

Experimental Evaluation of Qualitative Probability applied to Sensor Fusion and Intrusion Detection

Anonymous

Abstract

We experimentally analyze the accuracy of the System $Z+$ qualitative probability scheme of Goldszmidt and Pearl when used for diagnosis and information fusion. The Intrusion Detection System (IDS) fusion system Scyllarus, and its successor MIFD, use $Z+$ to assess the likelihood of various cyber attack events based on reports from IDSes. $Z+$ provides an order of magnitude approximation of conventional probability, similar to the order of magnitude approximation of computational complexity provided by big- O analysis. Scyllarus accurately identifies attacks and substantially reduces the false positives that are the bane of intrusion detection. In the work described here, we experimentally analyze the performance of MIFD in order to provide general conclusions about its behavior, complementing the results from field tests. Our experiments show that the qualitative probability scheme degrades gracefully in precision and recall as its order of magnitude approximation is a less and less accurate representation of true distributions. The system also degrades gracefully as its input sensors become less discriminating. Finally, we show that qualitatively fusing multiple IDSes successfully addresses base rate issues in intrusion detection. The interest of these results is not limited to intrusion detection: the method used in our systems is a general abductive scheme, based on qualitative Bayes networks, so the results are applicable to other information fusion and diagnostic applications. To the best of our knowledge, ours is the only experimental investigation of the accuracy of $Z+$ as an approximation of conventional probability.

Introduction

We experimentally analyze the accuracy of the System $Z+$ qualitative probability scheme when used for diagnosis and information fusion. In cyber defense, an “intrusion detection system (IDS) is a device or software application that monitors network or system activities for malicious activities or policy violations” (Wikipedia 2014). In previous work we developed a technique for IDS fusion, deployed in the Scyllarus system [citation withheld], and its successor MIFD [citation withheld]. These systems fuse together reports from multiple, heterogeneous IDSes, hypothesizing underlying events to explain those reports, and assessing the events’ likelihood, to detect cyber attacks. Their likelihood assessment is based on System $Z+$, a qualitative abstraction of probability theory (Goldszmidt and Pearl 1996). Scyllarus has been extensively tested in real networks, using both

real and synthetic data, and has shown its ability to accurately fuse reports from extremely noisy sensors. We also drastically reduce false alarm rates, the bane of intrusion detection systems, reducing the flood of incoming reports by multiple orders of magnitude. Unfortunately, it is difficult to draw crisp, *general* conclusions based on field evaluations alone. We need to demonstrate that the results are due to features of the algorithm, not simply artifacts of details of the test network, traffic, and attacks. This problem is particularly acute in the area of intrusion detection, as we explain below.

In this paper, we analyze the underlying reasoning machinery to complement earlier field studies. System $Z+$ models uncertain phenomena as falling into a small set of *qualitatively distinct* levels of likelihood, similar to the way that big- O methods abstract computational effort. To use the big- O analogy again, we check to make sure that our results degrade gracefully as the orders of magnitude become less important compared to the constant factors.

The experimental results we report show that System $Z+$ ’s accuracy degrades gracefully as the qualitative abstraction fits less and less well. We also show that the accuracy of our system degrades gracefully with decreasing sensor precision. Finally, MIFD is accurate even when detecting very rare events, not only when sensors fail independently, but also in the face of correlated false positives. These results confirm the results of our earlier field tests, and help explain why the qualitative scheme works so well.

Our results are generally relevant to qualitative probabilistic reasoning for information and sensor fusion. Our fusion problems are modeled as problems of causal explanation, or abduction: what are the events most likely to have caused the observations (the IDS reports) given certain causal relations? Therefore our results are also of interest to researchers in diagnosis. Qualitative probability systems like $Z+$ offer an attractive middle point between purely disjunctive reasoning in diagnosis, and full probabilistic reasoning.

To the best of our knowledge, ours is the only empirical work to explore the *accuracy* of reasoning with System $Z+$. Of about 250 works citing Goldszmidt and Pearl’s 1996 work (Google Scholar 2014), none of the few which detail applications of System $Z+$ reasoning conduct such investigations. Minock and Kraus (2002) investigate the *efficiency* of an implementation of System $Z+$, but not its accuracy.

In the next section we introduce the problem of IDS fusion and its challenges. Then we describe our approach to the problem, as implemented in the Scyllarus and MIFD systems. We describe our experimental designs, present the results and conclude with some proposals for future work.

