

Are Deeper Levels of Risk Analysis a Requirement for Enabling Optimal Tactical Responses in INFOSEC Alert Correlation Systems?

Stephen W. Neville

ECE Dept., University of Victoria

Stephen.Neville@ieee.org

Abstract

As network speeds and complexities increase, the development of automated systems that enact optimal tactical responses will be required. INFOSEC (information security) alert correlation systems provide a natural home for such capabilities. It can be asked whether the current generation of these systems has the technical capabilities required to enact optimal tactical responses. Specifically, is there a requirement to incorporate deeper levels of risk analysis within correlation systems? Currently, correlation systems only model attack risk via the generic attack severity metrics. Hence, these systems implicitly assume that: (a) all attacks are uniquely identifiable, or (b) the risk associated with the attacks is uniformly distributed across the set of plausible attacks. This work provides formal support for the intuitive supposition that such assumptions may not be supportable in the real-world and, hence, that integrated risk modeling is likely a necessity if optimal tactical attack response sub-systems are to be added.

1. Introduction

As the role of IT networks as primary mediums for commercial and governmental transactions has increased, the critical nature of cyber-security has been brought to the fore. Currently, cyber-security is primarily a human intensive effort with its core analytical and response tasks resting in the hands of highly trained security analysts. As has been experienced in other domains, such as network management, an industrial dependence on a small highly-skilled and specialized work force is not a scalable solution. Hence, scarcity of skilled analysts is likely to become a limiting factor in achieving strong cyber-security within corporate and governmental networks.

Training programs to address this issue have been initiated at the national level within the U.S. and in other countries. But, these efforts will not fully address the problem. As network speeds increase, the time frame within which tactical attack responses must be enacted

commensurately decreases, assuming total attack byte counts remain relatively constant. In addition, as networks and IT systems become more complex, more information must be analyzed within these decreasing time windows in order to enact correct tactical responses. Given that network size, complexity, and speeds are all increasing, the time is quickly approaching where direct reliance on human enacted tactical attack responses is untenable. This inevitability is underlined by the increasing capabilities and availability of automated attack tools, as discussed in [24] in terms of known malicious uses of Nessus [18].

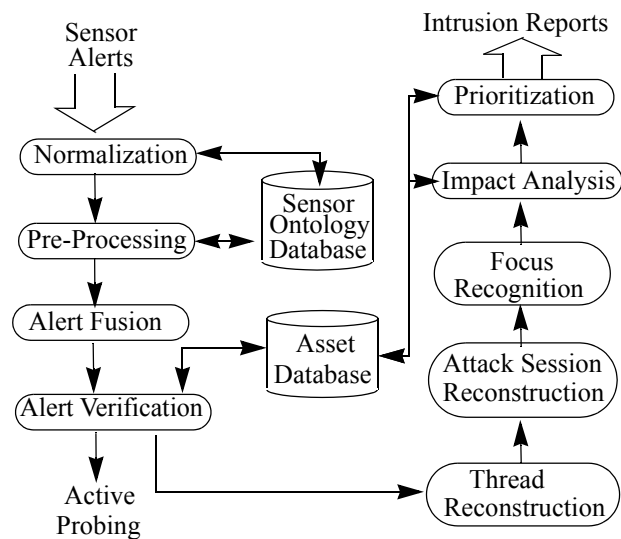


Figure 1: Block diagram of the alert correlation process, taken directly from Vigna's work of [1].

From a corporate perspective, cyber-security acts similar to insurance. The overall goal being to manage the risk of corporate losses due to malicious cyber-events against the costs of the cyber-security systems (*i.e.*, to maximize the return-on-investment (ROI) for cyber-security expenditures). For corporations these losses, or more precisely the expectation of loss, can usually be modeled in dollar terms; whereas, for governments such a singular metric would be incomplete. The overall increase

in cyber-security awareness has triggered a significant increase in the R&D of cyber-security sensors and systems, generally referred to under the umbrella term INFOSEC (information security) systems. The current generation of INFOSEC sensors are fundamentally low-level devices that when deployed in large-scale heavily used networks produce voluminous streams of low-level alerts that overwhelm the human security personnel tasked with their analyses.

Alert correlation systems have been specifically developed to address this problem. These systems are designed to roll-up (*i.e.*, combine) the generated low-level alert streams into higher level knowledge detailing the specific nature of the attacks and attack strategies being perpetrated against a given protected network. This higher level knowledge is passed on to human security analysts in the form of prioritized intrusion reports, and from these reports the human analysts then determine what tactical responses to enact. Figure 1, taken directly from [1], details the block diagram of a typical alert correlation system, which is a hierarchical multi-stage process beginning with the normalizing of the low-level alerts streams into a common format, and ending with the generation of a prioritized list of intrusion reports.

Since correlation systems exist at the systems level, they are an ideal home for automated tactical response systems. In fact, the current generation of correlation systems do enact limited responses, typically restricted to relatively simple facilities such as TCP/IP connection severing. In practice, though, these response mechanisms are normally disabled to mitigate self-inflicted DOS and attack amplification events. As the repertoire of available tactical responses increases, and under the suppositions outlined above, an obvious desire will be to build automated tactical response systems that perform optimally (*i.e.*, that guarantee the corporation achieves the optimal level of security available from its deployed defenses).

The fundamental questions asked by this work are: (a) Is the current generation of alert correlations technically capable of serving as the home for such response sub-systems? and (b) if not, what changes need to be made? The answer to the first question fundamentally rests on whether the output of the correlation systems is provably correct (*i.e.*, Do the generated intrusion reports provide a true representation of the attacker(s) activities up to the current point in time?). Optimal responses are trivial to achieve. Correctness holds in alert correlation systems if all the INFOSEC sensors generate unique alerts (a term formally defined in Section 6). But, as will be shown, proving uniqueness, even in under idealized cases, requires proving that there is a zero probability that any

other event (or events), known or unknown, could have generated the observed alerts.

The fundamental question is whether there is a one-to-one mapping between the generated alerts and the plausible attack set (*i.e.*, one and only one set of attacks could have given rise to the observed alerts). The current generation of alert correlation systems enforces an assumed one-to-one mapping between observed alerts and reported attacks. Hence, it is reasonable to ask if optimal tactical responses can be generated under this assumption. It should be noted that designing a correlation system which reports the most likely attacks given the observed evidence is not sufficient to allow for optimal tactical responses. In particular, as will be discussed in Section 4, the attacker can then gain an advantage by enacting unlikely attacks.

To select “optimal” responses requires a suitable optimality metric. To the author’s knowledge no such metric has been proposed in the literature. This work develops such a metric based on a game theoretic analysis of the attacker and defender behaviors, with the basis of the metric being the basic tenant that any tactical response should be effective, should not amplify the attack’s effects, and should be the “easiest” to enact of those that elicits the desired outcome.

