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

Robert P. Goldman and John Maraist

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 (Goldman and Harp 2009) and its successor MIFD (Burstein et al. 2012). 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 dramatically reduce false alarm rates, the bane of intrusion detection systems, reducing the flood of incoming reports by multiple orders of magnitude. Unfortunately, we cannot 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 (as opposed to theoretical investigations of the logic) 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.

Goldman and Harp discuss the first of these two questions (2009). In this paper we will focus on the second question, assuming a solution to the first.

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 to the user. Ideally, network administrators would have a suite of different IDSes active, since different IDS approaches bring different strengths and weaknesses, and have different "fields of view." For example, local privilege escalation attacks may be invisible to a NIDS; HIDS systems are more accurate for many purposes, but impose higher administrative burdens because they must be scattered around the defended network, rather that centrally deployed like NIDS. Signature-based IDSes provide fewer false positives, but are more likely to be blind to zero-day attacks than anomaly-based systems.

We will not discuss the varieties of IDSes further here – there is a voluminous literature on them. The *precise* nature of IDS weaknesses and strengths is not critical to our investigation here, but the fact that there *are* such strengths and weaknesses is what motivates our information fusion, and difficulties in modeling the events and sensors is what motivates our use of qualitative methods.

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. Although the sensor is accurate, the probability of having the disease given a positive test result is only 1%.

Some of our reviewers expressed the concern that we would have difficulty *modeling* the rare events in IDS fusion. There are three reasons when this problem does not arise. One is that these attack events are not rare in the world: they are rare with respect to our sampling rate. As we point out above, if we are sampling individual packets, the probability of any attack is vanishingly small, but that doesn't mean that attacks are rare at the "macroscopic" level. Second, our causal models are limited to models of sensing and fusion, part/whole, and class/subclass relations. These models are very simply generated by rules. Finally, the IDS developers provide us initial ontologies of events implicitly with their rules. Our knowledge engineering aims at smoothing inconsistencies between these implicit ontologies, and extending causal models by identifying cases where IDS developers have provided incorrect intuitions about the events their systems detect. At any rate, this concern is not relevant to the evaluation of the uncertainty calculus, which is our focus in this paper.

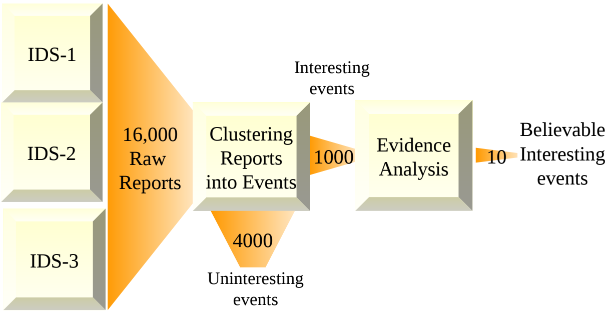

Figure 1: Scyllarus workload reduction.

Users regularly ignore or disable their IDSes, 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 nonexistent 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 card records (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. It shows how our earlier system was able to winnow the flow of reports in a small corporate network over a day of operation (Goldman and Harp 2009).

Related work. STAT and MetaSTAT (Vigna, Kemmerer, and Blix 2001) use finite-state models to detect and fuse events, but do 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 is closest to our clustering subsystem, but its knowledge resides in stateful rules instead of an ontology of attacks (Vandoorselaere 2008).

The Scyllarus and MIFD systems

Our original system Scyllarus performed IDS fusion on a diverse set of third party IDSes (Goldman et al. 2001). Its successor MIFD (Model-based Intrusion Fusion and Detection) enhances Scyllarus to integrate IDS fusion into a comprehensive suite for cyber defense (Thayer et al. 2013). 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 to address the lack of standardization in data formats and event taxonomies.

IDS fusion proper begins with **Clustering.** From the sensor reports, and based on an ontology detailing how (and with what probability) sensor reports may correspond to events, 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. Finally, Figure 2(c) shows a more complex example where an event has two component sub-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. Worse, some 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.

The second inference step is **Assessment,** in which 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.

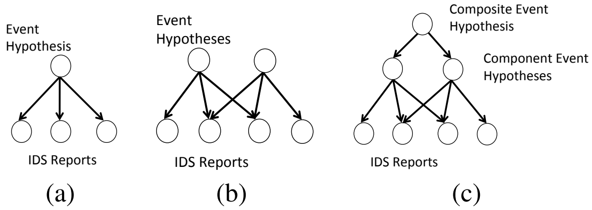

Figure 2: Bayes networks generated by MIFD.

Note that Scyllarus and MIFD are intended as long-running online services, not batch processes. Scyllarus has run for months at a time. One major requirement 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. Scyllarus also periodically flushes older events and reports out of working memory (they persist in a database).