Intrusion detection fusion

IDS fusion is the problem of creating a coherent overall picture of network status out of reports from multiple IDSes scattered around a computer network. An IDS fusion system must answer two questions:

1. How do the reports issued by the IDSes correspond to events? Multiple reports, from the same IDS or from different IDSes, might actually refer to the same event. This is the problem of *clustering* reports into events.
2. Which hypothesized events do we accept as actual? This is the problem of *assessing* the likelihood of event hypotheses.

In this paper we will focus on the second of these two questions, assuming a solution to the first. For discussions of the first question, see [citation withheld].

Existing IDSes are not designed to work together as part of a suite of sensors. Instead, each program generates a separate, often voluminous, stream of reports; fusing them into a coherent overview is left as an exercise for the user. Ideally, network administrators would have a suite of different IDSes active, since different IDS approaches — network (NIDS) vs. host-based (HIDS), signature vs. anomaly-based — bring different strengths and weaknesses. Briefly, more accurate IDSes have significantly higher software administration burdens, and there is an inevitable tradeoff between false positives and false negatives.

The most substantial challenge in managing IDSes is the information overload that they can impose. Much of this overload comes from the high false positive rate. This rate is partly due to inaccurate sensors. Another problem is the “base rate fallacy” (Axelsson 1999): in intrusion detection, as in other applications where very unlikely events must be detected, even a seemingly accurate sensor may yield too many false positives. Axelsson gives the example of a 99% accurate test detecting a disease that strikes only 1/10,000 patients. The probability of having the disease given a positive test result is only 1%, although the probability of the sensor labeling correctly is 99%. IDS users regularly either ignore or partially disable them, unable to absorb massive streams of false alarms: “Users attending an ‘ABCs of IDS’ event at London’s City University yesterday said more the 80 per cent of the alerts they received were false, with one citing 60 alerts he had received about non-existent problems that morning at 0300” (Leydon 2001). Recently, Target is reported to have ignored warnings about the data breach that resulted in theft of millions of credit cards (Schwartz 2014): “They are bombarded with alerts. They get so many that they just don’t respond to everything” (Finkle and Heavey 2014). Figure 1 gives a sense of the gravity of this problem, showing how our earlier system was able to winnow the flow of reports in a small corporate network [citation withheld]. Data reduction, in addition to building a comprehen-

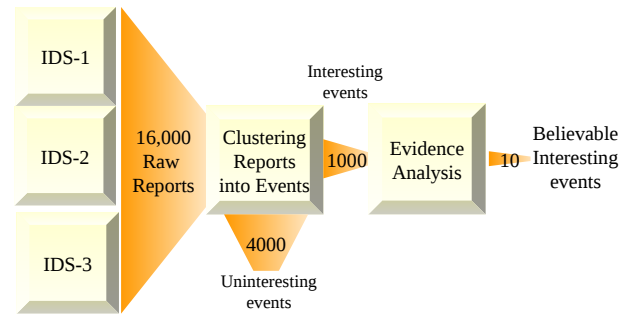


Figure 1: Scyllarus workload reduction [citation withheld].

sive overview and overcoming false positives and negatives, must be a goal of IDS fusion.

Related work. SecurityFocus has developed the Attack Registry and Intelligence Service (2003, ARIS). Its extractor consolidates IDS reports from four different IDSes into XML, and presents them in an incident console. However it makes no attempts to fuse the reports, nor weigh evidence for and against them. MetaSTAT is a fusion system built on a set of STAT-based IDSes (Vigna, Kemmerer, and Blix 2001). STAT is signature-based, and detects events by matching against extended finite-state event models. MetaSTAT uses finite-state models of across-sensor events to consume events from lower-level sensors at a higher level, but does not attempt to judge the plausibility of different events. EMERALD/eBayes (Valdes and Skinner 2001) fusion is the most similar to our systems. Their sensors are Bayes net-based, and the correlation approach allows “upstream” sensors to adjust the priors on “downstream” sensors. Its fusion is limited to clustering alerts that meet a similarity criterion; they do not have models of high-level events like ours (see below). To the best of our knowledge, they do not address the difficult issues of acquiring probability parameters for IDS fusion. The Prelude IDS’s Correlator allows users to analyze reports sent to Prelude from compatible IDSes (Vandoorselaere 2008) following user-provided rules. Its function is closest to our clustering subsystem, but knowledge resides in stateful rules instead of an ontology of attacks. The Arcsight Enterprise Security Manager (ArcSight 2008), also ties correlated IDS reports to an installation’s security goals and vulnerability information.