Via this optimality metric, the question is formally asked whether risk modeling is required to enact optimal tactical responses. In particular, an idealized defender advantageous model is constructed and the conditions under which risk modeling is not required are determined through analyzing the attacker and defender behaviors in terms of an extensive form game. As these conditions are derived in an idealized environment, they represent necessary conditions within the real-world. It is then shown that these conditions cannot be assumed to hold in the real-world; hence, the general conclusion must be that risk modeling is likely a real-world requirement. The obvious question is then the degree to which this requirement is a necessity (*i.e.*, do, or will, real attackers exploit the current lack of risk modeling). This question has been left as an area of future work, as the main focus of this work is to develop the required theoretical underpinnings.

Inherently, it is difficult to assess the real-world impact of the postulated issues in that the nature of the problem is such that evidence is not currently collected that would indicate whether the actual attacks are other than those reported by operational correlation systems. Alerts occur; they are clustered; the analyst enacts responses. It is likely that only long after this cycle has completed would evidence come to light that attacks other than those reported have indeed occurred and, most likely, such attacks would be attributed to missed alerts.

A simulation based approach would not help solve this problem since, as has been discussed by Vigna [1], the current simulation-based testing approaches for correlation systems do not provide complete test coverage. Hence, a mis-correlation event may be missed. Even if such events exist with low probabilities, game theory argues that they would be used once they become known to the attacker. Therefore, statistical testing methodologies are insufficient. Additionally, novel attacks exist and security is about future events. One can always argue that today's simulated results may not apply tomorrow. This is not to say that simulation and real-world studies are not important, and, in fact, this is a main focus of the future research deriving from this work. But, in reaching for the goal of optimal security, it is also important to have an analytical framework from which to base the security.

Section 2 begins the work by over viewing past works within the alert correlation domain, inclusive of the available analytical models and the reported testing approaches. Section 3 then provides a practical illustration of the issues of interest in this work. Sections 4 and 5 then define the attacker and defender models under consideration, with the formal optimality metric given as part of Section 5. Section 6 determines the conditions required for the unique identification of attack events under an idealized environment and shows that they cannot be assumed to hold in the real-world. Section 7 outlines the relationship between unique attack detection and the requirement for risk modeling. Section 8 presents the areas of future work flowing from this analysis. Section 9 then concludes the work. This work focuses solely on the issues involved in determining optimal tactical responses. Strategic responses are outside of the scope of this work.

2. Related Work

This work is most closely related to the work available on alert correlation systems and their evaluation. The principal work on formal modeling is that of Debar's M2D2 correlation data model [7]. This model denotes the information sources available to the correlation process and their interrelations. It does not model the details of how information is mapped between enacted attack space and the generated sensor alert space. This level of detail is required to determine the degree to which risk modeling is needed to enact optimal tactical responses.

Within the literature, a number of operational correlation systems have been described. The principal systems are those of Cuppens [4], Vigna [1], Julisch [2], Ning [5][11], and Valdes and Skinner [9]. These works

focus primarily on the correlation process itself. Hence, they do not generally address the problem of how to construct optimal attack responses. The response issue is addressed in part in Cuppens' work of [25] through two basic response mechanisms, those of *kill-login* and *close-connection*. These response mechanisms are initiated through *anti-correlation* analysis which employs a similar mechanism to the requires/provides model of computer attacks proposed by Levitt [10]. In this latter approach the observed attack sequences are analyzed to determine the attacker's likely next steps. Responses are then enacted that will result in the removal of the pre-conditions required for the attacker to proceed in their attack. In the work of Dain and Cunningham [26], the basic approach to requires/provides (or pre-condition/consequent) modeling is extended to a probabilistic domain. The work of Ning [5][11] extends the notion of alert correlation to the domain where it is assumed that there is missing and incomplete alert information. Ning's approach employs a pre-condition/consequent analysis approach to determine the nature of the missing information.

None of these works address the issue of the effects that the expectation of loss has on the correlation process. The focus is primarily from the perspective of determining the attacker's intent. As will be shown in Section 5 there is a distinction from this approach and a corporate objective of obtaining maximal security at minimal cost (i.e. achieving the best possible cyber-security return on investment). It should be noted that Julisch's work on alert correlation [2] is distinct from these other mentioned works in that it explicitly addresses only the issue of correlating non-malicious alerts. This distinction is important in that it facilitates a number of his work's proofs. Since non-malicious alerts do not require tactical responses, Julisch's work is not directly related to this work.

Three general approaches have been proposed within the literature to test correlation systems: (a) the re-playing of intrusion data sets, such as the MIT DARPA data sets [20] or the DEFCon data sets [21], into small-scale research networks [1][5][22], (b) the enacting of real attacks within small-scale networks [1][8], and (c) the placement of the clustering (or correlation) systems within operational network segments and evaluating its performance against observed attacks [2][9][23]. Typically, evaluation results are reported in terms of the data reduction achieved and the accuracy of the clustering performance against ground-truthed attacks. Obviously, full-scale testing of correlation systems within large-scale

unrestricted environments is impractical since the data volume issues would prohibit ground-truthing. These experimental tests have provided strong verification of the ability of alert clustering to address the data overload issue. But, they fall far short of providing formal proofs of correctness. In particular, as discussed by Vigna in [1], each of these testing approaches has significant weaknesses.

The work of Porras and Valdes [27] addresses the correlation issue from the basis of assessing mission impacts. But, supporting mission continuity is not equivalent to minimizing the expectation of loss. In particular, [27] has closer similarities to the survivability literature. Such approaches are only tangentially related to this work as they focus on maintaining network services and tasks in the presence of attacks and not on tactical responses. The difference is more clearly seen in the following scenario. An attacker enacts an attack that enables them to obtain a copy of a proprietary corporate report. The attack is such that it does not disturb any of the network's operations. Hence, survivability is maintained, but losses are incurred, the amount being dependent on the report's contents. A timely tactical response may prevent the loss but will not change the network survivability. Hence, the two issues are only tagentially related.

