concerning the effectiveness of an IDS in comparison to an IDS plus an administrator:

Hypothesis one: *A competent administrator using an IDS will have a significantly higher Bayesian detection rate than the IDS that is used*.

Hypothesis two: *A competent administrator using an IDS will have a significantly lower probability of detection than the IDS that is used*.

The first hypothesis is based on the idea that administrators will use their expertise to filter the output of the IDS. The second hypothesis is an expected consequence of the filtering process. In other words, when the administrator filters alarms, a portion of the true alarms will be incorrectly dismissed as false alarms.

The research question investigated in this paper stems from the qualitative studies of Goodall et al. (2009) and concern the type of expertise required to use IDSs effectively. It is: *Which types of expertise does the administrator use when correct respectively incorrect judgments are made*? Of particular interest is the use of domain expertise (in intrusion detection, security, and computer networking) and situated expertise (i.e. local knowledge grounded in the analyst's environment).

Data for this study was obtained through a four-day-long experiment where a system administrator used an IDS to monitor a set of computer networks under attack by security professionals. From this test it was possible to assess and compare the effectiveness of a system administrator analysing the IDS output to the raw output of the IDS.Section 2 describes the experiment setup in more detail. Section 3 describes the result. In section 4 the result is discussed. Finally, in section 5 conclusions are drawn.

2. Data collection and analysis

2.1. Target system

The experiment was conducted using a cluster of 160 computers held by the Swedish Defence Research Agency (FOI). In this cluster, virtual machines were installed and configured to represent computer networks of five organizations in the electric power industry. Each organisation's computer network comprised some 30 machines, of which about half were server machines and half were client machines. The five organisations' computer systems were connected to each other through an internet-like infrastructure containing 44 external servers, primarily web servers.

For all organisations the machines were divided into network zones in a manner representative for a small industrial organization with a demilitarized zone, an office network, and an internal "back office" network. The computers within these five organizations' networks differed significantly, both in terms of software products used and the versions of these software products. The target systems were constructed so that security vulnerabilities varied both between organisations and between the machines within each organisation. This was accomplished by randomizing the software versions to install and the security patches to apply. Variation was also introduced in terms of memory protection mechanisms used on the different machines. The purpose of this variation was to decrease the probability that a single vulnerability could be exploited on all systems and thus force attackers to use a wider range of attack-methods.

Realistic background traffic and activity is essential when the effectiveness of IDSs is tested (Mell et al. 2003; McHugh 2000; Ranum 2001), but also difficult to produce. In an attempt to create realistic background traffic, user behaviour was simulated on the client machines. This behaviour was emulated using scripts that generated keystroke-combinations that used software installed on the client machines (e.g., the web browser) to perform predefined tasks randomly in time according to a predefined scheme. Within the organisation's networks the client machines surfed

internal and external websites, read email and opened attachments, sent email to other users within and outside the organisation, and accessed and copied shared network files.

2.2. Attackers and their attacks

The attacks in this experiment were performed by staff at the Swedish Armed Forces Network and Telecommunications Unit. The attackers had the broad objective of disturbing the technical infrastructure of the five organisations. They worked on this endeavour during office hours for four days in October 2011. They primarily used the publicly available tool BackTrack 5 as when attacks were executed. They had no prior knowledge of how the five organizations' computer networks were designed, or where they were placed in the network infrastructure.

The attacks started with a reconnaissance phase where the target systems were identified and probed for vulnerabilities. Attacks and reconnaissance were thereafter performed iteratively during the four days. In total the attackers performed 63 attack actions (including network scans, password guessing, exploitation of configuration flaws, software vulnerability exploitation). On the fourth day the attackers had managed to penetrate machines in the computer networks of two organisations and severely decreased the security of all the networks by compromising their network firewalls.

2.3. Intrusion detection solution

The IDS used in this experiment was an integration of the host-based IDS OSSEC and the network-based IDS Snort. Both are public available, open source products commonly used in operational environments. They were installed in the organisations' computer networks and configured by the administrator who monitored them during the exercise. The administrator spent approximately one week to tune the intrusion detection solution in a manner seen as appropriate. Tuning in this context means to define rules that filter alarms in order to lower the number of false alarms and defining environment-specific signatures (Goodall et al. 2009) . The administrator had also been involved in designing and configuring the five organisations computer networks and had a considerable situated expertise of this environment.