The Scyllarus and MIFD systems

Our original system Scyllarus was an open system which performed IDS fusion on a diverse set of third party IDSes [citation withheld]. Its successor MIFD (Model-based Intrusion Fusion and Detection), enhances Scyllarus as part of the STRATUS system’s integration of IDS fusion into a comprehensive suite for cyber defense [citation withheld]. For simplicity we refer to our IDS fusion system as “MIFD,” but except where we specify otherwise, our account applies to Scyllarus as well.

The first tasks performed by MIFD are *collection* and *translation*: IDS reports must be collected and presented

to MIFD, and must be translated into a common sensor report data structure for processing. We will not discuss these less interesting preliminaries, except to mention that they involve a substantial amount of data rectification. There is no widely-accepted common data format (the proposed ID-MEF standard (Debar, Curry, and Feinstein 2007) has not caught on), nor is there an accepted taxonomy for events or event roles (e.g., the widely used “source” and “destination” are not consistently applied). These problems are exacerbated by the fact that intrusion detection “systems” are often simply a casually-assembled anthology of rules by different authors running in the same framework.

Core fusion inference begins with **Clustering**: From the sensor reports, MIFD generates *event hypotheses* to represent the underlying events of interest that may have caused the sensor reports. A sensor report that provides evidence for an event hypothesis is a *supporter*. The clustering process constructs a directed graph of event hypothesis and sensor report nodes forming a Bayesian belief network.

Figure 2(a) shows the Bayes net generated by MIFD for the very simple case of a single event hypothesized to be the cause of a number of sensor reports. In general sensor reports will be ambiguous, each supporting several different event hypotheses. Figure 2(b) shows a more complex example where sensor reports might be caused by two different underlying events.

IDS rules detect indirect features of intrusions, but they typically report only the intrusion that their designer believes has caused the feature. This creates a number of problems, first and foremost that the sensors have high false positive rates. For example, a local software update server may look like an attacker performing reconnaissance. What’s worse, some of the false positives are correlated: the normal activity of a print server may be mis-detected as a scan by several different sensors. To handle simple sensor failures, MIFD associates with each sensor a false positive likelihood. But to address the correlated false positives such as the scanning example, we introduce benign events, that can compete with attack and (simple) false positive explanations for results.¹ Another problem with indirect sensing is that the sensors are often imprecise: the features that they detect often occur not only in the absence of the advertised intrusion, but in the presence of *other* intrusions with similar characteristics.

MIFD also has *complex* events that have component sub-events, such as a phishing attack with component events phishing email and drive-by-download. Our current experiments do not address the role of complex events, so we will not discuss them further.

The final inference step is **Assessment**: MIFD assigns a degree of belief to each of the event hypotheses. MIFD does this by performing Bayesian updating on the belief network constructed by the clustering process. Scyllarus and MIFD are intended as long-running online services, not batch processes. Scyllarus has run for months at a time. One major re-

¹Where possible, we try to introduce sensors that can specifically detect the benign events. To return to our earlier example, we might introduce a “print server sensor” to help disambiguate detections. We do not address this issue in our experiments.

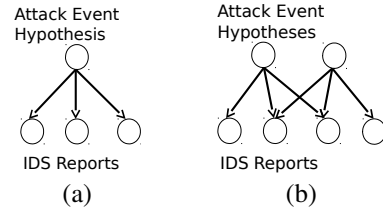


Figure 2: Bayes networks generated by MIFD.

quirement for such long-term operation is that the inference steps must be performed cyclically. MIFD repeatedly reads reports and clusters them, periodically interrupting that cycle to perform assessment.

MIFD follows Pearl’s exhortation (1988) that the core of probability theory is the patterns of inference it enables (explaining away, combining forward causal inference and evidential reasoning, etc.), rather than precise numerical calculations. He recommends these patterns be used even when precise numbers are not available, and that is what MIFD does. As outlined above, MIFD treats IDS fusion as an abductive problem, formalized using Bayes nets. But precise values for the parameters of these Bayes nets — probabilities of attacks, false positives, false negatives, etc. — are *never* available to us: populations of attacks and attackers change, different networks have little in common with each other, distributions are non-stationary, etc. Moreover these difficulties are not solvable via machine learning: Sommer and Paxson (2010) discuss the challenges IDS pose to ML.

