JISSec
Journal of Information System Security

QUANTIFYING THE EFFECTIVENESS OF INTRUSION DETECTION SYSTEMS IN OPERATION THROUGH DOMAIN EXPERTS

Teodor Sommestad
The Royal Institute of Technology (KTH), Sweden

Hannes Holm; Mathias Ekstedt; Nicholas Honeth
The Royal Institute of Technology (KTH), Sweden

teodor@sommestad.com; hannes.holm@ics.kth.se; mathias.ekstedt@ics.kth.se; nicholas.honeth@ics.kth.se

Journal of Information System Security is a publication of the Information Institute. The JISSec mission is to significantly expand the domain of information system security research to a wide and eclectic audience of academics, consultants and executives who are involved in the management of security and generally maintaining the integrity of the business operations.

Editor-in-Chief
Gurpreet Dhillon
Virginia Commonwealth University, USA

Managing Editor
Filipe de Sá-Soares
University of Minho, Portugal

ISSN: 1551-0123
Volume 10, Issue 2

www.jissec.org

Abstract

An intrusion detection system (IDS) is a security measure that can help system administrators in enterprise environments detect attacks made against computer networks. In order to be a good enterprise security measure, the IDS solution should be effective when it comes to making system operators aware of on-going cyber-attacks. However, it is difficult and costly to evaluate the effectiveness of IDSs by experiments or observations. This paper describes the result of an alternative approach to studying this topic. The effectiveness of 24 different IDS solution scenarios pertaining to remote arbitrary code exploits is evaluated by 165 domain experts. The respondents' answers were then combined according to Cooke's classical method, in which respondents are weighted based on how well they perform on a set of test questions. Results show that the single most important factor is whether either a host-based IDS, or a network-based IDS is in place. Assuming that either one or the other is in place, the most important course of action is to tune the IDS to its environment. The results also show that an updated signature database influences the effectiveness of the IDS less than if the vulnerability that is being exploited is well-known and is possible to patch or not.

Keywords: intrusion detection system; security architecture; expert judgment; incident handling; signature-based detection.

JISSec
Journal of Information System Security

1. Introduction

Intrusion detection systems (IDS) are promising security measures that are commonly used to defend information systems (Sumner, 2009). An IDS monitors a computer network or its hosts to detect attacks. Once attacks are identified, administrators can then be notified and appropriate actions can be carried out. However, IDSs are not perfect. They can fail to detect attacks that take place and can raise alarms for events that are not actually attacks. Thus, in practice, it is not sufficient to just receive an alarm from the IDS, as system administrators must also have confidence in the IDS and act on the alarm. In this paper, as in that of Axelsson, 2000b, effectiveness is defined as: the probability that the administrator reacts appropriately when an attack occurs.

The development of models and techniques for IDSs dates back three decades (Anderson, 1980; Denning, 1987) and even though there are a wide range of IDS solutions on the market today, IDSs are still a viable research field. A problem for both research and practice is that how effective an enterprise's IDS is in different operating conditions is largely unknown. Several variables are believed to impact the effectiveness of an IDS in operation. For example, if the rules or models of the IDS are updated, whether the IDS has been tuned for its environment, and if it is host-based, or network-based (Scarfone & Mell, 2007). For a decision-maker who considers installing or adjusting an IDS, the impact of such variables on effectiveness are of high relevance to guide them in making effective system design decisions. However, little is known about this impact.

One of the main reasons for the lack of knowledge concerning the effectiveness of different solutions are the difficulties and costs associated with evaluating configurations in realistic environments. Several challenges have been identified for empirical tests of IDSs: e.g., the generation of realistic background traffic (Barry & Chan, 2010; Mell, Hu, Lippmann, Haines, & Zissman, 2003). As a result, quantitative studies are typically made in artificial settings to assess the impact of one specific parameter. For example, such tests have been made regarding the impact of tuning configuration parameters in certain platforms (Salah & Kahtani, 2009), to see how the detection rate depends on its host's hardware performance (Alserhani et al., 2009), and to ascertain how well different IDS products detect network scans (Ktata, Kadhi, & Ghédira, 2009). Reliable empirical studies on the effectiveness of IDSs in operational settings are not available in the literature. Moreover, system administrators play an important role in operational environments as analysts of alarms. However, few studies consider the system administrator in their tests. In fact, the only study which addresses operational effectiveness and

includes the system administrator is the experiment carried out by Sommestad and Hunstad (2013).

Expert judgment is often used when quantitative data is difficult to obtain from empirical studies, or by other means. It has been used to assess the importance of attributes related to critical infrastructure risks (Cooke & Goossens, 2004), to quantify uncertainties related to crops (Krayer von Krauss, Casman, & Small, 2004) and recently to assess strategies related to security (McFadzean, Ezingeard, & Birchall, 2011), as well as much more (for more examples, see (Cooke 2008)). This paper describes a study in which a survey was used to collect expert judgment that quantifies the effectiveness of signature-based IDSs in different operational scenarios. The experts in this study are all researchers in the field of IDSs, who used their domain knowledge to assess whether arbitrary code execution attacks would be detected by an administrator in 24 different operational scenarios. The respondents' judgments were synthesized with an established method that assigns weights to domain experts' judgment, based on their replies to a number of test questions. Based on the synthesized result of the domain experts' assessments, recommendations to information security professionals and researchers are presented.

The paper is structured as follows. Section Two presents the operational scenarios that were investigated and the variables used to specify these. In Section Three, Cooke's classical method for expert judgement is explained. This method is used to sort out the experts that produced calibrated assessments and to determine whose answers should be trusted. Section Four presents the data collection method. Section Five presents the results on estimates of IDS's effectiveness in 24 operational scenarios and the influence of variables on this effectiveness. In Section Six, these results are discussed and in Section Seven conclusions are drawn.

2. Operational scenarios – a prediction model for IDS effectiveness

The quality of IDSs can be evaluated by a number of criteria (Biermann, 2001). In accordance with Axelsson's (2000a) definition of effectiveness, this research investigates the probability that actual attacks are detected and are reacted upon by the administrator who is monitoring the IDS. The effectiveness was studied for remotely executed arbitrary code attacks, and a number of operational scenarios for IDSs were investigated. These operational scenarios were specified using a number of variables identified as being important for IDS effectiveness, based on a literature review and after consultation with

three security experts who work in the field of IDSs. A summary of the variables identified in the literature review is presented in Section 0; the variables used in the present study are described in Section 0

2.1 Literature review

A plethora of detection methods and techniques have been introduced since studies on intrusion detection were first made in the 1980's, with studies such as those of Anderson (1980), and Denning (1987). A number of papers divide these methods into broad classification schemes. A common division is made between anomaly-based intrusion detection and signature-based (or misuse) intrusion detection (Axelsson, 2000a; Biermann, 2001; Garciateodoro, Diazverdejo, Maciafernandez, & Vazquez, 2009). Anomaly-based intrusion detection estimates the normal behaviour of a system and generates an alarm when the deviation from the normal behaviour exceeds a stipulated threshold (Garciateodoro et al., 2009). Signature-based schemes look for patterns (signatures) in the analysed data and raise an alarm if the patterns match a known attack (Garciateodoro et al., 2009). Some classifications also distinguish specification-based engines from the signature and anomaly based schemes (Xenakis, Panos, & Stavrakakis, 2010); others regard specification-based engines as a subset of anomaly-based detection (Axelsson, 2000a). In specification-based engines, any activity that deviates from predefined constraints (e.g., descriptions of correct behaviour) would cause alarm. Hybrids, or compound solutions, are also possible (Axelsson, 2000a; Scarfone & Mell, 2007).

Anomaly-based detection schemes have been given most of the attention in recent research on intrusion detection systems. (Garciateodoro et al., 2009) describe techniques used by these systems to detect anomalies such as: statistical-based, knowledge-based, or machine-learning based. (Biermann, 2001) divide them into: statistical, sequence matching and learning, predictive pattern generation, and neural network-based detection schemes.

Signature-based detection schemes also come in different variants. (Axelsson, 2000a) divide them into: state-modelling, expert system, string matching, and simple rule-based schemes. (Biermann, 2001) on the other hand divided them into: expert system, keystroke monitoring, model-based, state transition analysis, and pattern matching schemes. Whilst most research has been performed on anomaly-based detection in recent years, the majority of IDSs that are commercially available and are used in practice, are signature-based (Faysel & Haque, 2010). For example, Gartner observes that signature quality remains the primary selection factor on the market for IDSs with preventive capabilities (i.e., intrusion prevention systems) (Young & Pescatore, 2009).