3. A Practical Illustration

To illustrate the issues of interest in this work, assume that an attacker possesses two attacks against a corporate database. Attack A allows them to perform unauthorized reading of database entries but does not allow for modifications. Attack B is similar to attack A and it allows for deletions. Attack A has two stages which produce alerts a_1 , and a_2 , respectively. Attack B also has 2 stages, with stage 1 producing the same alert a_2 as stage 2 of A, while stage 2 produces alert a_3 . Assume that a_3 has a high false alarm rate such that only when it is seen in conjunction with a_2 is there enough grounds to assume attack B has occurred. a_1 also has a high false alarm rate. When the attacker performs attack A the company incurs an expected loss of \$1, while for B the expected loss is \$1M. The cost of responding optimally to A is \$1, while to respond to B costs \$50k, since a complete scan of the database must then be undertaken. Responding to A when the real attack is B reduces B's expected loss to \$100k. An intelligent attacker begins by enacting stage 1 of A

followed by stage 1 of B. The correlation system will initiate A's cluster once it sees a_1 . B's a_2 will then be added to this cluster once it arrives and the system will report attack A. The attacker can then perform stage 2 of B which will be reported as a false alarm. Risk modelling is required under this hypothetical situation if optimal defense is to be achieved as the correlation system possesses no information with which it can distinguish this event from a situation where only attack A occurs.

4. Attacker Model

The notation and assumptions regarding the attackers are as follows. A set of *atomic attacks*, denoted $\alpha = \{\alpha_1, \dots, \alpha_j, \dots, \alpha_{N_\alpha}\}$, exists and represents the complete space of attacks available to the attackers. Denote an individual attacker by A_k . Denote by $\alpha_k \subset \alpha$ A_k 's knowledge of α (*i.e.*, the atomic attacks available to A_k). It is assumed from the tactical perspective that A_k does not learn new attacks during the course of enacting attacks (*i.e.*, α_k is fixed for the duration of any given attack). For simplicity, each A_k is assumed to be attempting to reach a single selected goal G_k by enacting a step-by-step sequence of atomic attacks chosen from α_k . The resultant attack sequence, termed a *composite attack* and is denoted by $\alpha_J = \{\alpha_j\}_{j \in J}$, where J is an index set on α_k and is not restricted to possessing unique entries. If an attacker is simultaneously seeking multiple goals then this is modelled as multiple collaborative attackers through introducing additional A_k 's, one for each of the desired goals. Hence, A_k is more precisely viewed as identifying the attacker associated with a given α_j and not as representing the activities of that attacker. Collaborative attacks can be modeled by allowing the resulting shared atomic attacks to be members of all composite attack sequences that exploit the collaboration. The boundary of the composite attacks is therefore defined in terms of the attack's overall goal, with the atomic attacks being steps towards that goal. This approach is consistent with available works in attack sequence analysis and attack graph generation, for example [10]-[13], and it allows each composite attack to be dealt with individually.

Attacks are assumed observable to the defender only through the alerts generated by the deployed security sensors and collected at the correlation system. The arrival

times of these alerts are asynchronous events as viewed by the correlation system. Each atomic attack is assumed to be initiated at some time t_j and have a finite time period $T(\alpha_j)$ over which it is enacted. From the defender's perspective, $T(\alpha_j)$ appears as a random variable drawn from an unknown distribution. No assumption is made as to whether the attackers are internal or external to the defended network(s). This basic model is consistent with the standard attacker models and assumptions.

Additionally, it is assumed that the attackers are intelligent and rational as per game theory's definitions [14]. Specifically, rationality refers to the attackers enacting the attacks that maximize, at each step in the attack process, their perceived utility gain in reaching the goal. For each atomic attack $\alpha_j \in \alpha_k$ there is assumed to be a time dependent mapping $u_k(t): \alpha \rightarrow \mathfrak{R}$ conditioned on the attack's goal G_k , such that $u_k(\alpha_j, t|G_k)$ defines the *utility gain* achieved by the attacker if α_j is enacted at time t . For reasons outlined below the attacker cannot exactly know $u_k(\alpha_j, t|G_k)$; hence, the attacker is assumed to operate from an estimate $\hat{u}_k(\alpha_j, t|G_k)$, defined as the *perceived utility gain*. At each decision point, the attacker assesses the perceived utility gains and chooses the atomic attack that is estimated to bring them closer to their goal. Rationality therefore defines the non-random selection of atomic attacks at each step in the process of building a composite attack sequence designed to achieve a desired goal.

Intelligence within a strict game theory definition refers to both the attackers and defenders possessing the same understanding of the network and its defenses. Obviously, in the cyber-security domain such a level of knowledge is not generally achievable by an attacker unless they are also a defender. For the purposes of this work, a weaker definition will be employed, with intelligence denoting partial knowledge of the defense by the attackers. Denote the complete domain of all cyber-security related information about the target network as N , where this is assumed inclusive of all systems on the network, their locations, all executing software within the network, all defensive measures, etc. At time t , A_k possesses an estimate of N , denoted as $\hat{N}_k(t)$. Under the weaker intelligence definition $\hat{N}_k(t) \subset N$.

The attackers' and defender's actions are both modeled as enacting changes to N ; hence, both have as their ultimate goal the adaptation of N such that complete control over the opponent(s) is achieved. The attacker's

perceived utility gain is more precisely denoted as $\hat{u}_k(\alpha_j, t|G_k, \hat{N}_k(t))$. Each attack is assumed to return some information about N . This information gain, obtained by enacting α_j , is denoted as $I_k(\alpha_j, t+\Delta)$, where for simplicity $I_k(\alpha_j, t+\Delta) \geq 0$ (*i.e.*, active mis-information is not assumed to be part of the defender's repertoire) and $\Delta > 0$ represent the delay in obtaining this information gain. If it is assumed that the defender responds to the attacks such that $\hat{N}_k(t)$ remains valid, then

$$\hat{N}_k(t+\Delta) = \hat{N}_k(t) \cup I_k(\alpha_j, t+\Delta), \quad (1)$$

and $\hat{N}_k(t)$ will be a monotonically non-decreasing function of t . Obviously, this monotonic gain only holds when the defender does not enact tactical responses. Note that if $\hat{N}_k(t) = N$ then

$$\hat{u}_k(\alpha_j, t|G_k, \hat{N}_k(t)) = u_k(\alpha_j, t|G_k, N), \quad (2)$$

and the attacker can perfectly choose their next atomic attacks. Obviously, enacting each atomic attack comes at a cost to A_k . Denote this cost as $c_k(\alpha_j, t)$, where $c_k: \alpha \rightarrow \mathfrak{R}^+$. This cost is time dependent since both the attacker's and defender's actions influence N . Ostensibly, there exist times when certain attacks are easier (or harder) to accomplish. For example, if the network is heavily loaded then DOS attacks are easily enacted since the target systems are closer to their overload conditions.

If $c_k(\alpha_j, t)$ is viewed to be inclusive of both the direct and indirect costs associated with α_j then the attacker cannot be assumed to know $c_k(\alpha_j, t)$ perfectly. Within this context, indirect costs are defined as those costs associated with the effects of the defender's responses, up to and including the potential for attack attribution and any subsequent legal consequences. Hence, the attacker must proceed based on the estimate $\hat{c}_k(\alpha_j, t)$ and not on $c_k(\alpha_j, t)$ itself. The following scenario highlights this issue. A_k enacts attack α_j . This attack is detected by the defender and the nature of α_j allows for the trace back and identification of the attacker. This then leads to legal actions being commenced against A_k . In this case, A_k 's perceived cost, $\hat{c}_k(\alpha_j, t)$, may be sufficiently low that they choose to enact the attack, whereas had $c_k(\alpha_j, t)$ been available, the attack would not have been enacted. Obviously, $c_k(\alpha_j, t)$ is available only in hindsight.