Goldszmidt and Pearl’s System $Z+$ (1996), shares the basic structure of normal probability theory, but abstracts the actual probabilities. Events have a natural number rank κ corresponding to their degree of surprise (so a rank of one is more surprising than zero). The semantics of this scheme comes from a set of probability distributions in which the probabilities are polynomials over some infinitesimal ϵ . In this scheme, the κ corresponds to the exponent of the polynomial’s leading term. As we mentioned earlier, System $Z+$ is similar to the “big- O ” method used in analysis of algorithms. This semantics gives System $Z+$ a qualitative flavor by providing a “ladder” of events of qualitatively different orders of likelihood. Note that System $Z+$ is very different from Wellman’s qualitative probability networks (Wellman 1990), whose abstraction is based on direction of change, rather than order of magnitude relations “at rest,” as it were. Wellman’s approach is more relevant when choosing interventions (as in his work on medical diagnosis and treatment), and less for fusion.

Under System $Z+$ we may apply the normal operation of probability theory — conditionalization, Bayes’ law, etc. — but the arithmetic operations we use for them change. Rather than multiplying probabilities, we add degrees of surprise; rather than adding probabilities, we take the minimum surprise. Goldszmidt and Pearl (1996) provide the following substitutions between the quantitative probability function P and its qualitative counterpart κ for formulas ϕ and ψ ,

and primitive events, e .

$P(\psi) = \sum_{e \in \psi} P(e)$	$\kappa(\psi) = \min_{e \in \psi} \kappa(\phi)$
$P(\psi) + P(\neg\psi) = 1$	$\kappa(\psi) = 0 \vee \kappa(\neg\psi) = 0$
$P(\psi \phi) = P(\psi \wedge \phi)/P(\phi)$	$\kappa(\psi \phi) = \kappa(\psi \wedge \phi) - \kappa(\phi)$

In a Bayes network, with conditionally independent ω and ϕ we further have $\kappa(\omega \wedge \phi) = \kappa(\omega) + \kappa(\phi)$.

There are a number of efficient algorithms for finding the posterior distributions of Bayesian networks, conditional on observations of some of the random variables. These algorithms may readily be adapted to provide posterior κ rankings instead of probabilities. MIFD translates Bayes networks into an ATMS (Forbus and deKleer 1993); see (Charniak and Goldman 1988; Poole 1993; Provan 1989) and [citation withheld] for this encoding. This implementation is not especially efficient, but we have found it is I/O which dominates runtime, and we apply several optimizations to handle common special cases. To date (somewhat to our disappointment!) optimizing the general System $Z+$ inference method has not been key to system performance.

MIFD’s assessment ranks hypotheses as either *likely*, *plausible* or *unlikely*. A hypothesis h , conditioned on evidence ω , is *likely* if $\kappa(h|\omega) < \kappa(\neg h|\omega)$, *plausible* if $\kappa(h|\omega) = \kappa(\neg h|\omega)$ and *unlikely* otherwise.

The MIFD system requires comparatively few κ parameters given the above design. For sensors, we need $\kappa(\text{false-positive})$, and for events (attack and benign) we need $\kappa(\text{event})$. In practice, we assign global defaults for these, based on how generally accurate the input IDSes are: for example, we set these κ values so that it takes $n = 2$ or 3 sensors for us to judge an event as *likely*, with the following consistency constraints:

$$\begin{aligned} \kappa(\text{false-positive}), \kappa(\text{benign}) &< \kappa(\text{attack}) \\ &< n \cdot \kappa(\text{false-positive}) \end{aligned} \quad (1)$$

That is, a single sensor false positive is less surprising than an attack event; a benign event is also less surprising than an attack; and an attack is less surprising than false positives from n of the sensors detecting that event. With a few exceptions, this paper assumes $n = 3$ sensors, so we have $0 < \kappa(\text{benign}) < \kappa(\text{attack}) < 3\kappa(\text{false-positive})$. In actual deployments, we then nudge the false positive rankings up in the rare cases where we have a particularly good sensor, or down if we have a particularly bad sensor. We are less likely to adjust attack event probabilities, but there are exceptions such as scanning (very common), and some complex events.