This, as well as the input from the three domain experts, made signature-based IDSs the focus of this study.

The detection model can make a difference to the effectiveness of an IDS. It is often noted that signature-based detection only detects attacks that correspond to known signatures, whilst anomaly-based detection also can detect previously unknown attack types (see for example (Garciateodoro et al., 2009)). The coverage, i.e., the attack types the IDS can detect, is of obvious relevance to the effectiveness in practice (Mell et al., 2003)(McHugh, 2000). The algorithm used also makes a difference. For instance, (Ashfaq et al., 2008) compare a number of anomaly detection algorithms and show a clear difference in their performance. Much research effort has been dedicated to algorithms for detection. However, a wide range of other variables are also of importance for the effectiveness of operational IDSs. These are discussed below.

It can make a difference if sensors are placed on network hosts, in the network infrastructure, or on both (Scarfone & Mell, 2007). In addition, it can make a difference if sensors are placed so that they listen to the network passively (e.g., through a spanning port or network tap), or if they are placed inline, so that all traffic must go through them (Scarfone & Mell, 2007; Shaikh et al, 2008). Host-based sensors could also be placed inside, or outside, the code they are supposed to monitor, and as this influences what the sensor can monitor, it will also have an impact on effectiveness (Shaikh et al., 2008). Protocols, protocol layers and the amount of traffic the an IDS can handle are also of relevance (Mell et al., 2003; Scarfone & Mell, 2007). Furthermore, the environment in which the sensors are placed can be expected to influence effectiveness, and also interact with the factors listed above (McHugh, 2000; Mell et al., 2003). For example, complex and intensive network traffic may give rise to higher instances of false alarms and make it difficult for the IDS to identify actual attacks.

In addition to detection mechanism, the placement of sensors, and the environment, there are a number of further variables that are related to deployment and management of IDS, which can be expected to influence their performance. Configuration and tuning is of importance to both anomaly-based and signature-based intrusion detection (Scarfone & Mell, 2007). Configuration parameters include: thresholds and alert settings to optimize false positives and false negatives (Scarfone & Mell, 2007; Werlinger, Hawkey, Muldner, Jaferian, & Beznosov, 2008), tuning and customizing the system for its environment (Scarfone & Mell, 2007; Werlinger et al., 2008) and securing the actual IDS from attacks (Mell et al., 2003; Scarfone & Mell, 2007).

Deploying an IDS correctly is generally challenging, and as a consequence, the competence of system administrators is an important factor (Scarfone & Mell, 2007; Werlinger et al., 2008). For example, the administrators' programming skills and their knowledge about the environment where they are supposed to deploy IDSs are of relevance (Scarfone & Mell, 2007; Werlinger et al., 2008).

After deployment, the detection system needs to be maintained and managed. Updating the system and its engine to the latest version is part of this management process (Scarfone & Mell, 2007). For signature-based detection systems, it is essencial to continually update the signature database (Scarfone & Mell, 2007). Periodic testing of IDS's functionality has also been suggested (Scarfone & Mell, 2007).

Alarm lists may be comprised by as many as 99 % false alarms, and methods that assist the administrator in identifying actual attacks are therefore important (Julisch & Dacier, 2002). The abovementioned variables influence the amount of false alarms that are raised by IDS and also the amount of attacks that they miss. In the end, however, an administrator must be able to distinguish actual attacks from false positives, and decide how to react. It is the effectiveness achieved at this stage that is investigated in this paper. Thompson et al. (2006) present design recommendations to ease the cognitive burden placed on administrators, using visualization and different proposed techniques for visualization (e.g., (Itoh, Takakura, Sawada, & Koyamada, 2006; Thompson, Rantanen, Yurcik, & Bailey, 2007)). While visualization of alarms and the network's status can help the administrator, the competence of this person is also an important factor. Goodall et al. (2009) found that administrators require expertise in networking, security, and a portion of situation expertise (e.g., about the business that they work in) to carry out their task. Moreover, they are often faced with problems that are not predefined and change as environments evolve (Goodall et al., 2009).

2.2 Variables specified in the assessed scenarios

As described in 0, there are numerous variables that may influence the effectiveness of an IDS in operation. One could specify operational scenarios by assigning values to all of these variables, e.g., with regards to the employed algorithm(s), the competence of operators and the profile of background traffic. However, doing so would only show the value of these exact configurations and would limit the validity of the result for these particular cases. Collecting such detailed information in an enterprise-context would also be extremely expensive, and prediction models requiring this level of detail would be expensive to use. Also, as shown by the critique against experimental efforts

(McHugh, 2000), it is difficult to identify all the variables that are of relevance and to make sure that they are representative of typical operations.

This study aims to provide approximate values for IDSs' effectiveness and the approximate importance of a number of important variables related to them. To maintain generality of the estimates produced, it focuses on a number of carefully selected variables and let the greater majority of variables vary, just as they typically do so in an enterprise's environment. Because the variables that influence effectiveness (e.g. how competent administrators are) vary between enterprises, the exact effectiveness will be uncertain when they are left unspecified. This uncertainty is managed by expressing the effectiveness through a probability distribution which captures the uncertainty caused by this noise. Hence, for each operational scenario, the experts were asked to provide estimates of effectiveness in terms of a probability distribution that was representative for their enterprises, given that unspecified variables vary, just as they do in practice.

The selection of variables to be included in the operational scenarios was made by consulting three experts on IDSs. These domain experts were presented with a list of variables and were then asked to complement this list with other variables that they had found to be important. They were then asked to prioritize these variables, based on how much they simplified predictions on the effectiveness of an IDS. Because the variables were supposed to be used for predictions, they also discussed whether the system owner would be able to identify their values for an installation. Based on this prioritization procedure, the model depicted in Figure 1 was used.

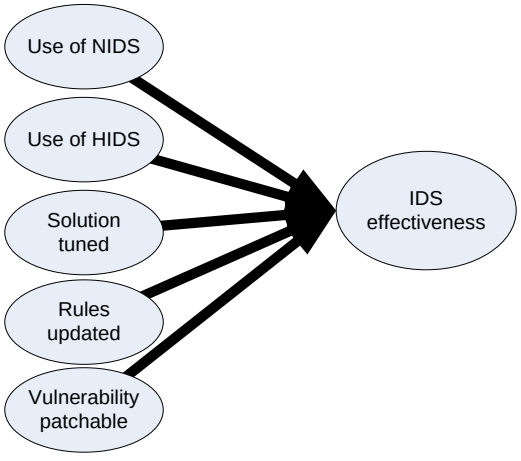

Figure 1. Variables studied.

Table 1 describes the five variables used to describe the different operational scenarios. In total, 24 different operational scenarios are investigated, each corresponding to a specific configuration of the five variables. Two of the five variables concern the placement of sensors – i.e., whether the IDS is host, or network-based. One variable selected by the domain experts was the tuning of IDSs. Updates of signatures was also regarded as an important variable. This was operationalized as to whether the signature of the system is fully updated, or not. Finally, the type of vulnerability that was to be exploited was judged to be important. A signature-based IDS is presumably less effective in scenarios with unknown attacks, as noted in 0. In this model, the vulnerability-type exploited is captured by considering scenarios where it is possible to patch the attack with a software update (i.e., are well-known vulnerabilities), as well as scenarios where the exploit uses software vulnerabilities for which no patch is available. This variable was more highly prioritized than the exact signature match for exploits used by the attacker, due to the fact that details on the latter are difficult to collect in practice.

Variable	Description
NIDS	Whether a network-based intrusion detection system is used, or not.
HIDS	Whether a host-based intrusion detection system is used, or not.
Tuned	Whether the intrusion detection systems used have been tuned for their environment, or not.
Updated	Whether the signatures used by the intrusion detection systems are fully updated, or not.
Patchable	Whether the exploit they are supposed to detect use a vulnerability that can be patched, or not.

Table 1. Variables included in the model.

In addition to the five variables, the respondents helped to identify two assumptions which would have a limited effect on the usability of the result. Firstly, assumptions were made concerning the attack scenario to match a common threat. It was assumed that the attack scenario was specified as a remote arbitrary code exploit, performed by an external professional penetration tester, with the possibility to spend one week in preparing the attack. Thus, the attacker exploits a software vulnerability in order to execute code on the targeted system. Secondly, it was assumed that the detection scheme was signature-based. The second assumption was made because the experience of the three domain experts was that the vast majority of IDSs installed in enterprises today use this detection scheme, and it is therefore more interesting.