The above model allows the attack process to be described as an iterative process of selection and enacting atomic attacks over a sequence of decision points $\{t_1, \dots, t_m, \dots, t_n\}$ such that at each step t_m an $\alpha_j \in \alpha_k$ is selected which satisfies

$$\arg \max_{\alpha_j} (\hat{u}_k(\alpha_j, t_m|G_k, \hat{N}_k(t_m))). \quad (3)$$

If the sequence of atomic attacks enacted by A_k up to time t is denoted as $\alpha_j(t)$, where each $\alpha_j \in \alpha_j(t)$ is such that it satisfied Eq. 3 at its decision point t_j (*i.e.*, when it was selected), then A_k 's estimated total cost of enacting $\alpha_j(t)$ is

$$\hat{c}_k(t|\alpha_j) = \sum_{\alpha_j \in \alpha_j(t)} \hat{c}_k(\alpha_j, t_j). \quad (4)$$

Obviously, there exists a maximum cost for the composite attack that is palatable to the attacker (*i.e.*, the attacker is resource limited in some sense). Denote this limit on the total cost as C_k^* . The attacker's iterative search process, under the above game theoretic model, ends when: (a) the goal G_k is reached, or (b) C_k^* is exceeded. Note that since the index set J is assumed inclusive of repeated entries, theoretically the attacker can construct infinitely many composite attacks to try. The situations where the defenses are immune to the attacks known to the attacker or where the attacker merely gives up trying are modeled as case (b), as it is assumed that the attacker can modify C_k^* during the search. Hence, the attacker's goal at each decision step is more precisely modeled as choosing the α_j which satisfies Eq. 3 while simultaneously minimizing Eq. 4. Balancing these concerns is up to the attacker. Obviously, the above cost framework assumes the attackers intend to play the game to win, as opposed to moving on to softer targets should penetrating their chosen targets' defenses exceed their attack expertise.

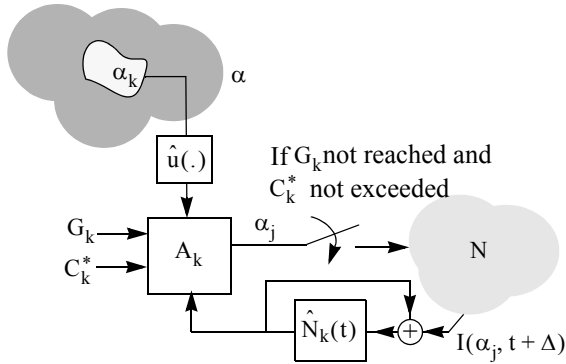


Figure 2: The attacker's information sources.

The attacker's behavior can therefore be modeled as an extensive form game with partial knowledge, where at each decision node the attacker enacts the atomic attack that satisfies Eq. 3 and minimizes Eq. 4. The game continues until one of the two terminal conditions is reached. Unlike standard extensive form games, the utility function is assumed to dynamically changes in response to

changes in N , where such changes occur due to the actions of all of the attackers and defenders.

Figure 2 shows the block diagram of the information sources assumed available to the attacker and how they change in time. Figure 3 shows how this block diagram maps into an extensive form game with a set of decision points where the attacker chooses which atomic attack to enact next. It is important to note that the standard mapping of the game as one of alternating turns is not appropriate in that both attacker and defender actions are asynchronous events with respect to each other's viewpoint. Hence, an attacker may proceed through several decision nodes before the defender enacts a response, or the defender may enact several responses before the attacker transitions to the next decision node. This can be modeled as viewing the game, from each attacker's perspective, as the playing of a single player game in which the environment randomly perturbs prior to each decision node.

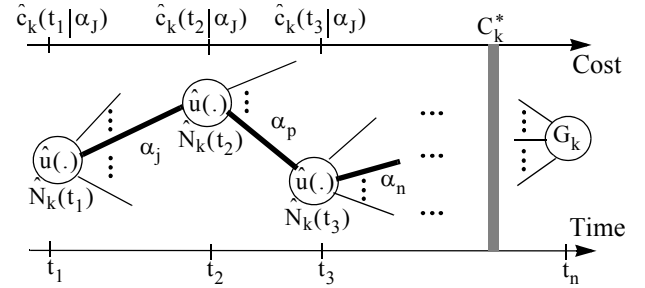


Figure 3: Attacker behavior as an extensive form game under a randomly perturbing environment.

5. Defender Model

The defender model is similar to the attacker model in that the defender is playing the opposition side of the game described above. The defender's task is to detect and then respond optimally to each enacted atomic attack. A single defender is assumed to be playing simultaneously against an unknown number of attackers. The question is whether the defender requires risk analysis to play optimally. Of key interest are the processes by which the defender detects attacks and both their internal and composite timing characteristics, as this gives rise to the situations previously illustrated in Section 3.

It is assumed that the defender has distributed a heterogeneous set of cyber-security sensors $\mathbf{s} = \{s_1, \dots, s_n\}$ within the network enclave under their protection. These sensors are assumed to provide the sole information about the attackers' actions. Each sensor is

assumed to contain a set of triggers $\{x_k\}$. For simplicity, these triggers are assumed to be binary decision functions that map the information space they observe, denoted as E_{s_j} , into the domain $\{0, 1\}$ (*i.e.*, $x_k: E_{s_j} \rightarrow \{0, 1\}$). It is assumed that $\bigcup_{\forall j} E_{s_j} = E$ where E is the space of all events occurring within the network, $e_m \in E$ being one such event and by definition $E \subset N$.

More precisely, each trigger x_k defines both a mapping g_k of the event space E into its own detection (or feature) space X_k (*i.e.*, $g_k: E \rightarrow X_k$) and a sub-space within this detection space, $x_k \subseteq X_k$ that the event e_m must cover for the trigger to fire. Define an indicator function $1_k(e_m)$ such that

$$1_k(e_m) = \begin{cases} 1 & \text{if } g_k(e_m) \supseteq x_k \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

When $1_k(e_m) = 1$ the associated trigger generates the alert a_k . All a_k 's are assumed to be complete in the sense that they report all of the information available from their respective x_k sub-spaces inclusive of the triggering event's source and destination information and the sensor's id. Denote the complete set of alerts generated by α_j as $\text{alerts}(\alpha_j) = \bigcup_{s_j} 1_{x_k}(\alpha_j) \cdot a_k$ where $1 \cdot a_k = a_k$ and $0 \cdot a_k = \emptyset$. The event space E is assumed to be composed of both normal (or authorized) events and attack events. The effects of false negatives are modelled as causing changes in N for which the defender cannot enact tactical responses since the attacks went unobserved.