Monitoring was performed via a web-based user interface built in Zabbix, a solution for monitoring the availability and performance of computer systems. The administrator received alarms live via a number of predefined views focusing on systems at varying levels of abstraction. In addition to the possibility of viewing these alarms the administrator also had administrative rights on the attacked systems. As the focus of this experiment was the system administrator's analysis of alarms produced by an IDS, no attempts were made to prevent any of the attacks.

2.4. Measurement instrument

Both the attackers and the administrator maintained logs during the experiment. The attackers logged the attacks performed together with their success and outcome. The administrator logged anomalies which the administrator believed were related to attacks. The administrator also entered records in the log if he believed that a host had been compromised. The output of the IDS alarms was also saved. All logs were time synchronized. The data set, as well as other data collected during the exercise (e.g.., raw network traffic), are can be obtained from the authors on request.

2.5. Comparison of detection rates

Hypothesis one concerns the Bayesian detection rate of the IDS alone compared to the IDS administrator. The Bayesian detection rate is the probability that an alarm is raised because of an actual attack. In this study the attackers documented all attacks they performed and the Bayesian detection rate is assessed as the portion of all alarms which can be connected to any of these attacks. An experienced network security expert was used to judge whether this was the case, based on all information available in the logs.

Hypothesis two concerns the probability of detection, i.e., the probability that an attack will cause an alarm. As for the Bayesian detection rate the attackers' log was used to compare the performance of the IDS and its administrator. For each entry in the attackers' log book it was identified if an alarm had been produced because of the actions associated with the attack. The criterion was that at least one alarm should have been raised which could be tied to the actions taken by the attackers. Hence, it was not required that an alarm described everything the attackers did (e.g., all ports that was scanned) in order for the attack to be considered detected.

For both the IDS and the system administrator Bayesian detection rate was obtained as the ratio between the number of alarms raised because of the actions of the attackers and the number of alarms raised in total. The probability of detection was obtained as the ratio between the number of attacks with an alarm tied to it and the number of attacks performed in total.

Fisher's exact test (Fisher 1922) tests if there are non-random associations between two categorical variables. In other words, it can be used to tests if categorical variables have distributions that are different. In this paper it is used to test hypotheses one and two by comparing the result of the IDS and its administrator. A significance level of 0.05 is used in this experiment.

2.6. Interviews concerning the use of expertise

To obtain knowledge of the role of expertise in intrusion detection, an interview with the administrator monitoring the IDS during the experiment, was performed. Due to the large amount of data collected it was necessary to choose a small data subset to focus on for the interview. For each alarm/decision category a number of cases were

identified as of special interest to consider. These cases included scenarios where the IDS and/or administrator made correct as well as incorrect judgments. In particular, the interview identified three scenarios for each of the following cases: the system produced several alarms and the administrator made a correct decision; the system produced several alarms and the administrator made an incorrect decision; the system produced few or no alarms and the administrator made a correct decision; the system produced few or no alarms and the administrator made an incorrect decision.

The interview was semi-structured in the sense that the categorization and cases guided the interview, but apart from that no formal interview guide with pre-defined questions was used. The interview was documented with notes which were later confirmed by the respondent.

3. Results

3.1. The performance of the administrator and IDS

During the experiment the administrator produced 70 alarms and the IDS produced 2107 alarms. Meanwhile, the attackers performed 63 actions involving reconnaissance (often network scans) or direct attack (e.g., password guessing). The alarms raised by the IDS pointed to 44 of these actions (i.e., 19 were missed); the alarms raised by the administrator pointed to 37 of these actions (i.e. 26 were missed). The IDS thus outperforms the administrator with when it comes to the probability that an attack is detected (69% vs. 58%). However, with respect to the Bayesian detection rate the system administrator outperformed the IDS. Of the 70 alarms raised by the administrator 40 (57%) was found to be causes of actual attacks; of the 2107 alarms raised by the IDS only 233 (11%) were due to actual attacks.

Hypothesis one states that the administrator will produce a higher Bayesian detection rate than the IDS. To test if the results can be from the same probability distribution Fisher's exact test (Fisher 1922) is applied on the contingency table. The test shows that the difference is significant ($p<0.0001$). Hypothesis two states that the administrator will significantly lower probability of detection compared to the IDS. However, Fisher's exact test does show a significant difference ($p=0.2645$) between the IDS and system administrator with regards to probability of detection.