3. Method used to synthesize expert judgments

This paper uses the judgment of domain experts to produce quantitative estimates of IDSs' effectiveness in different scenarios. There is a substantial amount of research on techniques to combine, or synthesize, the judgment of multiple experts to increase the calibration of the estimates used. These techniques include the following: consensus methods (Fink et al., 1984; Ashton, 1985), the Cochran-Weiss-Shanteau index (Weiss & Shanteau, 2003), self-proclaimed expertise (Abdolmohammadi & Shanteau, 1992), experience (Shanteau et al., 2002), certifications (Shanteau et al., 2002), peer-recommendations (Shanteau et al., 2002), and Cooke's classical method (Cooke, 1991). There is little research that compares the accuracy obtained by these methods. However, research has shown that groups of individuals assess an uncertain quantity better than the average expert, whilst the best individuals in the group are often better calibrated than the group as a whole (Clemen and Winkler, 1999). Research has also shown that consensus is related to accuracy, but that the relationship between accuracy experience and the relationship between self-proclamation is less clear (Holm, Sommestad, Ekstedt, & Honeth, 2013).

The scheme used to combine judgments in this research is the one used in the classical model of Cooke (Cooke, 1991). Cooke's model is a generic method for combining expert judgments that has been applied to a number of different domains. Applications of Cooke's classical method show that it outperforms both the best expert and the equal-weight-combination of experts' estimates. In an evaluation involving 45 studies, it performed significantly better than both alternatives in 27 studies, and equally well as the best expert in 15 of the studies (Cooke, 2008).

In Cooke's classical method, two scores are calculated for the respondents for the purpose of weighting them. One for calibration, and another for information. These two scores are based on the respondents' answers to a set of seed questions, i.e., questions for which the true answer is known at the time of the analysis. The calibration score shows how correctly a respondent's answers reflect the true value, and the information score shows how precise a respondent's answer is. A decision maker is formed by assigning weights based on their scores. The weights defined by this decision maker are then used to weight the respondents' answers to the questions of interest.

Thus, the method filters out individuals who are true experts from a pool of potential experts, given the accuracy and preciseness of their answers to a set of test questions. Only those filtered out as true experts are used to estimate the probabilities of questions of interest (i.e., the operational scenarios

described in Section 2). In sections 3.1, 3.2 and in 3.3, Cooke's classical method is explained. For a more detailed explanation, the reader is referred to (Cooke, 1991).

3.1 Calibration score

In the elicitation phase, the experts provide individual answers to the seed questions. The seed questions require that the respondents specify a probability distribution for an uncertain continuous variable. This distribution is typically specified by stating its 5th, 50th, and 95th percentile values, which yield four intervals over the percentiles *[0-5, 5-50, 50-95, 95-100,* with probabilities of *p= [0.05, 0.45, 0.45, 0.05]*. As the seeds are realizations of these variables, the well-calibrated expert will have approximately 5% of their realizations in the first interval, 45 % of their realizations in the second interval, 45 % of their realizations in the third interval and 5% of their realizations in the fourth interval. If *s* is the distribution of the seed over the intervals, the relative information of *s* with respect to *p,* is: $I(s,p) = \sum_{i=1}^{4} \ln(s_i/p_i)$. This value indicates how surprised someone would be if one believed that the distribution was *p,* and then learnt that it was *s*.

If *N* is the number of samples/seeds, the statistic of 2*NI*(*s*, *p*) is asymptotically Chi-square distributed, with three degrees of freedom. This asymptotic behaviour is used to calculate the calibration *Cal* of expert *e* as: $Cal(e) = 1 - \chi_3^2(2N\, I(s,p))$. Calibration measures the statistical likelihood (i.e., p-value) of the hypothesis that the realizations of the seeds (*s*) are sampled independently from distributions agreeing with the expert's assessments (*p*).

3.2 Information score

The second score used to weight experts, is the information score, i.e., how precise and informative the expert's distributions are. This score is calculated as the deviation of the expert's distribution to some meaningful base distribution. In this study, the base distribution is the uniform distribution over [0,1], which represent no knowledge at all about likely outcomes.

If b_i is the background density for seed $i \in \{1,\ldots,N\}$, and $d_{e,i}$ is the density of expert *e* on seed *I*, then the information score for expert *e* is calculated as: $\inf(e) = \frac{1}{N}\sum_{i=1}^{N} I(d_{e,i}, b_i)$, i.e., as the relative information of the experts distribution, with respect to the base distribution. It should be noted that the information score does not reflect calibration and does not depend on the realization of the seed questions. So, regardless of what the correct answer is to a seed question, a respondent will receive a low information score for an answer which is similar to the base distribution, i.e., the answer is distributed

evenly over the variable's range. Conversely, an answer which is more certain, and assigns most of the probability density to a few values, will yield a high information score.

3.3 Constructing a decision maker

Cooke's classical method rewards experts who produce answers with a high calibration (high statistical likelihood) and a high information value (high precision). A strictly proper scoring rule is used to calculate the weights that the decision maker should use. If the calibration score of the expert *e* is at least as high as a threshold value (α), then the expert's weight is obtained by $w(e) = Cal(e) * Inf(e)$. If the experts' calibration is less than the threshold value (α), then the expert's weight is set to zero, which is a situation which commonly occurs for a substantial portion of experts in practical applications.

The threshold value α corresponds to the significance level for rejection of the hypothesis that the expert is well calibrated. The best value for α is identified by resolving the value that would optimize a virtual decision maker. This virtual decision maker combines the experts' answers (probability distributions) based on the weights obtained at the chosen threshold value (α). The optimal level for α is where this virtual expert would receive the highest possible weight, if it were added to the expert pool and had its calibration and information scored as the actual experts.

When α has been resolved, the normalized value of the experts' weights *w(e)* are used to combine their estimates of the uncertain quantities of interest.

4. Data collection method

This section presents how the data was collected, by explaining: what population and sample of experts were chosen, how the measurement instrument was developed and tested, how seed questions for Cooke's classical method were assessed, and the result from applying Cooke's classical method.

4.1 The domain experts

Because this research aims to identify quantities related to IDSs, the respondents needed to demonstrate both the ability to evaluate aspects in the domain, as well as the ability to reason in terms of probabilities. In terms of the expert categories described in (Weiss & Shanteau, 2003), individuals that are expert judges are desirable. Studies of experts' calibration have concluded

that experts are well calibrated in situations with learnability and with ecological validity (Bolger & Wright, 1994). Learnability comes with models about the domain, the possibility to express judgment in a coherent quantifiable manner, and the opportunity to learn from historic predictions and outcomes. Ecological validity is present if the expert is used to making judgments of the type they are asked about in the survey.

Respondents that have had the opportunity to learn the effectiveness of IDSs are likely to be those that have performed tests on different solutions in a quantifiable manner. Researchers in the intrusion detection field have performed and disseminated a number of empirical studies related to effectiveness of different solutions. While these studies are questionable with respect to generality (see (McHugh, 2000)), they do however offer input to specific scenarios. Practitioners (e.g., system administrators) will probably not have the same opportunity to learn the effect of different scenarios, as they typically only have experience from a few installations, and rarely perform stringent evaluations of effectiveness. Furthermore, researchers are more used to estimating probability distribution and reason in terms of probabilities, and will thus provide a better ecological validity. For these reasons, IDS researchers were chosen as respondents for the survey.

To identify IDS researchers, articles published in the SCOPUS database (Elsevier B.V., 2011) between January 2005 and September 2010 were reviewed. Authors who had written articles about information technology with "intrusion detection" in the title, abstract or keywords were then identified. If their contact information could be found, they were added to the list of potential respondents, resulting in a sample of 13,561 respondents. After reviewing respondents with respect to their research topic, and the availability of their contact information, a sample of 6,269 individuals was identified. Of these, the contact information of approximately 1,550 turned out to be incorrect, or out-dated. A pilot study involving 500 respondents (described in 0) reduced the number of respondents who received the final survey to approximately 4,200 individuals.

Out of approximately 4,200 researchers invited to participate in the survey, 1,355 opened it, and 243 submitted answers to the survey's questions. A response rate of this magnitude is to be expected of a slightly more advanced survey. As recommended by (Cavusgil & Elvey-Kirk, 1998), motivators were presented to the respondents invited to participate in the survey, namely: (i) helping the research community as whole, (ii) the possibility to win a gift certificate for literature, and, (iii) being able to compare their answers to other experts after the survey was completed. A number of respondents provided

input for less than half of the questions, i.e., they answered with the pre-set background measure on more than half of the questions. These were excluded from further analysis, resulting in 165 usable surveys that were completed by IDS researchers.