The goals of the alert correlation system is to group all the a_k 's generated by each α_j up to time t into a distinct cluster $\theta_j(t)$. The defender then analyses the set of all such clusters (or cluster information space), denoted as $\Theta(t) = \{\theta_j(t)\}_{\forall j}$, in order to estimate the atomic attacks that have been enacted. Denoting this set of estimated atomic attacks as $\{\hat{\alpha}_j\}$, the defender's goal is to have $\hat{\alpha}_j = \alpha_j \quad \forall j$ (*i.e.*, perfect identification of all attacks). Section 6 discusses the conditions under which this objective can be met within an idealize environment.

Once this set $\{\hat{\alpha}_j\}$ is arrived at, the current generation of alert correlation systems passes it up to human security analysts, packaged as prioritized intrusion reports, for the enactment of the appropriate tactical and strategic responses. From a tactical perspective this may include re-tuning of

firewall rules and IDS rule sets, the dropping of the suspect TCP/IP connection, etc. Whereas, from a strategic perspective this may include, but is not limited to, the deployment of new system patches, the re-structuring of the network topology, the deployment of new security sensors, etc.

In order to achieve optimal tactical responses a formal definition optimality metric is required. To the author's knowledge, no previous formalization of this problem have been presented in the alert correlation literature. Section 2. To derive such a metric first, observe that, as in the attacker case, the defender does not know N precisely and, hence, must work from a time dependent estimate, denoted as $\hat{N}_d(t)$, which, as in the attacker case, improves via defender information gain $I_d(t + \Delta) \geq 0$. In general, one expects $\hat{N}_d(t)$ is a better estimate than $\hat{N}_k(t)$, especially if the attacker is external to the network. But this may not be the case if the attacker is an insider, and particularly if they are one of the employed security analysts. Additionally, the defender does not know the true utility gain that α_j provides A_k and it must be estimated as $\hat{u}_d(\hat{\alpha}_j, t | \hat{G}_k, \hat{N}_d(t))$, \hat{G}_k being the defender's estimate of A_k 's goal obtained through analysis of $\Theta(t)$.

The defender can base their tactical response on $\hat{u}_d(\hat{\alpha}_j, t | \hat{G}_k, \hat{N}_d(t))$. But, from a corporate perspective, it can be argued that this is not a desirable approach since this does not take into account any measure of the expectation of loss. For corporations such losses are bottom line issues whereas, strictly speaking, minimizing the attack utilities may not be. Specifically, some attacks may have a high utility to the attacker but only inflict minimal losses, such as a defaced web site. The defender is better off in this case focusing their resources on prioritizing attack responses to minimize the expectant losses. The overall corporate goal is assumed to be one of maximizing cyber-security at minimal cost (*i.e.*, obtaining the best possible ROI from cyber-security investments). Obviously, this objective stems from the assumption that the defender is resource limited in some fashion. Extending the analysis past the corporate context would require non-monetary evaluations of loss to be incorporated, which is outside of this work's scope.

Denote the financial loss of attack α_j as $\lambda(\alpha_j)$, where $\lambda: \alpha \rightarrow \mathfrak{R}^+$. In practice, this loss can be defined, for example, as the total cost of recovery from α_j . Obviously, the defender can only know $\lambda(\alpha_j)$ perfectly in hindsight as the true extent of the losses or damage caused by an attack may not be apparent for some time. For example, an attack may result in the loss of intellectual property and this may not

be immediately discovered. Hence, at the time when the defensive response is enacted the defender must operate based on a time dependent estimate of $\lambda(\alpha_j)$. Denote this estimated loss by $\hat{\lambda}(\hat{\alpha}_j, t)$, where obviously this loss is also a function of the attack estimate. If it is assumed for simplicity that attack losses accumulate linearly (*i.e.*, $\lambda(\alpha_j) = \sum_{\forall \alpha_j \in \alpha_j} \lambda(\alpha_j)$) then the total estimated expected losses at time t are $\hat{L}(t|\Theta(t)) = \sum_{\forall \hat{\alpha}_j \in \Theta(t)} \hat{\lambda}(\hat{\alpha}_j, t)$. This is compared to the true loss at t which is $L(t) = \sum_{\forall \alpha_j \text{ enacted}} \sum_{\forall \alpha_j \in \alpha_j} \lambda(\alpha_j)$.

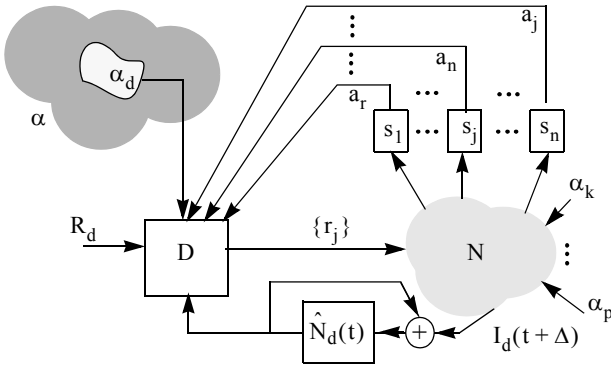


Figure 4: The defender's information sources.

With regards to the tactical responses, assume that the defender possesses a set of known responses that can be enacted and that this response set is static for the duration of each composite attack. Denote this response set by R_d , and an element of this set by $r_j \in R_d$. Each response is assumed to effect the attack by enacting a change in N , for example through updating a firewall or IDS rule set, serving a TCP/IP connection, changing routing tables, etc. Assume for each $\alpha_j \in \alpha_d$ there exists a $r_j \in R_d$ such that this r_j is the best known response for α_j , α_d representing the defender's knowledge of α . Denote such best responses by r_j^* .

Obviously, each response entails some cost to the defender. In particular, the response may effect the normal (or authorized) activities on the network. Denote this defensive cost as $c_d(r_j, t|N)$. Obviously, this cost must be conditioned on N as N includes information regarding the normal activities. The true cost of enacting a set of responses $\{r_j\}$ at time t is therefore

$$C(\{r_j\}, t|N) = \sum_{\forall r_j \in \{r_j\}} c_d(r_j, t|N), \quad (6)$$

assuming linearity applies. The defender though can only possess an estimate of this true cost, denoted as

$$\hat{C}(\{r_j\}, t|N_d(t), \{\hat{\alpha}_j\}) = \sum_{\forall r_j \in \{r_j\}} \hat{c}_d(r_j, t|N_d(t), \{\hat{\alpha}_j\}), \quad (7)$$

where $\{\hat{\alpha}_j\}$ is the set of estimated atomic attacks obtained from $\Theta(t)$, and \hat{c}_d is the estimated cost associated with each response.