Experimental Design

Our experiments probe limits of accuracy of System $Z+$. Our first two experiments consider what happens as our qualitative abstraction is a more or less accurate reflection of reality. It is relatively accurate when the probabilities of events it treats as qualitatively distinct are far apart, and less and less accurate as they come closer and closer together. Our third experiment considers how MIFD degrades as its input sensors become less and less precise. Finally, we consider how well MIFD addresses base rate problems – to what extent, and under what circumstances fusing redundant sensors qualitatively will increase accuracy.

In all of these experiments, we define sensor arrangements, and probabilities for attacks, confounding benign events, and sensor false positives and false negatives. We sample from them to find ground truth and sensor report sets, run MIFD on the sensor reports, and evaluate its accuracy. These experiments are necessarily artificial, since they assume omniscience about parameters which are even in principle inaccessible.

For each experiment we create a *configuration*, parameters specifying how a set of events and sensors (the *setting*) is generated. A setting comprises a set of possible attacks, a set of sensors, and a set of sensing relations between attacks and sensors. In each *run* of a setting we sample from the attack and benign events. In each experiment we generated 1,000 settings with 100 runs per setting. After sampling events, we sample from the sensors according to their false positive and false negative probabilities.

Having generated the events and sensor reports for a run, we then use the reports as input to MIFD, and assess the resulting Bayes networks. We extract the set of attack event hypotheses that have been labeled as *likely* by MIFD. We evaluate MIFD’s performance in terms of *precision* and *recall*, comparing the set of events that MIFD considers likely with ground truth generated for the run. Recall is the percentage of actual attack events which are labeled as likely. Precision is the percentage of attack events labeled as likely which actually occurred. Because of the high rate of reports and the low rate of events, precision and recall must be in the high nineties, or the sensing system will not be usable. We report the effect of our experiments on precision and recall in our results section below.

Configurations. In each of our experiments, we start with a configuration that reflects the realities of IDS fusion, and whose parameter values are broadly consistent with the simplifying assumptions of System $Z+$. Configuration parameters may be either constants or parameterized random variables. We then vary some configuration parameter in ways which degrade MIFD’s performance. These experiments show us MIFD’s performance under best-case conditions, and how that performance degrades.

- *num-events* — The total number of events (both benign and attacks) to be detected by each sensor, a measure of the sensor’s specificity. Initially we use a random variable which returns 1 80% of the time, and 2 otherwise.
- *sensor-to-attack-ratio* — The number of sensors which should detect each attack event, initially 3.
- *sensor-overlap* — The number of types of attack to be detected by each sensor, specified as a probability checked when deciding whether to associate an additional attack type with a sensor, initially 0.2.
- $\kappa(\text{false-positive})$ — The qualitative probability of false-positives for the sensors. We choose these to be 1 in configurations where it takes 3 reports to rank a hypothesis likely, or 2 where it takes only 2.
- $\kappa(\text{attack})$ and $\kappa(\text{benign})$ — The qualitative probabilities of attack and benign events, initially 2 and 1 respectively.
- *sensor-to-benign-ratio* — The number of sensors which should detect each benign event, initially 3.
- *kappa-translations* — Translations (into a constant or into

a random variable) of qualitative probability values to real values in $[0, 1]$. Initially we translate to constants, $\kappa(0) = 0.5$, $\kappa(1) = 0.01$, $\kappa(2) = 0.001$.

- **num-attacks** — The number of attacks to take place, initially 4.

Settings. When creating a setting, we first consult the *num-attacks* parameter to determine the number of attack prototypes. Each prototype is associated with an occurrence probability given by (or sampled for each prototype from) $\kappa(\text{attack})$. The number of sensors is not fixed, but instead depends on several configuration parameters. For each attack, we keep track of the number of sensors which we must assign to detect it; this value is initially set from *sensor-to-attack-ratio*. We create new sensors as long as any attack requires assignment to an additional sensor. A new sensor is initially assigned a total number of events which it will detect from *num-events*, and some set of attack events. Each sensor is assigned at least one attack, plus additional attacks up to its total number of events as consecutive samples of a Bernoulli random variable with success probability *sensor-overlap* return 1. After all sensor prototypes are created and assigned attacks, we assign benign events to the sensors to reach their total number of events. We create enough benign event prototypes to satisfy both *sensor-to-benign-ratio* and each sensor’s total event count, and distribute them randomly among the sensors according to their total event counts.

The experiments.