	IDS	Administrator
Alarms raised	2107	70
Probability of detection	69%	58%
Bayesian detection rate	11%	57%

Table 1: Performance of IDS and the administrator.

3.2. The role of expertise

The interview with the human administrator gave some insight in the reasoning and knowledge that applied when IDS are used.

A number of the analysed cases correspond to when *the administrator was facing several alarms and made a correct decision.* This includes scenarios where the administrator correctly identified that the alarms are caused by attack and when the administrator correctly identifies them as false alarms. When identifying the alarms correctly the primary pieces of information used was: the involvement of an external IP-address and the unusually large number of alarms which matched the typical traces of attack-tools. Thus, both situated expertise and knowledge about security is used. In all cases when t*he administrator was facing several alarms and made an incorrect decision* the reason was simply that the alarms were missed by the administrator due to the high amount of alarms.

In many cases the administrator correctly identified a comparable amount of alarms as a false positive. In these cases the administrator correctly identified the alarms as causes of normal network traffic (general network expertise) or hypothesized that it would be unlikely that the attackers had access to the machine involved (situated expertise). In one case the administrator correctly ignored alarms because they would have originated from attacks the administrator considered the attackers incapable of. In other words, based on knowledge of which vulnerabilities that were present and a good idea of which resources the attacker had the alarms could be dismissed.

In other cases *the administrator was facing few or no alarms and makes a correct decision.* When few alarms was correctly associated with an attack the information used to identify maliciousness was the presence of external IP-addresses and the indirect effects of attacks, e.g., when a machine starts to execute strange requests to other machines or extraordinary network loads appear. Thus, situated expertise was the primary input to decisions made in such situations.

In a few cases t*he administrator faced few or no alarms and made an incorrect decision.* The reason was that the administrator was misdirected by another, high priority, false alarm. In one case the reason was an incorrect hypothesis concerning the privileges acquired by the attackers

4. Discussion

The primary findings of this experiment are described in section 4.1. This experiment has several limitations and its result should therefore be interpreted with care. Some issues with making broad generalizations from this experiment are described in sections 4.2 and 4.3.

4.1. Primary findings

The primary findings in this experiment is that the IDS administrator produces significantly better filtered output than IDS systems do alone, and the administrator's filtering does not significantly impact the detection rate. As suggested in Goodall et al. (2009), the administrator do so by using situated expertise, general expertise about computer networks, and general expertise about security and attacks. In this experiment the administrator also used knowledge (or guesses) about which attack

methods and tools the attackers had access to and would use. Knowledge about the capabilities of the threat agent does not correspond to any of the expertise-types identified by Goodall et al. (2009), but can apparently be effectively used to filter alarm lists.

4.2. Control and sampling of nuisance variables

A number of nuisance variables which can be expected to influence the result are kept constant in this test. This includes variables that are given by the administrator (e.g. the competence), the IDS (e.g. signatures and tuning), and the interface between the administrator and the IDS (e.g. how alarms are visualized). As these are kept constant, the result does not reveal how these influence the result. A reasonable hypothesis is that they all have an impact on the result. For instance, it appears likely that a better tuned IDS would increase the IDS's Bayesian detection rate and decrease the difference between the administrator and system. On the other hand, the result from this setup is so clear that it appears unlikely that the overall conclusion would be different. In particular, it appears unlikely that the Bayesian detection rate of the system would come close to that of a competent administrator monitoring it.

A number of nuisance variables in this environment were sampled to produce meaningful variation. Meaningful in this case means that they are varied to represent conditions which make the result generalizable to a realistic context. Attacks executed by the attackers varied over the experiment, the configuration of attacked networks and computers differed, and background traffic varied in the experiment.

Since the attackers actually performed attacks with an explicit objective it appears likely that they represent a set of steps which resembles those attacks undertaken when a computer network is attacked. Likewise, the systems under attack were designed to resemble those of typical organizations in the electrical power industry, with different network zones and types of computer machines. However, it is unclear if the somewhat artificial variation influences the result.

An important factor for the Bayesian detection rate is the background traffic. Effort was therefore made to produce background traffic which resembles real actions in the sense that real software applications were used and real requests were made to a diverse set of websites, etc. As stated in (Mell et al. 2003), "there is no such thing as a 'standard' network", which makes it difficult to produce background traffic so that the result is generalizable to a wide context. The scripts used in this study to generate user actions (i.e. background traffic) and other records from the experiment can be obtained from the authors.