The above model allows the defender's behavior to be described as an iterative process of selection and enacting tactical responses over a sequence of decision points $\{t_1, \dots, t_m, \dots\}$ such that at each step t_m a set $\{r_j\} \subset R_d$ is selected which satisfies

$$\operatorname{argmin}_{\{r_j\}} (g[\hat{L}(t_m|\Theta(t)), \hat{C}(\{r_j\}, t_m|\hat{N}, \{\hat{\alpha}_j\})]) \quad (8)$$

where $g(\cdot)$ is the defender's chosen functional weighting between the expected losses and the response costs. When $r_j = r_j^* \forall r_j$ then, by definition, Eq. 8 is optimized with respect to the information that known to the defender at the time when the response is enacted. In an idealized situation better responses may be possible but such responses would require the defender to know unobtainable information. The above model represents the best the defender can do based on the information they have. From the above analysis it can be seen that, as in the attackers' case, the defender's behavior can be modelled as playing an extensive form game where the environment randomly perturbs prior to each decision node.

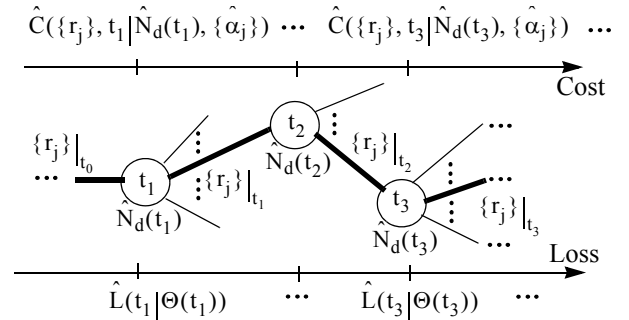


Figure 5: Defender behavior as an extensive form game under the randomly perturbing environment N .

In this case, though, the defender must play *ad infinitum* as there exists no "winning" node from the defender's perspective. "Winning" would require total dominance of all attackers to be achievable. This is not feasible under the rational and intelligent assumption unless the defender can guarantee no attacks exist which exceed a palatable loss threshold. Such a guarantee must

be inclusive of novel attacks. Figures 4 shows the information sources available to the defender and how they update. Figure 5 shows how this block diagram maps to the extensive form game in which the environment randomly perturbs.

6. Conditions the for Unique Attack Identification

Obviously, if all attacks are uniquely identifiable then determining optimal tactical responses is trivial under the above model, since it would always be the case that $\hat{\alpha}_j = \alpha_j$. Hence, an important question is: “What are the conditions required for unique attack identification to occur within real-world large-scale networks?” To derive these conditions the characteristics of the alert generation and atomic attack estimation processes must be explored in more detail. Obviously, the existence of unique alerts directly implies unique attack identification, where the formal definition of uniqueness is,

Definition: An alert alert_k is *unique* iff for some $\alpha_j \in \alpha$ $\text{Prob}(\text{alert}_k \in \text{alerts}(\alpha_j)) = 1$ and $\forall \alpha_m \in \alpha$ $m \neq j$ $\text{Prob}(\text{alert}_k \in \text{alerts}(\alpha_m)) = 0$. All alerts that are not unique are defined as *non-unique*.

Under this definition, and under the assumptions that: (a) all alerts contain complete information, (b) all attacks are detected via the deployed sensors, and (c) no triggers are redundant or superfluous, the causes of non-unique alerts can be identified.

Theorem 1: Under binary triggers all alerts are unique iff $\forall x_k \in X$ $\neg \exists x_j$ $k \neq j$ such that $x_k \subset x_j$ when x_k and x_j are co-located.

Proof by contradiction. Assume that all alerts are unique and there exists two co-located triggers such that $x_k \subset x_j$ for $k \neq j$. There must exist an $\alpha_m \in \alpha$ such that $I_j(\alpha_m) = 1$ otherwise x_j would be superfluous. Since $x_k \subset x_j$ then it is also the case that $I_k(\alpha_m) = 1$, hence $\text{alert}_k \in \text{alerts}(\alpha_m)$. Given that x_k is neither redundant nor superfluous, there must also exist an α_n with $n \neq m$ such that $I_k(\alpha_n) = 1$ and $I_j(\alpha_n) = 0$. Hence, $\text{alert}_k \in \text{alerts}(\alpha_n)$. Therefore, alert_k is non-unique contradicting the original assumption. If there does not exist a co-located x_k such that $x_k \subset x_j$ then, since all alerts are defined to contain complete information, all alerts must be unique. Hence, alerts are unique iff there does not exist a co-located trigger x_k such that $x_k \subset x_j$ $\forall x_j \in X$. ■

When only unique alerts exist, the optimization of Eq. 8 is trivially achieved. Hence, the more interesting case is when it is assumed that there exist at least some non-unique alerts. In this case, without a loss of generality, the focus can be placed solely on the subset of attacks that generate

non-unique alerts. Obviously, the potential for non-unique alerts increases if the generated alerts do not report complete information.

For this subset of attacks, the defender must either (a) await the first unique alert to arrive before enacting the response, or (b) achieve unique identification by analysing the composite set of arriving non-unique alerts. In either case, the response must be timely, where from a tactical perspective “timely” implies at a point before the composite attack completes (*i.e.*, before $T(\alpha_j)$). Obviously, if the last alert generated by the composite attack is required to uniquely identify the final atomic attack then the defender cannot provide a timely defence. Additionally, it can be assumed that expected losses increase as atomic attacks complete. Hence, it is assumed that the defender enacts responses based on the analysis of subsets of the atomic attack’s generated alerts and does not wait for the complete set of alerts to arrive before acting.

Obviously, the case (a) above reduces to the case of all alerts being unique if the defender chooses only to focus their attention on the set of unique alerts. Hence, security can also be trivially achieved in this case, assuming that the defender has identified *a priori* the best responses for each unique alert. Case (b) is therefore more relevant to real-world operational security in that it is not reducible to the situation where all attacks can be directly and uniquely identified. The question then is: “If non-unique alerts are assumed, then what conditions are required, under the idealized model, to guarantee unique atomic attack identification can be performed solely through the analysis of subsets of the generated alerts?”

Theorem 2: Unique atomic attack identification is guaranteeable in this case, under the game theoretic assumptions, iff it can be proven that the observed subset of alerts has a zero probability of having been generated by any other atomic attack or any possible combination of atomic attacks.