Varying probabilities. Our first experiment examines how MIFD’s performance degrades as we progressively violate the assumption that different levels of the stratified likelihood ranking qualitatively differ. Specifically we consider a sequence of settings in which we assign probabilities to each κ , always with $P(\kappa = 0) = 0.5$ but varying the values for 1 and 2. We test with *sensor-to-attack-ratios* of 2 and 3 to see how much corroboration is necessary. In the case of two sensors/attack we took $\kappa(\text{false-positive})=2$, $\kappa(\text{benign})=2$, $\kappa(\text{attack})=3$ instead of the defaults above, in order to satisfy Equation 1. We bring the probabilities corresponding to the κ s closer and closer to see how MIFD’s performance degrades.

Non point-value distributions. The above experiment still represents a considerable abstraction: in general the set of events to which we assign the same qualitative likelihood will not all have the same probability. Our second experiment examines how variation in the probabilities of the events at the same qualitative likelihood affects the accuracy of qualitative fusion. To do so, we define a second order probability distribution for each κ ranking using a *beta distribution*, the Bayesian prior for Bernoulli distributions. The β has two parameters, α and β . The sum $\alpha + \beta$ corresponds to increasing the “virtual sample size,” and controls the variance. To assess the sensitivity of MIFD to variance in the probabilities, we fix the mean values of the distributions for $\kappa(0)$, $\kappa(1)$ and $\kappa(2)$ at 0.5, 0.01 and 0.001, and increase the variance by decreasing the number of virtual samples.

Sensor imprecision. We next explore how the performance of our techniques degrade as the sensors encounter an increasing overlap of possible events in their “field of

$P(\kappa = 1)$	$P(\kappa = 2)$	$s2a = 3$		$s2a = 2$	
		P	R	P	R
0.005	0.0005	99.97	96.73	99.28	98.32
0.01	0.001	99.93	96.39	98.81	98.04
0.05	0.005	96.36	92.94	94.37	97.33
0.1	0.01	86.71	89.47	89.92	96.23
0.25	0.125	84.16	56.93	74.53	70.57
0.33333	0.22222	75.87	39.52	66.06	53.78
0.375	0.28125	71.56	31.28	61.84	45.34
0.46975	0.43945	58.57	15.54	53.56	27.76

Figure 3: Precision (P) and recall (R) percentages from the **varying probabilities** experiment. The second result set varies the initial sensor-to-attack ratio (*s2a*) from 3 to 2.

view.” For this experiment we fix the *sensor-overlap* to be 0.9, and vary *num-events*. This value for *sensor-overlap* is significantly higher than in earlier runs, but we do not expect performance in Setting 1 of this experiment to differ greatly from Setting 1 of the earlier experiments because of the low initial values of *num-events*: 80% of the time in this first run there will be only a single event associated with a sensor, in which cases the *sensor-overlap* is not consulted at all. With the higher values from *num-events* in subsequent settings, more attacks will be associated with each sensor, increasing the opportunity for MIFD to misdiagnose the true cause of sensor reports.

Base rate. Our fourth experiment examines how our performance degrades as the rate of true attacks goes down. Such base rate problems, where even a high accuracy sensor can perform unacceptably for very unlikely events, plague intrusion detection (Axelsson 1999). We examine these issues by reducing the probability of generating attack events (but not the probability of generating benign events).

Results

Varying probabilities. Figure 3 shows MIFD’s performance as the stratified likelihood rankings become less distinct. We see that MIFD’s performance degrades gracefully as the probabilities corresponding to $\kappa(0)$, $\kappa(1)$ and $\kappa(2)$ approach each other. Although the theoretical basis of the qualitative probability levels is infinitesimal, MIFD performs reasonably for $P(\kappa = 1)$ and $P(\kappa = 2)$ taken as high as 0.05 and 0.005 respectively. MIFD also degrades gracefully as its input sensors become less discriminating, simulated by raising the sensor-to-attack ratio from 2 to 3. We had expected MIFD to be more brittle under sensor-to-attack ratio 2, but this was only partly true; precision declines more quickly with ratio 2 than with ratio 3. Recall declines less quickly, because the corroboration of several sensors is more relevant to suppressing spurious additional event diagnoses.