4.2 Elicitation instrument

A web survey was used to collect the probability distributions from the invited respondents. The survey comprised four parts, each beginning with a short introduction to the section. Firstly, the respondents were given an introduction to the survey which explained the purpose of the survey and its outline. In this introduction they also confirmed that they were the person who had been invited, and provided information about themselves, e.g., years of experience in the field of research. Secondly, the respondents received training regarding the answering format used in the survey. After confirming that this format was understood, the respondents then proceeded to the third part. In the third part, both the seed questions and the questions of the study were presented to the respondents. Finally, the respondents were asked to provide qualitative feedback on the survey and the variables covered by it.

The questions in Section Three of the survey were each described through a scenario entailing a number of conditions. Scenarios and conditions for the seed questions can be found in Table 2; scenarios and conditions for the questions at issue in this study are described in Table 3. For each scenario, the respondent was asked to provide a probability distribution that expressed their belief. This probability distribution was specified by setting the 5^{th} percentile, the 50^{th} percentile (the median), and the 95^{th} percentile for the probability distribution. In the survey, the respondents specified their distribution by adjusting sliders, or by entering values, to draw a dynamically updated graph over their probability distribution. The three points specified by the respondents defines four intervals over the range [0 to 100]. The graphs displayed the probability density as a histogram, which was instantly updated upon the change of input values.

The use of graphical formats is known to improve the accuracy of elicitation (Garthwaite, Kadane, & O'Hagan, 2005). Figures and colours were also used to complement the textual questions and to make the questions easier to understand. In Figure 1, the format presented to the respondents is exemplified. A more thorough description of the survey instrument is given in the Appendix.

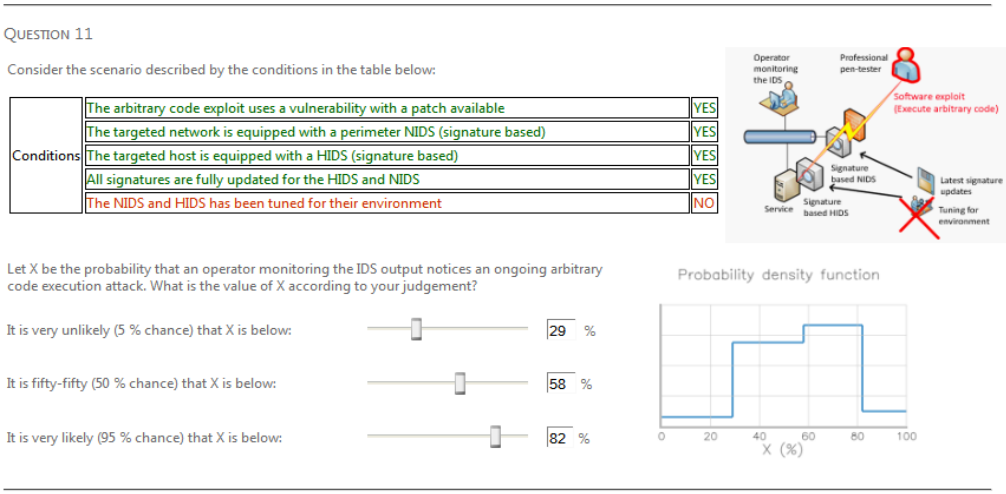

Figure 2. An example of question and answering format in the survey

Elicitation of probability distributions is associated with a number of issues (Garthwaite et al., 2005). Effort was therefore spent on ensuring that the measurement instrument maintained sufficient quality. The survey was, after careful construction, qualitatively reviewed during personal sessions with two external respondents who were representative of the population. These sessions contained two parts. Firstly the respondents were given the task to fill in the survey, and were given the same amount of information as someone doing so remotely. After this, discussions followed regarding instrument quality. These sessions resulted in several improvements with respect to language and the phrasing of questions.

However, the main part of the instrument review took place during the next phase: a pilot study using a randomized sample of 500 respondents from the previously mentioned 6,269 screened subjects. This pilot survey was opened by 123 persons, and completed by 34 during the week it was opened. Cronbach's alpha (Cronbach & Shavelson, 2004; Cronbach, 1951) is often used to test the reliability of a survey instrument and whether respondents understand its questions. A reliability test using Cronbach's alpha was carried out, using one variable (four different versions of the fourth seed question). Measuring the reliability of more than one question was considered unnecessary, as all sections and questions were formatted in the same way. Results from this test showed a reliability value of 0.817, which indicates good internal consistency of the instrument. Qualitative comments also confirmed that respondents understood the questions. A few possible improvements

were identified however. After these changes had been implemented, the survey was again qualitatively reviewed by the two external reviewers.

4.3 Seed questions

In this study, Cooke's classical method is used to synthesize experts' judgements. This method assigns weight to the experts, based on their calibration and the information score to the seed questions. As an expert's performance on answering the seed questions is used to weight them, it is critical that the correct answer to seeds are known, and that they lie in the same domain as the studied variables. Thus, the seeds should represent the truth, and it should be difficult to tell them apart from the questions in the study. However, they do not necessarily need to be directly related to questions of the study (Cooke, 1991).

Naturally, the robustness of the weights attributed to individual experts depends on the number of seeds used. Experience shows that around eight seed questions are enough to see a substantial difference in calibration (Cooke, 1991).

For this study, two types of seed questions were used (cf. Table 2). The first type (questions 1-3) concerned the detection rate of different IDS products when faced with a seven types of commands produced with Nmap, which is a network discovery tool. The actual detection rates (the realization values) were drawn from an empirical test described in (Ktata et al., 2009). The second type of seed questions (4-8) concerned the coverage of software vulnerabilities in the IDS ruleset maintained by the Sourcefire Vulnerability Research Team. This ruleset is used in the popular signature-based IDS product Snort, amongst others. Statistics on how well this ruleset covered vulnerabilities in different products and timeframes was obtained by cross referencing this ruleset's coverage to the National Vulnerability Database (NIST Computer Security Resource Center (CSRC), 2011). The Common Vulnerability Scoring System (CVSS) (Mell, Scarfone, & Romanosky, 2007) is a well-established system for rating a software vulnerability's severity. Vulnerabilities rated with high severity, according to the Common Vulnerability Scoring System (CVSS) (Mell et al., 2007), were used, as such vulnerabilities are those that could be used for arbitrary code exploits.

#	Question	Realization (%)
1	If one of the seven NMAP commands was randomly selected and then executed, how probable do you think it is that a default configured Snort intrusion detection system would detect it?	72
2	If one of the seven NMAP commands was randomly selected and then executed, how probable do you think it is that a default configured Tamandua intrusion detection system would detect it?	29
3	If one of the seven NMAP commands was randomly selected and then executed, how probable do you think it is that a default configured Firestorm intrusion detection system would detect it?	29
4	Consider vulnerabilities of high severity (according to CVSS) that impact Windows 7 and were published during 2010. What proportion of these vulnerabilities has a corresponding signature in Snort's default ruleset?	40
5	Consider vulnerabilities of high severity (according to CVSS) that impact MySQL and were published during 2004-2009. What proportion of these vulnerabilities has a corresponding signature in Snort's default ruleset?	87
6	Consider vulnerabilities of high severity (according to CVSS) that impact Windows 7 and were published during 2009. What proportion of these vulnerabilities has a corresponding signature in Snort's default ruleset?	37
7	Consider vulnerabilities of high severity (according to CVSS) that impact Windows 7 and were published during the last 6 months. What proportion of these vulnerabilities has a corresponding signature in Snort's default ruleset?	35
8	Consider vulnerabilities of high severity (according to CVSS) that impact Samba and were published during 2010. What proportion of these vulnerabilities has a corresponding signature in Snort's default ruleset?	33

Table 2. Seed questions used in abbreviated format. The seven NMAP commands can be found in (Ktata et al., 2009).

A threat to the validity of this study is the fact that these sources are also available to the respondents, who could have used them identify the answers to the seed questions. However, it appears unlikely that any of them did so.

None of the respondents to the survey gave comments that indicated that they had realized that the correct answer could be found in this way, and neither did the qualitative reviewers realize this during the dry runs. Further analysis of the answers received did not show any answers based on these sources. Naturally, Ktata et al. were excluded from the list of potential respondents.

4.4 Respondents' performance

The weight was calculated from the answers of each respondent to the seed questions. All 165 of them completed the survey in less than one hour. As in many other studies involving expert judgment, many of the experts were poorly calibrated. Their calibration score varied between $2.200*10^{-10}$ and 0.6638, with a mean of 0.1575 and their information score varied between $8.620*10^{-7}$ and 3.293, with a mean of 0.8630. Figure 2 shows the information and calibration scores of the respondents (c.f. Section 3 for an explanation of these values).