Proof by contradiction. Assume the above condition does not hold and there exists an atomic attack that can still be uniquely identified solely by a subset of the non-unique alerts it generates. Then by definition there is a non-zero probability that these alerts were in fact generated by either a different atomic attack or a set of attacks. Since it is assumed that all of the information available to the defender about attacks comes solely by way of the generated alerts, the defender has no means of distinguishing these two cases. Hence, the defender cannot uniquely identify the attack. If the above condition holds then by definition the atomic attack is uniquely identifiable. Hence, unique identification by the analysis of non-unique alerts requires that the probability of the observed alerts being generated in any other way is provably zero. ■

Of course, in general, the defender has no means constructing the required proofs since this would require complete knowledge of the attack space α . Additionally, the order by which alerts arrive at the correlation system is non-deterministic. The alert arrival timing is characterized by inherent delays from the attack propagating through the network, from the processing time for the security sensors to generate the alerts, and from the propagation time required for the alerts to transit the network from the generating sensor to the correlation system. Hence, the defender cannot use assumptions about alert orderings in their analyses. A knowledgeable attacker can also enact coordinated attacks to further influence the ordering of the attack generation process. It is possible to mitigate the attacker's direct effects on the sensors and to utilize a secure path to transmit the alerts, but by definition the attacker has the access required to enact such coordinated attacks. Outside of malicious influence, delays in large-scale networks have been shown to be non-trivial to accurately model [15][16], particularly since what is required from the tactical defence viewpoint is knowledge of the instantaneous delays and not just the delay distributions. Hence, the defender cannot hinge optimal tactical defences on statistical models of the alert orderings.

The above analysis does not address the question of how common such non-unique alerts may be in practice, or how exploitable this approach may be in the real-world. Determining the non-uniqueness of alerts requires knowledge of the sensor triggers; such information is not released for COTs products for obvious reasons. Obtaining it would require reverse engineering of all of the COTs systems deployed in the defence. In open source products this information is obtainable and, for example, it is well known that the SNORT IDS [17] contains a number of internal non-unique alerts which give rise to the need to know the employed rule orderings. Given the known high false alarm rates of operationally deployed INFOSEC sensors, it is unlikely that non-unique alerts are rare. Whether they can be exploited with reasonable computational costs has been left as an area of future work. What the above analysis does show is that, even in an idealized environment, correctness of the correlation system cannot be proven.

7. Implications for Risk Modeling

The above analysis therefore implies that the defender must deal with the situation where the set of observed alerts maps to multiple possible sets of estimated atomic attacks. Hence, the formal analysis does not support the defender assuming that the mapping between observed alerts and estimated attacks is one-to-one. No assumption was made in Section 5 that the responses r_j mapped one-

to-one to atomic attacks. Hence, it may be possible that for all of the plausible sets of atomic attack estimates $\{\hat{\alpha}_j\}$ derivable from the observed alerts there exists a singular set of best response set $\{r_j\}$ (*i.e.*, regardless of which set of attack estimates is deemed "true", the same set of responses provides the best tactical response). If this is the case then the defender can minimize Eq. 8 without needing to model the expected losses. For the case of non-unique alerts, this is the only time when these losses can be ignored and optimal tactical responses can still be enacted. If a singular attack set cannot cover all of the plausible attacks, then Eq. 8 cannot be optimized without requiring that the estimated losses be considered. The fundamental issue is that the defender has no information from which to distinguish which of the plausible attacks is actually occurring at the point in time when the tactical response must be enacted. Hence, minimizing Eq. 8 requires risk modeling.

An argument can be made that it is unlikely for the attackers to be able to exploit the defender's inability to uniquely identify attacks, or that such a deficiency would not be known to the attackers. If it is assumed that what is under discussion are COTs based defenses then these arguments are not supportable under the assumptions of Section 4. Specifically, a motivated attacker can purchase and experiment with the same defensive COTs available to the defender. Rationality was defined as each attacker's step by step decision process being one of selecting the atomic attack that satisfies Eq. 3 based on their level of knowledge about the attack target and about attacks. Obviously, if the defender can be placed in a situation where multiple response choices exist then this gives the attacker an advantage. Rationality drives the attacker to generate such a situation if possible. Vigna has shown the practical feasibility of reverse engineering the "secret" signature of COTs IDS [19].

If the defender chooses not to perform risk analysis during the selection of tactical responses, they are inherently assuming (a) that all attacks are uniquely identifiable, or (b) that the risk is uniformly distributed across the set of potentially detected attacks. Either of which would need to be formally proved, within the context of real-world large-scale networks, if optimal tactical responses were sought and the need for risk modeling was discounted.

8. Future Work

Several areas of future work are apparent from the above analysis. First, the game theoretic based attacker and defender models are not complete in the sense that no methodologies are presented for arriving at the various

estimates that are required. Nor are the forms of such estimates given. Second, the above analysis implies a necessity to have correlation systems support the generation of the complete set of plausible attack estimates based on the observed evidence. The computational cost of this implication is not addressed. Third, it was assumed that the losses and costs can be linearly summed but this is conjecture. Fourth, the attacker model assumes that the attacker is concerned only with reaching a singular goal, therefore, it cannot model opportunistic attacks. In some cases, such attacks could be the majority of what is faced. Finally, possibilities obviously exist to enact optimal responses via the analysis of the composite attacks as they develop, as per the work in attack sequence and attack graph analyses, provided that the resultant sequences of atomic attacks are unique (*i.e.*, that the attacker's objective can be uniquely identified). The accounting for this possibility has not been included. Such possibilities fundamentally rest on the need for the defender to possess complete knowledge of what attacks the attacker can enact. Hence, its lack of inclusion does not limit the relevance of this work.

Most importantly, the issue of the verification of the above theoretical analysis needs to be addressed through simulation and real-world studies. As was stated in the introduction, the purpose of this work was to lay the theoretical groundwork for those studies. A facility capable of performing real-time simulations and the physical reproduction of the observed network traffic behaviors of corporate-scale (3000+ host) networks is under construction to enable these studies. This future work will include addressing the issue as to the degree to which the postulated issues are in fact exploitable by attacks.

The issues brought forward by this work are also a small part of the overall problem of constructing a formal methodology to assess alert correlation system correctness. It is hoped that further work in combining game theory and correlation analysis may lead to advances in this area. A significant impediment, though, is the lack of a generally agreed upon formal model for alert correlation systems which is inclusive of all of the pertinent issues.