Non point-value distributions. Figure 4 (i) shows MIFD’s performance when we interpret the qualitative likelihood levels as various beta distributions. The results confirm those of the previous experiment, and show that MIFD is

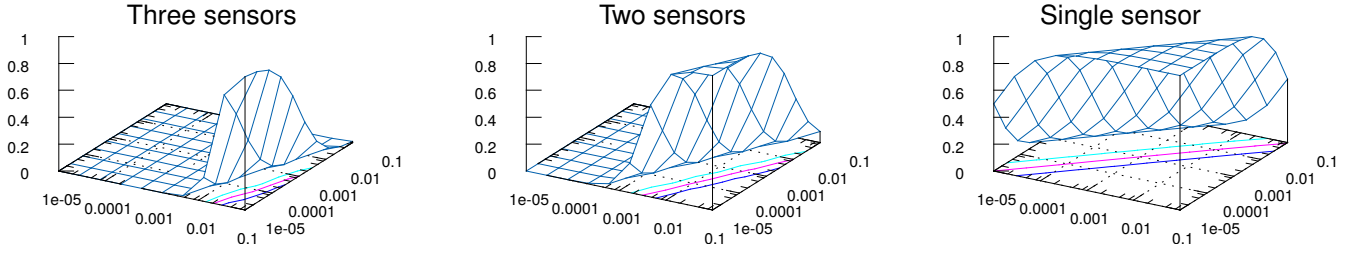
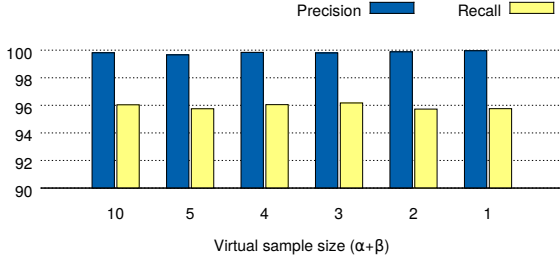
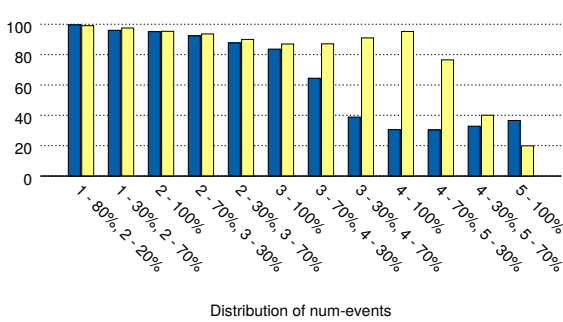


Figure 5: Base rate problems and sensor fusion. The graphs show the false positive probabilities (on the z-axis) as a function of the false alarm probability (on the x-axis, running from the origin to the lower right on each graph) and attack probability (on the y-axis). The diagonal contour lines project onto the xy-plane where the surface intersects $z=0.25$, $z=0.5$ and $z=0.75$.

(i) Beta distributions



(ii) Sensor imprecision



(iii) Base rate

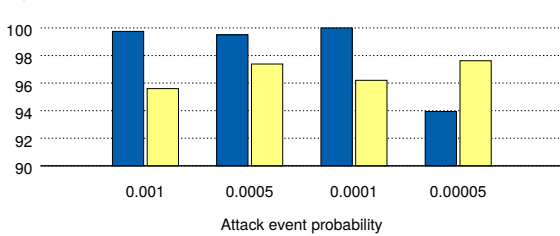


Figure 4: Precision and recall percentages for the **beta distribution**, **sensor imprecision** and **base rate** experiments.

not sensitive to variance in the probabilities of events at the same qualitative surprise level.

Sensor imprecision. Figure 4 (ii) shows how MIFD's accuracy declines when sensors correspond to more than one event. Recall and precision stay well above 90% for sensors responding to two or fewer distinct events. Unsurprisingly, precision declines sharply as sensors respond to three or more events. This decline is not particularly serious in practice: actual IDS sensors rarely detect large numbers of events: if they do we introduce new abstractions.

Base rate. Figure 5 shows the challenge of base rates, and how MIFD addresses it through sensor fusion. In the rightmost plot, we see that the false positive probability is extremely high with only one sensor, even a very accurate one. However, requiring corroboration from two or three sensors, as MIFD does, reduces the false alarm rate substantially. Note that this requires the sensors to fail independently. If sensors' false positives are correlated, then corroboration can fail to lower the false positive rate. It was to handle such correlated failures that we introduced benign events into Scyllarus and MIFD.

Figure 4 (iii) shows how MIFD's false positive rate rises with event rarity. The high recall shows that MIFD does find the attacks (the bounce up at the right is because of the extremely low sample size), but declining precision is from mistakenly hypothesizing additional attacks. Precision declined unacceptably below 90% for smaller attack probabilities than shown on this chart.