Cooke's classical method aims to identify those respondents whose judgment is well calibrated and informative. The virtual decision maker was optimized at a significance level (α) of 0.6638. Consequently, the 12 rightmost respondents in Figure 2 received a weight higher than zero, and the other 153 respondents received a weight of zero. As noted above, it is not uncommon for a substantial number of respondents to receive the weight of zero with this method.

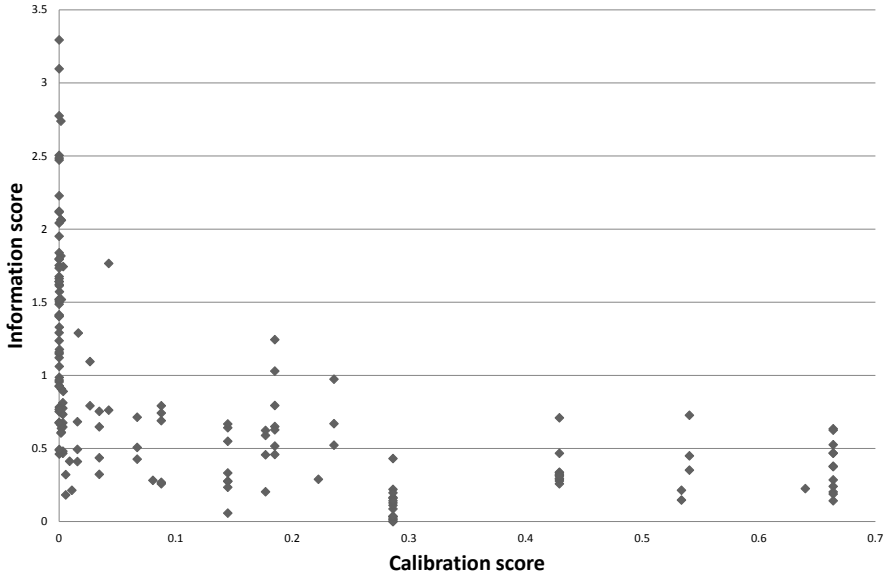

Figure 3. Information and calibration scores of the respondents.

The twelve respondents who received a positive weight all had the same calibration score (0.6638). Their weights are therefore directly proportional to their information score (cf. Section 0). They received weights of between 0.0313 and 0.1401 after normalization.

5. Results

This section presents the results of the analysis performed on the judgment of the 165 researchers. The synthesized estimates of those respondents who were assigned weights are presented in Section 0. In Section 0 the influence that each of the five individual variable has on effectiveness is described.

5.1 Detection rate in the scenarios

To identify the probability distribution which the virtual decision maker assigns to the effectiveness in the 24 scenarios, the individual estimates were combined using their weights. The estimated distributions were assumed to be distributed in the same way as they were presented to the respondents (c.f. Section 0), i.e., as depicted in the histograms over the four ranges that they constructed with their answers.

As depicted in Table 3, the synthesized estimates show clear differences among the scenarios. The median for the scenarios varies between 32% and 65%, the value at the 5th percentile varies between 2% and 13% and the value at the 95th percentile varies between 80% and 97%. Scenario 1 (where all variables are true) has the highest median (65%) and mean (58%) effectiveness. Scenario 17, which is the same as Scenario 1, but without the network IDS, is the second most effective, judging from the median (63%) and mean (55%). Scenarios 4, 6, 8, 12, 14, 16 and 24 are at the other end of the scale, with medians or means of 40 % or below.

Scenario	NIDS	HIDS	Patchable	Updated	Tuned	Low (5%)	Median (50%)	High (95%)	Expected
1	Y	Y	Y	Y	Y	13	65	91	58
2	Y	Y	Y	Y	N	8	43	93	48
3	Y	Y	Y	N	Y	12	59	96	54
4	Y	Y	Y	N	N	5	39	82	41
5	Y	N	Y	Y	Y	6	48	91	47
6	Y	N	Y	Y	N	6	38	91	41
7	Y	N	Y	N	Y	8	44	88	44
8	Y	N	Y	N	N	4	32	92	39
9	Y	Y	N	Y	Y	9	51	91	48
10	Y	Y	N	Y	N	8	45	89	43
11	Y	Y	N	N	Y	10	49	90	46
12	Y	Y	N	N	N	2	39	80	38
13	Y	N	N	Y	Y	2	40	90	41
14	Y	N	N	Y	N	7	37	85	38
15	Y	N	N	N	Y	10	42	88	42
16	Y	N	N	N	N	2	39	93	43
17	N	Y	Y	Y	Y	8	63	94	55
18	N	Y	Y	Y	N	7	51	91	50
19	N	Y	Y	N	Y	9	53	92	50
20	N	Y	Y	N	N	4	48	97	47
21	N	Y	N	Y	Y	8	50	89	45
22	N	Y	N	Y	N	7	48	87	44
23	N	Y	N	N	Y	9	51	92	48
24	N	Y	N	N	N	2	40	84	40

Table 3. The scenarios, their variable configuration and their estimated effectiveness.

5.2 Variables' influence on the effectiveness of intrusion detection

This study identified five variables as being relevant to effectiveness through the literature and by the interviews with domain experts. The variation over scenarios on effectiveness supports this hypothesis. A relevant question is then, how important are these variables for the IDSs' effectiveness and do certain variable combinations have a particular effect, i.e., if the variables are independent or interact. Table 4 shows the mean influence that the five variables have on probability distribution. It also shows the variable interactions that have the highest influence on effectiveness.

The values in Table 4 show the weight of variables alone, and also in combination, calculated as in a full factorial experiment (Montgomery, 2008). These calculations are made under the assumption that the effectiveness is zero, without either a HIDS or NIDS. The values thus represent the mean influence a variable, or variable combination, has on effectiveness. For instance, the values for NIDS are obtained as follows: $\frac{1}{16}\sum_{i=1}^{16} Scenario_i - Scenario_{i+16}$, where scenario 25-32 have zero values (there is no detection system in place).

As can be seen in Table 4, the variables with the highest influence are the NIDS and HIDS, i.e., to actually have an IDS. A NIDS does on average increase the expected effectiveness, with 20.75 percentiles, whilst a HIDS increases the expected effectiveness by 26.25 percentiles. The relatively high influence of these variables should be seen in the light of the fact that without them, effectiveness is zero. Given that a NIDS or HIDS is in place, the most rewarding change is to tune the intrusion detection solution to its environment (4.125 percentiles). If the vulnerability that is exploited is patchable the expected effectiveness increase by 3.625 percentiles and if the IDS has the latest signatures the expected increase is 1.625 percentiles.

	Low (5%)	Median (50%)	High (95%)	Expected value
NIDS	3.625	19.125	44	20.75
HIDS	4.75	29.625	45	26.25
Tuned	2.625	7.25	1.75	4.125
Updated	0.75	2.75	0.5	1.625
Patchable	0.875	3.25	2.5	3.625
NIDS & HIDS	-2	-20.875	-45.75	-21.125
NIDS & Tuned	0.875	3.5	0.75	2
HIDS & Updated	1.125	2	1	1.75
HIDS & Patchable	0.5	2.75	1.75	2.75
Patchable & Updated	0	1.375	0	1.375

Table 4. The influence the strength of individual variables and selected variable combinations.

As can be seen from Table 4, the combination of a NIDS and HIDS has a substantial negative impact on effectiveness. The interaction even exceeds the positive influence a NIDS has on expected effectiveness. In other words, having both a NIDS and a HIDS is on average less effective than only having a HIDS. Looking at the scenarios in Table 3, the negative numbers can be explained by the comparison of scenarios where no tuning has been made to the solution, i.e., if an untuned NIDS is removed, and only an untuned HIDS is used, then effectiveness increases. The negative value resulting from this interaction also exceeds the positive value a HIDS has on the 95th percentiles. The explanation for this negative influence can also be found in conjunction with untuned solutions. When the solution is neither updated, nor tuned (as in Scenario 4 and 12), then the 95th percentile's value increases, if the host based component is removed, given that a NIDS is in place.

Other variables also interact, but to a lesser extent in absolute numbers. Table 4 shows those interactions that have influences greater than 1.25 percentiles (positive or negative) on the expected effectiveness. As can be seen, tuning appears to be of particular importance in the case where a NIDS is used. The expected value on effectiveness then increases by two percentiles, in addition to the 4.125 percentiles that tuning otherwise add, i.e., tuning adds 50 percent

more effectiveness if a NIDS is used. For HIDS, signatures that are updated increase the expected effectiveness by an extra 1.75 percentiles, and HIDS also appears to be more helped by a scenario where the exploited vulnerability is possible to patch (i.e., is well known). The interaction is 2.75 percentiles between updates and vulnerability-type. The positive interaction between updating a system and being attacked with known (patchable) exploits is intuitive – updates can be expected to have a limited impact on effectiveness against new attacks (which there is seldom a patch for).