9. Conclusions

This work has answered the general question as to whether desiring optimal tactical responses inherently presupposes the need to employ deeper levels of risk analysis in alert correlation systems. The arguments supporting this supposition were derived by first developing game theoretic frameworks to formally

describe both the attackers' and defender's motivations and behaviors. These models lead to the description of each player's behavior in terms of single player extensive form games where, prior to each decision node, the game environment randomly perturbs in accordance to the other player's actions. At each decision node, the attacker selects the atomic attack that maximizes their perceived utility gain based on the information they know. Similarly, the defender selects the "optimal" responses based on their estimates of the enacted attacks, as determined by the alerts they observe, such that their responses minimize the total expected losses and the response costs. The attacker was assumed to only be concerned with enacting singular composite attack sequences. The defender was modeled as defending against an unknown number of attackers. Hence, although the attacker plays a finite game which ends when the goal is reached or the attack's costs exceed what is palatable, the defender plays on *ad infinitum*.

The details of the alert generation process and its timing characteristics were then analyzed in order to determine the condition required for attacks to be uniquely identifiable. It was shown that if all atomic attacks possess at least some unique alerts, then the problem reduces to the trivial case of the defender focussing solely on the unique alerts and enacting predetermined best tactical responses. It was shown that, if attacks exist which generate non-unique alerts, it cannot be guaranteed, particularly within large-scale real-world networks, that the defender can correctly identify the enacted attack. This was determined by analyzing an idealized case and showing that, even within this defender advantageous situations, such a proof was not possible. From this result, it can be concluded that desiring optimal tactical responses implies a need for risk analysis.

10. References

- [1] F. Valeur, G. Vigna, C. Kruegel, and R.A. Kremmer, "A Comprehensive Approach to Intrusion Detection Alert Correlation", IEEE Transactions on Dependable and Secure Computing, Vol. 1, No. 3, pp. 146-169, 2004.
- [2] K. Julisch, "Clustering Intrusion Detection Alarms to Support Root Cause Analysis", ACM Transactions on Information and System Security, Vol. 6, No. 4, Nov. 2003, pp. 443-471.
- [3] H. Debar and A. Wespi, "Aggregation and Alert Correlation of Intrusion Detection Alerts", Conference on Recent Advances in Intrusion Detection (RAID 2001), pp.85-103, Oct., 2001.
- [4] F. Cuppens, "Managing Alerts in A Multi-Intrusion Detection Environment", 17th Annual Computer Security Applications Conference, Dec. 10-14, 2001, pp.22-31.

- [5] P. Ning, Y. Cui, D. Reeves and D. Xu, "Tools and Techniques for Analysing Intrusion Alerts", *ACM Transactions on Information and System Security*, Vol. 7, No. 2, pp. 273-318, May 2004.
- [6] H. Debar and A. Wespi, "The Intrusion-Detection Console Correlation Mechanism", 4th Workshop on Recent Advances in Intrusion Detection (RAID '2001), Oct. 2001.
- [7] B. Morin, L. Me, H. Debar, and M. Ducasse, "M2D2: A Formal Model for IDS Alert Correlation", 2002 International Symposium on Recent Advances on Intrusion Detection, pp. 115-137, 2002.
- [8] D. Xu, and P. Ning, "Alert Correlation through Triggering Events and Common Resources", 20th Computer Security Applications Conference (ACSAC '04), 2004.
- [9] A. Valdes and K. Skinner, "Probabilistic Alert Correlation", 2001 International Symposium on Recent Advances in Intrusion Detection (RAID 2001), pp.54-68, Oct., 2001.
- [10] S. J. Templeton and K. Levitt, "A Requires/Provides Model for Computer Attacks", New Security Paradigms Workshop 2000, Cork Ireland, Sept. 19-21, 2000
- [11] P. Ning and D. Xu, "Hypothesizing and Reasoning about Attacks Missed by Intrusion Detection Systems", *ACM Transactions on Information and System Security (TISSEC)*, Vol. 7, No. 4, pp. 591-627, November, 2004.
- [12] S. Noel, E. Robertson, and S. Jajodia, "Correlating Intrusion Events and Building Attack Scenarios Through Attack Graph Distances", 20th Computer Security Applications Conference, pp. 350-359, Dec., 2004.
- [13] O. Sheyner, J. Haines, S., Jha, R. Lippmann, and J.M. Wing, "Automated Generation and Analysis of Attack Graphs", 2002 IEEE Symposium on Security and Privacy, pp. 273-284, 2002.
- [14] R. B. Myerson, *Game Theory: Analysis of Conflict*, Harvard University Press, 6th Ed., 2004.
- [15] S. Floyd, "Difficulties in Simulating the Internet", *IEEE/ACM Transactions on Networking*, Vol.9, No.4, August, 2001, pp. 392-403.
- [16] V. Paxson, "End-to-End Internet Packet Dynamics", *IEEE/ACM Transactions on Networking*, Vol.7, No.3, pp. 277-292, June, 1999.
- [17] Snort - The Open Source Network Intrusion Detection System, <http://www.snort.org>, 2005.
- [18] M. Arboi, *The Nessus Attack Scripting Language Reference Guide*, 2002, http://www.nessus.org/doc/nasl2_reference.pdf.
- [19] C. Kruegel, D. Mutz, W. Robertson, G. Vigna, and R. Kemmerer, "Reverse Engineering of Network Signatures", to appear in *Proceedings of the AusCERT Asia Pacific Information Technology Security Conference*, Gold Coast, Australia, May 2005.
- [20] MIT Lincoln Laboratory, *Lincoln Lab Data Sets*, http://www.ll.mit.edu/IST/ideval,data,data_index.html, 2000.
- [21] DEFCON, *Def Con capture the flag contest*, <http://www.defcon.org>
- [22] J. Haines, D.K. Ryder, L. Tinnel, and S. Taylor, "Validation of Sensor Alert Correlators", *IEEE Security & Privacy*, January/February, 2003, pp. 46-56.
- [23] D. Anderson, M.Fong, and A. Valdes, "Heterogeneous Sensor Correlation: A Case Study of Live Traffic Analysis", 3rd IEEE Information Assurance Workshop, June, 2002.
- [24] A. Householder, K. Houle, and C. Dougherty, "Computer attack trends challenge Internet security", *IEEE Computer*, Vol. 35, No. 4, pp. 5-7, April, 2002.
- [25] F. Cuppens, S. Gombault, and T. Sans, "Selecting Appropriate Counter-Measures in an Intrusion Detection Framework", *Proceedings of the 17th IEEE Security Foundations Workshop (CSFW'04)*, 2004.
- [26] O.M. Dain and R.K. Cunningham, "Building Scenarios for a Heterogeneous Alert Stream", *Proceedings of the 2001 IEEE Workshop on Information Assurance and Security*, West Point, New York, pp. 231-235, June 5-6, 2001.
- [27] P.A. Porras, M.W. Fong, and A. Valdes, "A Mission-Impact-Based Approach to Alarm Correlation", 2002 International Symposium on Recent Advances on Intrusion Detection, pp. 95-114, 2002.