4.3. The measurement instrument and unit of analysis

The unit of analysis in this study is the entries in the attackers' log book. Alarms which are a causal effect of the entries in this log book are correct alarms (true positives); alarms that do not match an entry in this log book are incorrect alarms (false positives). In conventional tests of IDSs the unit of analysis is the unit which

the IDS bases decisions on, typically a network session or operating system event. This study uses a coarser unit of analysis since the more coarse decisions of the administrator is to be compared to those of the IDS. While intrusive actions are intuitively meaningful as a unit to detect, its definition is less rigid than conventional units of analysis. For instance, some entries in the log book involves hundreds of network sessions (e.g. a network scan) while other only involve one (e.g. a software vulnerability exploitation attempt).

While it is possible that another set of attackers (with other opinions on what intrusive behaviour is) would have produced a different log book, it appears unlikely that it would result in a dramatically different result. Manual inspections of a subset of the alarms marked as "false positives" strengthen this belief – no apparent traces could be found to actions taken by the attacker (e.g., to machines they had control over).

5. Conclusions

This experiment confirms earlier findings concerning the importance of expertise in the use of intrusion detection systems. In this experiment the intrusion detection system administrator achieves a significantly better Bayesian detection rate than the intrusion detection system, and the administrator's detection rate is not significantly different from the intrusion detection system. An administrator can achieve this effective filtering by using situated expertise, computer networks expertise, computer security expertise, and knowledge about what the threat agent is capable of.

6. References

Axelsson, S., 2000. The base-rate fallacy and the difficulty of intrusion detection. *ACM Transactions on Information and System Security*, 3(3), pp.186-205.

Biermann, E., 2001. A comparison of Intrusion Detection systems. *Computers & Security*, 20(8), pp.676-683.

Branlat, M., 2011. *Challenges to Adversarial Interplay Under High Uncertainty: Staged-World Study of a Cyber Security Event*. The Ohio State University.

Fisher, R.A., 1922. On the interpretation of chi2 from contingency tables, and the calculation of P. *Journal of the Royal Statistical Society*, 85(1), pp.87–94.

Goodall, J.R., Lutters, W.G. & Komlodi, A., 2009. Developing expertise for network intrusion detection. *Information Technology & People*, 22(2), pp.92–108.

Goodall, J.R., Lutters, W.G. & Komlodi, Anita, 2004. I know my network: collaboration and expertise in intrusion detection. In *Proceedings of the 2004 ACM conference on Computer supported cooperative work*. ACM, pp. 342–345.

McHugh, J., 2000. Testing Intrusion detection systems: a critique of the 1998 and 1999 DARPA intrusion detection system evaluations as performed by Lincoln Laboratory. *ACM Transactions on Information and System Security*, 3(4), pp.262-294.

Mell, P. et al., 2003. *An overview of issues in testing intrusion detection systems, (NIST IR 7007)*, National Institute of Standard and Technology.

Ranum, M.J., 2001. Experiences Benchmarking Intrusion Detection Systems. *Security*, NFR Security, pp.1-10.

Sourour, M., Adel, B. & Tarek, A., 2009. Environmental awareness intrusion detection and prevention system toward reducing false positives and false negatives. In *2009 IEEE Symposium on Computational Intelligence in Cyber Security*. IEEE, pp. 107-114.

Spathoulas, G.P. & Katsikas, S.K., 2010. Reducing false positives in intrusion detection systems. *Computers & Security*, 29(1), pp.35-44.

Thompson, R.S. et al., 2007. Command line or pretty lines?: comparing textual and visual interfaces for intrusion detection. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, p. 1205.

Werlinger, R. et al., 2010. Preparation, detection, and analysis: the diagnostic work of IT security incident response. *Information Management & Computer Security*, 18(1), pp.26–42.

Werlinger, R. et al., 2009. Towards Understanding Diagnostic Work During the Detection and Investigation of Security Incidents. In *Proceedings of the Third International Symposium on Human Aspects of Information Security & Assurance (HAISA 2009)*. Lulu.

Werlinger, R., Hawkey, K. & Muldner, K., 2008. The challenges of using an intrusion detection system: is it worth the effort? *SOUPS '08 Proceedings of the 4th symposium on Usable privacy and security*, (1).