6. Discussion

The outline of the discussion is as follows. Section 6.1 discusses the validity and reliability of the survey as experts' judgments and also as a knowledge elicitation instrument. Section 6.2 gives recommendations to practitioners, based on the research findings, and Section 6.3 gives recommendations for future research.

6.1 Validity and reliability

Estimation of probabilities is known to be difficult. Estimation of a set of variables' influence on an outcome's probability is likely to be even more challenging in most situations. In this research, we defined a number of operational scenarios that describe fairly concrete cases with variables of interest in a well-defined state. We believe that this can remove a significant cognitive burden from the respondents, as they do not assess variables' influence directly. In other words, respondents only need to relate to real cases that they have experience of, and does not require them to consider variables' interdependencies and relative frequency. Furthermore, it likely that use of operational scenarios removes cognitive biases in the responses. More specifically, the use of operational scenarios does not require the respondents to compare different scenarios, and there is little risk that their answers are influenced by the base rate fallacy described by Kahneman and Tversky (1973).

When it comes to synthesis of expert judgments, this study used Cooke's classical method (Cooke, 1991). This performance-based method aims to select the experts that are well calibrated and to combine their judgments in an optimal way. The track record of this method (Cooke, 2008) positions it as a best-practice when it comes to eliciting expert judgment of uncertain quantities.

Cooke (1991) provides a list of guidelines for how to elicit data from experts: 1) questions must be clear and unambiguous, 2) a dry run should be carried out before the actual study, 3) an attractive graphical format should be used and there should be a brief explanation of the elicitation format, 4) elicitation should not exceed one hour , 5) coaching should be avoided, and, 6) an analyst should be present when respondents answer the questions. As described in Section 0, all guidelines except for 6) are met in this study, i.e., no analyst were present when respondents answered the questions. With a web survey, this was obviously not possible. The respondents were given the contact information of the research group when they were invited to the survey, and were encouraged to contact them if any questions arose. While this ensures that no coaching occurred during the elicitation, it is possible that it suppressed potential questions from being asked. To identify potential issues of this type, the respondents were asked to comment on the clarity of the questions and the question format used. Based on the comments received, no distressing issues relating to the formulation of the questions arose. Several respondents did, however, comment on the difficulty of expressing knowledge quantitatively, or the difficulty in estimating the effectiveness of IDSs in general (as there little empirical data on it). However, this issue is not surprising, and is a part of the reason why this study was carried out in the first place.

When using Cooke's classical method, it is appropriate to perform a robustness test with respect to the seed variables and the experts, by removing one expert and investigating the impact of this removal (Cooke, 1991). Such tests were performed and they indicate that the solution is robust to changes, with regards to both seed questions and experts. However, the answers to the seed questions show that many experts in the intrusion detection field are poorly trained in calibration (as in many other domains), i.e., their estimates do not match empirical observations well. This can be seen by the calibration scores to the seed questions used in this study (c.f. Table 2), and show the importance of assigning different weights to experts' judgment. Twelve respondents were assigned a weight when the virtual decision maker was optimized. The estimates from the twelve respondents was relatively uninformative when compared to the respondents' estimates overall. This should not be seen as surprising. Overconfidence is a well-known cause for poor calibration in expert judgments (Lin, 2008). Nevertheless, this uncertainty suggests that the research community's knowledge about effectiveness of IDS is lacking.

The cost of obtaining observational data on the effectiveness of operational IDSs (where administrators use the system) was the main motivation for the use of IDS experts' judgment to cover the broad scope of this study. The only

observational data about this that was found in the literature is the one described by Sommestad and Hunstad (2013). Although extensive efforts were made to organize this experiment (e.g., construction of fictive networks, the installation and tuning of an IDS, as well as time spent by attackers and administrators), it is associated with several assumptions and delimitations which threaten the representativeness of the result. It roughly corresponds to Scenario 1, which the experts in this study assessed as being the most ideal scenario. The experiment gave an effectiveness of 58% and a mean value predicted by the domain experts of 59% (cf. Table 3). Thus, the experiment (executed after this expert survey) corroborates the experts' assessment.

6.2 Recommendations for information system decision-makers

From a practitioners' point of view, these results provide input as to which actions should be taken in order to use an IDS effectively.

Firstly, the results show that experts are uncertain about IDS effectiveness, and that many of them are poorly trained in calibration (incorrect and uncertain) of the test questions used to weight them. In other words, if a decision maker would ask a randomly selected IDS expert for advice, they are likely to receive vague or incorrect suggestions, and if multiple experts are asked for advice, then their recommendations will probably differ. This study synthesized the judgment of a large number of security experts (from whom those with more experience of calibration have been carefully selected). The synthesized results are uncertain, but it is unlikely that the decision maker can gain more precise knowledge (at this level of abstraction) from a random security expert, or a random set of security experts. Furthermore, knowing the uncertainty of the effectiveness in an IDS scenario will help the decision maker to make informed decisions and to appreciate the effectiveness of those countermeasures that are not covered by this study.

Secondly, tuning the IDS to its environment is expected to increase the detection rate. However, tuning an IDS in an enterprise context is a continuous process: as soon as there has been a change in any parameter that is under surveillance, the IDS needs to be tuned to reflect this change. For example, if the organization has installed a new FTP server, or bought new computer systems, then traffic patterns will change and the IDS will need to be tuned again. Since tuning requires constant adaptation of the IDS it will require that system administrators regularly spend time analysing recent changes to the enterprise system architecture, and that they adapt the IDS accordingly (Scarfone & Mell, 2007) (Werlinger et al., 2008). Of course, these costs can be neglected if the IDS is deployed in a static and documented environment, e.g., in an industrial facility's control system network.

Thirdly, if the IDS uses the most recent ruleset, then the effectiveness will also increase. In comparison to tuning, keeping the IDS updated with a recent ruleset is a relatively straightforward process, which does not require administrators to analyse the current architecture or to spend significant effort in programming the IDS solution. On the other hand, adherence to new rules is often associated with some cost, and the impact on effectiveness is considerably less than that of tuning.

Fourthly, host-based solutions (HIDS) give better effectiveness than network-based solutions (NIDS). However, a problem of HIDSs is that they are required to be implemented at a host-level, which could involve significant costs. For example, each HIDS might have to be manually installed in each supervised system, and perhaps have to the tuned manually for the context of each such system. A NIDS-solution is not as effective as a HIDS-solution. As such, a cost effective architecture is likely to use a HIDS solution on the most sensitive systems in the enterprise, and a NIDS solution to monitor less sensitive systems. For instance, a HIDS solution could be used to monitor critical business servers, and a NIDS solution could be used to monitor office clients.

Fifthly, combined solutions (with both HIDS and NIDS) are not recommended. They have been presented in literature as a way to increase effectiveness, however, the results from this study suggest the opposite – a combination of a HIDS and a NIDS is not believed to increase the effectiveness of intrusion detection. In fact, if a HIDS is already in use, then experts believe that the effectiveness will decrease if an NIDS is also installed. One reason behind this could be the fact that the output and the HIDS will overlap, and that large amounts of information (and false alarms) need to be processed by the administrator in order to detect attacks with multiple sensors.

Sixthly, signature-based systems can also detect novel attacks. An interesting result of this study is that the possibility to patch exploited vulnerability has the lowest impact of the assessed variables. This suggests that signature-based systems can detect novel attack-types, just as anomaly based systems can.

Finally, organizational decision makers should reflect on whether IDSs really are needed in their environments. This study show that such tools are believed to only provide modest effectiveness, and that it is costly to implement and maintain an IDS solution. The tools do not only require technical costs (installation/maintenance), but also investment in time by network administrators, who need to carefully study the output of the solution to be able to detect real attacks.

JISSec
Journal of Information System Security

6.3 Recommendations for researchers and the theoretical contribution

Observational studies and experiments are costly to perform and should therefore be carefully planned. This study identifies a number of variables and variable-interactions that are believed to be important by a carefully selected group of domain experts. The results should be interpreted cautiously, as the study is based on experts' estimates, and these experts express a great deal of uncertainty. Their uncertain opinions also suggest either: a) a lack of knowledge regarding the effectiveness of IDSs, or, b) that important conditions (i.e., variables) are missing in the operational scenarios used in this study. Based on the responses to this survey, the latter appears more likely. Only a small portion of the respondents were able to identify more important variables than those that were already included. These suggested that anomaly-based intrusion detection should be added and that further details on the variable relating to the exploited vulnerability type were desirable.