Conclusions

The results presented in this paper clarify why Scyllarus and MIFD have been successful in practical deployments. They show that the qualitative scheme they use is not sensitive to the actual probabilities of events, and that its performance degrades gracefully. The results also justify the claim that by combining the output of multiple sensors, even very noisy sensors, IDS fusion can tame the high false positive rates that plague the field of intrusion detection. Since our results predict the circumstances under which IDS fusion will work and will fail, they can also be used to inform the design and deployment of IDSes for effective incorporation in a fusion system. Finally, our results should encourage prospective users of qualitative schemes based on probabilistic reasoning, and promote deeper examination of such systems.

References

- ArcSight. 2008. Arcsight enterprise security manager. http://www.arcsight.com/product_info_esm.htm.
- ARIS. 2003. Attack registry & intelligence service. <http://aris.securityfocus.com/AboutAris.asp>. ARIS analyzer Data Sheet.
- Axelsson, S. 1999. The base-rate fallacy and its implications for the difficulty of intrusion detection. In *Proceedings of the 6th ACM Conference on Computer and Communications Security*, CCS '99, 1–7. New York, NY, USA: ACM.
- Charniak, E., and Goldman, R. P. 1988. A logic for semantic interpretation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 87–94.
- Debar, H.; Curry, D.; and Feinstein, B. 2007. The intrusion detection message exchange format (IDMEF). RFC 4765, IETF NWG.
- Finkle, J., and Heavey, S. 2014. Target says it declined to act on early alert of cyber breach. Reuters web site. <http://www.reuters.com/article/2014/03/13/us-target-breach-idUSBREA2C14F20140313>.
- Forbus, K. D., and deKleer, J. 1993. *Building Problem Solvers*. Cambridge, Massachusetts: MIT Press.
- Goldszmidt, M., and Pearl, J. 1996. Qualitative probabilities for default reasoning, belief revision and causal modeling. *Artificial Intelligence* 84(1–2):57–112.
- Google Scholar. 2014. Citations of (Goldszmidt and Pearl 1996). Query: http://scholar.google.com/scholar?cites=11462195722076826989&as_sdt=5,24&sciodt=0,24&hl=en.
- Lee, W.; Mè, L.; and Wespi, A., eds. 2001. *Recent Advances in Intrusion Detection (RAID 2001)*, number 2212 in LNCS. Springer-Verlag.
- Leydon, J. 2001. IDS users swamped with false alerts. *The Register*. http://www.theregister.co.uk/2001/12/15/ids_users_swamped_with_false/.
- Minock, M., and Kraus, H. 2002. Z-log: Applying system-z. In Flesca, S.; Greco, S.; Ianni, G.; and Leone, N., eds., *Logics in Artificial Intelligence*, volume 2424 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg. 545–548.
- Pearl, J. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Los Altos, CA: Morgan Kaufmann Publishers, Inc.
- Poole, D. 1993. Probabilistic horn abduction and Bayesian networks. *Artificial Intelligence* 64:81–129.
- Provan, G. 1989. An Analysis of ATMS-based Techniques for Computing Dempster-Shafer Belief Functions. In *Proceedings of the 11th International Joint Conference on Artificial Intelligence*, 1115–1120. Morgan Kaufmann Publishers, Inc.
- Schwartz, M. 2014. Target ignored data breach alarms. InformationWeek web site, <https://www.informationweek.com/security/attacks-and-breaches/target-ignored-data-breach-alarms/d-d-id/1127712>.
- Sommer, R., and Paxson, V. 2010. Outside the closed world: On using machine learning for network intrusion detection. In *Proceedings of the IEEE Symposium on Security and Privacy*.
- Valdes, A., and Skinner, K. 2001. Probabilistic alert correlation. In Lee et al. (2001).
- Vandoorselaere, Y. 2008. Prelude correlator. <https://trac.prelude-ids.org/wiki/PreludeCorrelator>. Prelude Correlator online documentation.
- Vigna, G.; Kemmerer, R. A.; and Blix, P. 2001. Designing a Web of Highly-Configurable Intrusion Detection Sensors. In Lee et al. (2001), 69–84.
- Wellman, M. P. 1990. Graphical inference in qualitative probabilistic networks. *Networks* 20:687–701.
- Wikipedia. 2014. Intrusion detection systems. http://en.wikipedia.org/wiki/Intrusion_detection_system.