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. MIFD treats IDS fusion as an abductive problem, formalized using Bayes nets. But precise values for the parameters of these Bayes nets 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 IDSes 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. System $Z+$ is similar to the "big-O" method used in analysis of algorithms. With this semantics System $Z+$ provides a "ladder" of events of qualitatively different orders of likelihood.

Under System $Z+$ we may apply the normal operation of probability theory but the arithmetic operations we use for them change. Rather than multiplying probabilities, we add degrees of 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.

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

In a Bayes network, with conditionally independent ω and ϕ

[1] Where possible, we try to introduce sensors that can specifically detect the benign events.

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; Goldman and Harp 2009) 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.

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. That is, a hypothesis is likely (unlikely) if some maximally likely scenario labels it as true (false), and no maximally likely scenario labels it as false (true). A hypothesis is plausible in any other situation. Note that this restriction to three classes applies to MIFD's output, not its inputs — the κ values of MIFD's underlying model range over arbitrary natural numbers as needed to distinguish qualitatively distinct likelihoods. It would be possible to alter the assessment process to provide more classes of likelihood, using the difference in kappa ranks between the maximally likely scenario in which a hypothesis is labeled as true (false) and the one in which it is labeled as false (true). We have not seen the need to do so, and indeed worry that additional precision would come at the cost of validity.

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:

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

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.

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 as well as the probabilities for attacks, confounding benign events, 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. Note that while the models are generated artificially, they have the same structure as models used in real deployments of the Scyllarus system, although we have limited ourselves to more simple cases.

For each experiment we create a *configuration*, a set of parameters specifying how a set of events and sensors, a *setting*, is generated by sampling random variables and comparing to these parameters. 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, for each configuration, we generated 1,000 settings and conducted 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 event types (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.

- κ(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.
- κ(attack) and κ(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 types. Each type is associated with an occurrence probability dictated by the κ(attack) parameter. 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 based on a Bernoulli random variable with parameter *sensor-overlap*. After all sensor types are created and assigned attacks, we assign benign events to the sensors to reach their total number of events. We create enough benign event types 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-ratio*s of 2 and 3 to see how much corroboration is necessary. In the case of two sensors/attack we took κ(false-positive)=2, κ(benign)=2, κ(attack)=3 instead of the defaults above, in order to satisfy Equation 1. We bring the probabilities corresponding to the κs 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.

		$s2a = 3$		$s2a = 2$	
$P(\kappa = 1)$	$P(\kappa = 2)$	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.

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 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

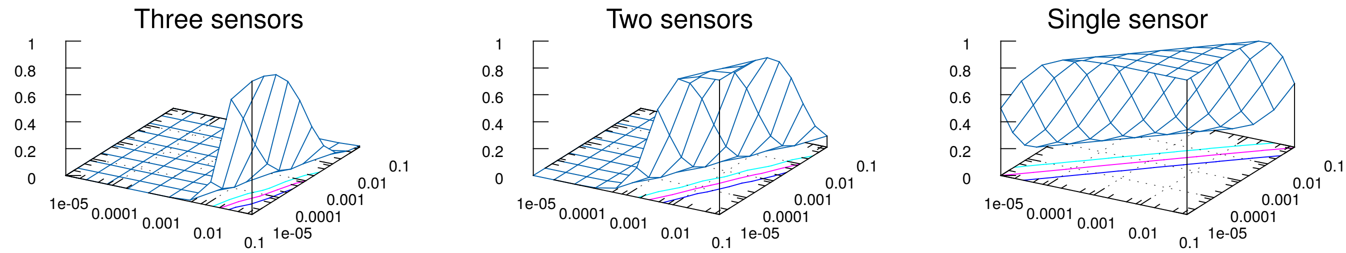

Figure 5: Graphs of quantitative false positive probabilities (on the z-axis) as function of the probability of an individual sensor giving a false positive reading (on the x-axis, running from the origin to the lower right on each graph) and attack probability (on the y-axis). Specifically, each graph plots $z = P(\text{no attack}|\text{alarm raised}) = \frac{x^s(1-y)}{y+x^s(1-y)}$ for the given number of sensors s.