Based on the results from this expert study, it is possible to make broader hypotheses about the relationships between effective IDS and the included variables. We present three hypotheses below.

Firstly, there is a widespread belief that the combination of a NIDS and a HIDS will create an effective IDS solution. For instance, in SANS FAQ on intrusion detection, the conclusion is that "[a] truly effective IDS will use a combination of network and host-based intrusion detection" (Zirkle, 2008). On the contrary, our expert survey suggests that the combination of a HIDS and a NIDS will decrease the effectiveness under certain circumstances. We suggest that this could be because a combined solution will flood the operator with false alarms. Support for this hypothesis can be found in the interaction with tuning, which is known to decrease false alarms. Combined solution performs significantly worse in just those operational scenarios where the solution has not been tuned (Scenarios 2 vs. 18, 4 vs. 20, 10 vs. 22 and 12 vs. 24).

Secondly, the literature often labels signature-based solutions as being largely ineffective in scenarios with attacks that use zero-day exploits (see for example (Wang, Cretu, & Stolfo, 2006) or (Kanoun, Cuppens-Boulahia, Cuppens, Dubus, & Martin, 2009)). Our results suggests otherwise, showing a decreased effectiveness of only 3.6 percentiles when an unpatchable vulnerability is attacked. One explanation is the similarities between the exploit-code used for different vulnerabilities. In other words, a new exploit-type could utilize an attack vector which already has a signature in the IDS ruleset. For instance, a buffer overflow using a NOOP-sled can be detected

through a long sequence of characters, no matter whether it is a zero-day vulnerability, or not (I. Kim et al., 2009). A recent study by Holm (2014) provides further support for this theory.

A third interesting finding is that the variables are all rather independent (except for the usage of HIDS and NIDS together). This suggests that future research could be reasonably optimized in each variable domain independently. In other words, experiments concerning effectiveness can be focussed on one variable at a time, and the consideration of interactions is not crucial.

7. Conclusion

Reliable data on intrusion detection effectiveness from observations or experiments is not available. The synthesized judgment of researchers in the intrusion detection field shows a great deal of uncertainty when estimating the effectiveness of IDSs for different scenarios. Some of this uncertainty stems from natural variation between enterprises, but it appears reasonable that a portion also come from epistemic uncertainty and is strongly related to the lack of empirical studies in this field, i.e., the community is not certain as to how well intrusion detection actually works.

This study provides indicators for the effectiveness of intrusion detection in different scenarios. In particular, host-based solutions are associated with higher effectiveness than network-based ones. Furthermore, tuning is a measure with a comparably high impact on effectiveness, yet it is not of great importance for effectiveness, if the vulnerability exploited is well-known and patchable, or even if it is not. These quantitative results are based on the synthesized judgment of researchers in the field and indicate the importance of different variables and the effectiveness of solutions as a whole.

8. References

Abdolmohammadi, M. J., & Shanteau, J. (1992). Personal attributes of expert auditors. *Organizational Behavior and Human Decision Processes*, 53(2), 158–172.

Alserhani, F., Akhlaq, M., Awan, I. U., Mellor, J., Cullen, A. J., & Mirchandani, P. (2009). Evaluating Intrusion Detection Systems in High Speed Networks. *2009 Fifth International Conference on Information Assurance and Security*, 454–459. doi:10.1109/IAS.2009.276

Anderson, J. P. (1980). *Computer security threat monitoring and surveillance*. Forth Washington: Technical report, James P. Anderson Company, Fort Washington, Pennsylvania.

Ashfaq, A., Robert, M., Mumtaz, A., Ali, M., Sajjad, A., & Khayam, S. (2008). A comparative evaluation of anomaly detectors under portscan attacks. In *Recent Advances in Intrusion Detection* (pp. 351–371). Springer. Retrieved from http://www.springerlink.com/index/x8643207t2174l34.pdf

Ashton, A. H. (1985). Does consensus imply accuracy in accounting studies of decision making? *The Accounting Review*, 60(2), 173–185.

Axelsson, S. (2000a). Intrusion detection systems: A survey and taxonomy. Technical Report (Vol. 99, pp. 1–15). Göteborg, Sweden.

Axelsson, S. (2000b). The base-rate fallacy and the difficulty of intrusion detection. *ACM Transactions on Information and System Security*, 3(3), 186–205. doi:10.1145/357830.357849

Barry, B. I. A., & Chan, H. A. (2010). Intrusion detection systems. In P. Stavroulakis & M. Stamp (Eds.), *Handbook of Information and Communication Security* (Vol. 2001, pp. 193–205). Springer. doi:10.1016/S1361-3723(01)00614-5

Biermann, E. (2001). A comparison of Intrusion Detection systems. *Computers & Security*, 20(8), 676–683. doi:10.1016/S0167-4048(01)00806-9

Bolger, F., & Wright, G. (1994). Assessing the quality of expert judgment: Issues and analysis. *Decision Support Systems*, 11(1), 1–24. doi:10.1016/0167-9236(94)90061-2

Cavusgil, S. T., & Elvey-Kirk, L. A. (1998). Mail survey response behavior: A conceptualization of motivating factors and an empirical study. *European Journal of Marketing,* 32(11/12), 1165–1192. doi:10.1108/03090569810243776

Clemen, R. T., & Winkler, R. L. (1999). Combining probability distributions from experts in risk analysis. *Risk Analysis*, 19(187), 187–204.

Cooke, R. M. (1991). Experts in Uncertainty: *Opinions and Subjective Probability in Science*. New York, New York, USA: Open University Press.

Cooke, R. M. (2008). TU Delft expert judgment data base. *Reliability Engineering & System Safety*, 93(5), 657–674. doi:10.1016/j.ress.2007.03.005

Cooke, R. M., & Goossens, L. (2004). Expert judgement elicitation for risk assessments of critical infrastructures. *Journal of Risk Research*, 7(6), 643–656. Retrieved from http://www.ingentaconnect.com/content/routledg/rjrr/2004/00000007/00000006/art00008

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297–334. doi:10.1007/BF02310555

Cronbach, L. J., & Shavelson, R. J. (2004). My Current Thoughts on Coefficient Alpha and Successor Procedures. *Educational and Psychological Measurement*, 64(3), 391–418. doi:10.1177/0013164404266386

Denning, D. E. (1987). An Intrusion-Detection Model. *IEEE Transactions on Software Engineering*, SE-13(2), 222–232. doi:10.1109/TSE.1987.232894

Elsevier B.V. (2011). Scopus. Retrieved from http://www.scopus.com/

Faysel, M. A., & Haque, S. S. (2010). Towards Cyber Defense : Research in Intrusion Detection and Intrusion Prevention Systems. *Journal of Computer Science*, 10(7), 316–325.

Fink, A., Kosecoff, J., Chassin, M., & Brook, R. H. (1984). Consensus methods: characteristics and guidelines for use. *American Journal of Public Health*, 74(9), 979–983. doi:10.2105/AJPH.74.9.979

Garciateodoro, P., Diazverdejo, J., Maciafernandez, G., & Vazquez, E. (2009). Anomaly-based network intrusion detection: Techniques, systems and challenges. *Computers & Security*, 28(1-2), 18–28. doi:10.1016/j.cose.2008.08.003

Garthwaite, P. H., Kadane, J. B., & O'Hagan, A. (2005). Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association,* 100(470), 680–701.

Goodall, J. R., Lutters, W. G., & Komlodi, A. (2009). Developing expertise for network intrusion detection. *Information Technology & People*, 22(2), 92–108. Retrieved from http://www.emeraldinsight.com/journals.htm?articleid=1793305&show=abstract

Holm, H. (2014). Signature Based Intrusion Detection for Zero-Day Attacks: (Not) A Closed Chapter? In *2014 47th Hawaii International Conference on System Sciences* (pp. 4895–4904). Big Island, HI, United states: IEEE. doi:10.1109/HICSS.2014.600

Holm, H., Sommestad, T., Ekstedt, M., & Honeth, N. (2013). Indicators of expert judgement and their significance: an empirical investigation in the area of cyber security. *Expert Systems, (Accepted)*, n/a–n/a. doi:10.1111/exsy.12039

Itoh, T., Takakura, H., Sawada, A., & Koyamada, K. (2006). Visualization of Network Intrusion Detection Data. *IEEE Computer Graphics and Applications*, 26(2), 40–47.

Julisch, K., & Dacier, M. (2002). Mining intrusion detection alarms for actionable knowledge. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 366–375). New York, New York, USA: ACM. doi:10.1145/775094.775101

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, 80(4), 237–251. doi:10.1037/h0034747

Kanoun, W., Cuppens-Boulahia, N., Cuppens, F., Dubus, S., & Martin, A. (2009). Success Likelihood of Ongoing Attacks for Intrusion Detection and Response Systems. *2009 International Conference on Computational Science and Engineering*, 83–91. doi:10.1109/CSE.2009.233

Krayer von Krauss, M. P., Casman, E. a, & Small, M. J. (2004). Elicitation of expert judgments of uncertainty in the risk assessment of herbicide-tolerant oilseed crops. Risk Analysis : *An Official Publication of the Society for Risk Analysis,* 24(6), 1515–27. doi:10.1111/j.0272-4332.2004.00546.x

Ktata, F. B., Kadhi, N. El, & Ghédira, K. (2009). Agent IDS based on Misuse Approach. *Journal of Software*, 4(6), 495–507. doi:10.4304/jsw.4.6.495-507

Lin, S. (2008). A study of expert overconfidence. *Reliability Engineering & System Safet*y, 93(5), 711–721. doi:10.1016/j.ress.2007.03.014

McFadzean, E., Ezingeard, J.-N., & Birchall, D. (2011). Information Assurance and Corporate Strategy: A Delphi Study of Choices, Challenges, and Developments for the Future. *Information Systems Management*, 28(2), 102–129. doi:10.1080/10580530.2011.562127

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), 262–294. doi:10.1145/382912.382923

Mell, P., Hu, V., Lippmann, R., Haines, J. W., & Zissman, M. (2003). A*n overview of issues in testing intrusion detection systems*, (NIST IR 7007). *Citeseer*. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.8.5163

Mell, P., Scarfone, K., & Romanosky, S. (2007). A complete guide to the common vulnerability scoring system version 2.0. *Published by FIRST-Forum of Incident Response and Security Teams*. Retrieved January 09, 2014, from http://www.first.org/cvss/cvss-guide.pdf

Montgomery, D. C. (2008). *Design and analysis of experiments*. Hoboken, NJ: John Wiley & Sons Inc.

NIST Computer Security Resource Center (CSRC). (2011). National Vulnerability Database. Retrieved February 13, 2011, from www.nvd.nist.org

Salah, K., & Kahtani, a. (2009). Improving Snort performance under Linux. *IET Communications*, 3(12), 1883. doi:10.1049/iet-com.2009.0114

Scarfone, K., & Mell, P. (2007). *Guide to intrusion detection and prevention systems*. *Nist Special Publications* (Vol. 800). Gaithersburg, MD, USA.

Shaikh, S., Chivers, H., Nobles, P., Clark, J., & Chen, H. (2008). Characterising intrusion detection sensors. *Network Security, 2008(9)*, 10–12. doi:10.1016/S1353-4858(08)70107-7

Shanteau, J., Weiss, D. J., Thomas, R. P., & Pounds, J. C. (2002). Performance-based assessment of expertise: How to decide if someone is an expert or not. *European Journal of Operational Research*, 136(2), 253–263. doi:10.1016/S0377-2217(01)00113-8

Sommestad, T., & Hunstad, A. (2013). Intrusion detection and the role of the system administrator. *Information Management & Computer Security*, 21(1), 30 – 40. doi:10.1108/09685221311314400

Sumner, M. (2009). Information Security Threats: A Comparative Analysis of Impact, Probability, and Preparedness. *Information Systems Management, 26*(1), 2–12. doi:10.1080/10580530802384639

Thompson, R. S., Rantanen, E. M., & Yurcik, W. (2006). Network intrusion detection cognitive task analysis: Textual and visual tool usage and recommendations. In *Human Factors and Ergonomics Society Annual Meeting Proceedings* (Vol. 50, pp. 669–673). Human Factors and Ergonomics Society. Retrieved from http://www.ingentaconnect.com/content/hfes/hfproc/2006/00000050/00000005/art00011

Thompson, R. S., Rantanen, E. M., Yurcik, W., & Bailey, B. P. (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* (p. 1205). ACM. Retrieved from http://dl.acm.org/citation.cfm?id=1240807

Wang, K., Cretu, G., & Stolfo, S. (2006). Anomalous Payload-Based Worm Detection and Signature Generation. In *Recent Advances in Intrusion Detection* (pp. 227–246). Springer. Retrieved from http://www.springerlink.com/index/75h308806288v3p1.pdf

Weiss, D. J. D. J., & Shanteau, J. (2003). Empirical Assessment of Expertise. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 45(1), 104–116. doi:10.1518/hfes.45.1.104.27233

Werlinger, R., Hawkey, K., Muldner, K., Jaferian, P., & Beznosov, 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), 107–118. Retrieved from http://portal.acm.org/citation.cfm?id=1408679

Xenakis, C., Panos, C., & Stavrakakis, I. (2010). A comparative evaluation of intrusion detection architectures for mobile ad hoc networks. *Computers & Security*, 30(ii), 1–18. doi:10.1016/j.cose.2010.10.008

Young, G., & Pescatore, J. (2009). *Magic quadrant for network intrusion prevention system appliances*. Retrieved from http://www.adexsus.com/v2/pdf/Detectores de Intrusos/Gartner/Cuadrante Magico.pdf

Zirkle, L. (2008). What is host-based intrusion detection? *Intrusion Detection FAQ*. Retrieved from http://www.sans.org/security-resources/idfaq/host_based.php

JISSec
Journal of Information System Security

Appendix – Survey instrument

8.1 Introductory section

The following text introduced the respondents to the survey and explained its purpose.

WELCOME!

This survey focuses on properties related to intrusion detection systems and is distributed to number of selected experts in the research community. The survey comprises 36 questions, where you will be asked to quantify your answer in terms of a probability for each question.

Many of these 36 questions are difficult. They will ask you for probabilities which to some extent are unknown (both to you and the community at large). However, the fact that the probabilities are unknown is also the reason why this survey is being sent out to IDS experts in an attempt to approximate them. If all these probabilities were known, or easy to identify, then there would be no reason to approximate them through domain experts. Consequently, we ask you here (as a domain expert) to provide **your own belief and best guess**.

By completing this survey you will:

- **Help the research community to approximate the quantifiable properties of IDS.**
- **Be able to compare your answers with the answers of other IDS experts.**
- **Have the chance to win a 100 USD gift certificate at Amazon.**

The survey comprise of 6 pages (including this one). You can use your invitation link to return to an uncompleted survey later – you do not have to complete it right now. After completing the survey, you will receive a link that displays your answers and compares them to the aggregate of all answers.

8.2 Seed questions

The seed questions are given in Table 2 of the paper above. They were answered by specifying three probabilities: one for the 5th percentile, one for the 50th percentile, and one for the 95th percentile.

Let X be the probability of detection. What is the value of X according to your judgment?

It is very unlikely (5% chance) that the value is below: [0-100%]

There is a fifty-fifty (50% chance) that the value is below: [0-100%]

It is very likely (95% chance) that value is below: [0-100%]

8.3 Questions of interest

Instructions at the top of each page of questions reminded respondents of the following:

For all questions on this page the attacker should be thought of as:

- *A professional penetration tester, with access to tools that are free and/or commercially available.*
- *An outsider who has spent one week prior in preparing the attack.*

All questions were answered by specifying three probabilities: one for the 5th percentile, one for the 50th percentile, and one for the 95th percentile. The respondent could use a slider or a text field to enter values. When a value was entered, a probability density function next to the input field was updated to reflect how probable different values were, according to the respondents' answers.

Scenarios 1-24 were formulated in the same way, but asked the respondent to answer it under different conditions. These conditions were:

- *The arbitrary code exploit uses as vulnerability with a patch available* [YES/NO]
- *The targeted network is equipped with a perimeter NIDS (signature-based)* [YES/NO]
- *The targeted host is equipped with a perimeter HIDS (signature-based)* [YES/NO]
- *All signatures are fully updated for the HIDS and NIDS* [YES/NO]
- *The NIDS and HIDS have been tuned for their environment* [YES/NO]

All 24 permutations, except those with neither a NIDS or a HIDS was considered. Figure 2 of the paper demonstrates the disposition with a table that contains conditions color-based for their state, a schematic figure of the scenario, the question, and the three probabilities that make up the response. The formulation was:

Let X be the probability that an operator monitoring the IDS output notices an ongoing arbitrary code execution attack. What is the value of X, according to your judgment?

It is very unlikely (5% chance) that X is below: [0-100%]

There is a fifty-fifty (50% chance) that X is below: [0-100%]

It is very likely (95% chance) that X is below: [0-100%]