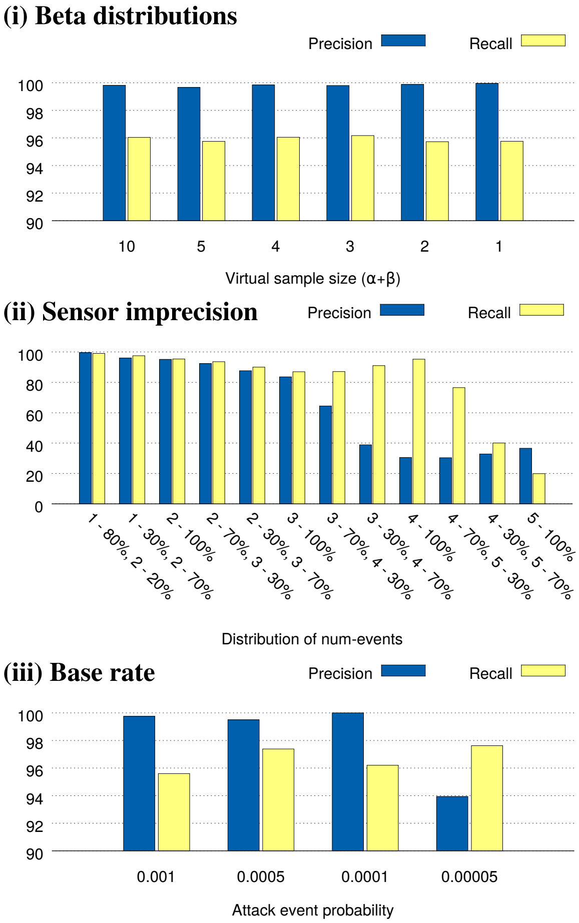

Figure 4: Precision and recall percentages for the **beta distribution**, **sensor imprecision** and **base rate** experiments.

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 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.

We would like to remind the reader that imprecise probabilities are forced upon us by the nature of the cyber intrusion detection domain: it's not just a matter of not having the right machine learning technique. The true probabilities are non-stationary, they involve an adversarial process, vary from location to location, and are difficult if not impossible to learn because of the absence of valid labeled training data. The use of a small number of qualitatively distinct likelihood levels also aids us in the knowledge engineering process.

In practice, we also find that the results of System $Z+$ calculations are easy to understand. We have found this through experience explaining the output of our systems to users, and through experience debugging. When we draw out the Bayes networks, we can think of using assumptions to assemble the "cheapest" explanation for a set of observations. This can be simpler to understand than the exact computations in a Bayes net.

Our experimental 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. These results are somewhat independent from the question of the adequacy of the qualitative calculus. In a system that performed diagnosis/fusion using conventional probability theory, it can be seen that multiple sensors that fail independently will tend to perform well: it's easier to drive the probability of error down by multiplying failure probabilities (p^n) than to try to drive down a single sensor's p of failure. However, our results show that substituting System $Z+$ for conventional probability theory preserves this desirable characteristic of probabilistic reasoning. Our results also show resistance to a moderate degree of correlation in sensor failures when using System $Z+$.

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. In future work, we would like to examine more complex inference patterns that arise from "knotty" Bayes nets, and the accuracy of the assessment process in models that take into account how attacks spread through the topology of the underlying computer network.

Acknowledgments. This paper was supported by DARPA and the U.S. Air Force under contract number FA8650-11-C-7191. Approved for Public Release, Distribution Unlimited. The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

References

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.

Burstein, M.; Goldman, R. P.; Robertson, P.; Laddaga, R.; Balzer, R.; Goldman, N.; Geib, C.; Kuter, U.; McDonald, D.; Maraist, J.; Keller, P.; and Wile, D. 2012. STRATUS: Strategic and tactical resiliency against threats to ubiquitous systems. In *Proceedings of the Adaptive Host and Network Security Workshop*.

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.

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.

Goldman, R. P., and Harp, S. A. 2009. Model-based intrusion assessment in Common Lisp. In *Proc. Int'l Lisp Conference*.

Goldman, R. P.; Heimerdinger, W.; Harp, S. A.; Geib, C. W.; Thomas, V.; and Carter, R. L. 2001. Information modeling for intrusion report aggregation. In *DARPA Information Survivability Conference and Exposition(DISCEX-2001)*, 329–342. DARPA and the IEEE Computer Society.

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). 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*. Waltham, MA: Morgan Kaufmann.

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.

Schwartz, M. 2014. Target ignored data breach alarms. InformationWeek Dark Reading web site. http://www.darkreading.com/attacks-breaches/target-ignored-data-breach-alarms/d/d-id/1141468.

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*.

Thayer, J.; Burstein, M.; Goldman, R. P.; Kuter, U.; Robertson, P.; and Laddaga, R. 2013. Comparing strategic and tactical responses to cyber threats. In *SASO Workshop on Adaptive Host and Network Security AHANS*.

Valdes, A., and Skinner, K. 2001. Probabilistic alert correlation. In Lee et al. (2001).

Vandoorselaere, Y. 2008. Prelude correlator. `https://www.prelude-siem.org/projects/prelude/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.

Wikipedia. 2014. Intrusion detection systems. `http://en.wikipedia.org/wiki/Intrusion_detection_system`.