High level information fusion for tracking and projection of multistage cyber attacks

Shanchieh J. Yang a,*, Adam Stotz b, Jared Holsopple b, Moises Sudit c, Michael Kuhl d

a Department of Computer Engineering, Rochester Institute of Technology, 83 Lomb Memorial Dr., Rochester, NY 14623, USA
b Information Fusion Division, Calspan-UB Research Center, 4455 Genesee Street, Buffalo, NY 14225, USA
c Department of Industrial and Systems Engineering, University at Buffalo, 438 Bell Hall, Buffalo, NY 14260-2050, USA
d Department of Industrial and Systems Engineering, Rochester Institute of Technology, 83 Lomb Memorial Dr., Rochester, NY 14623, USA

Received 30 June 2006; received in revised form 11 June 2007; accepted 21 June 2007
Available online 28 June 2007

Abstract

The use of computer networks has become a necessity for government, industry, and personal businesses. Protection and defense against cyber attacks on computer networks, however, are becoming inadequate as attackers become more sophisticated and as the networks and systems become more complex. Drawing analogies from other application domains, this paper introduces information fusion to provide situation awareness and threat prediction from massive volumes of sensed data. An in-depth discussion is provided to define fusion tasks for cyber defense. A novel cyber fusion system is proposed to address specifically the tracking and projection of multistage attacks. Critical assessments of the developed attack tracking and threat projection sub-components are provided with simulation results. This pioneering work elaborates the benefits, limitations, and future challenges of high level information fusion for cyber security.

Keywords: Cyber security; Information fusion; Situation and threat assessment; Alert correlation

1. Introduction

High level information fusion is arguably still in its infancy for not only cyber security but also many other application domains. There has been work promoting high level information fusion. Glinton et al. [1] attempted to predict future enemy moves by fusing sensed activities, doctrine and terrain information. Sidenbladh et al. [2] proposed to enhance the level of threat awareness by comparing different threat predictions. A task force was formed to improve speaker recognition by fusing heterogeneous knowledge sources [3]. Lee and Llinas [4] investigated means to combine probabilistic and logical models to predict the intent of military attacks. While it is debatable whether the above studies are indeed high level information fusion, they provide glimpses of benefits and challenges in identifying and tracking more than ‘objects’.

Indeed, fusion to provide information and knowledge beyond identifying objects is, perhaps the only certain definition for high level information fusion. Relatively extensive work has been conducted to enhance the situation awareness and threat prediction for application domains where systems operate in a physical environment. In the cyber domain, however, little has been accomplished. Bass, in his paper published in 2000 [5], advocates the need of information fusion to fill the void of intrusion detection. Since then, much work, e.g., [6–14], has been devoted to correlate Intrusion Detection System (IDS) alerts, so as to provide better situation awareness of cyber attacks. Correlating alerts provides the network security analysts a view of the attacks, instead of a view containing an enormous volume of alerts, many of which could be irrelevant. While this set of work certainly helps advance cyber defense, it is still far from accomplishing the goal of revealing the
relationship between the attacks and the projection and impact of the attacks.

How may we accomplish high level information fusion in the cyber security domain? This is the question this paper intends to answer. We shall, however, first discuss in what sense high level information fusion may benefit cyber defense. The term ‘high level information fusion’ comes from the Joint Directors of Laboratories’ (JDL’s) definition of a fusion model [15,16] and its recent revision in 2004 [17]. A general interpretation of high level fusion based on JDL’s model, includes Level 2 situation assessment and Level 3 threat and impact assessment. Based on Endsley’s model [18], high level fusion may involve comprehension and anticipation. These definitions hint at the type of questions information fusion may address: “How many and where are the current attacks?”; “Are assets X and Y threatened?”; “By which adversary and with what attack method?”; “What will be the damage if an attack is successful?”; and “Is the defense mechanism sufficient to neutralize each attack?”.

A common denominator in the above questions is the object for assessment – cyber attacks. Note that, assuming a reasonably secured network, each cyber attack may take multiple steps to gain privilege(s) and penetrate into the network core(s). Tracking and projecting these multistage attacks, thus, are the critical enablers for answering the aforementioned questions. Unlike traditional target tracking problems, vulnerabilities in cyber space, however, are open to the world, available to anyone, anywhere, with the capability and intent to exploit them. In addition, network and system configurations are constantly changing due to security patches and the needs for new software and features. As a result, both the network and computer system vulnerabilities (the blue team) and the attacker motivation and attack methods (the red team) are diverse and constantly evolving. These characteristics present additional challenges to the problem of tracking and projecting cyber attack actions – not to mention the overwhelming volume of data from continuously sensed malicious activities in different parts of a network. This work proposes and discusses an integrated information fusion system to assist analysts in uncovering information and knowledge from the immense data generated within the dynamic, unbounded cyber space.

The contributions and organization of the paper are as follows. Recognizing cyber attack characteristics, Section 2 breaks down fusion tasks in assisting cyber defense. This discussion is based on our experience working on high level information fusion, and offers an unique and first-of-its-kind view of technical issues and potential solutions for cyber security. Section 3 illustrates a proposed high level fusion based cyber defense system, that aims at providing situation, threat and impact assessment of cyber attacks. The system is based upon our work on INFormation Fusion Engine for Real-time Decision-making (INFERD) [12,19] and Threat Assessment for Network Data and Information (TANDI) [20,21]. An inside look of INFERD and TANDI, in how they track and project cyber attack actions, is also provided in Section 3. Section 4 presents simulation results revealing the capabilities and shortcomings of INFERD and TANDI. This section discusses not only the simulation results but also a challenging, yet rarely discussed issue – the evaluation of a high level fusion system for cyber security. The paper is concluded in Section 5.

2. Cyber defense as an information fusion problem

Effectively combating cyber attacks involves improving upon current capabilities in a spectrum of cyber related problems. These problems include malicious activity/software detection and identification, trace-back, (e)ncryption, authentication, and network management among others. Solutions to these individual problems may not be satisfactory on their own, but a combined usage of them can provide a robust cyber defense capability never before realized. Research associated with each of these individual problems deserves entire papers to do them proper justice. This work assumes well administered networks and many of the security components are in place, so that even sophisticated attackers will need to execute multiple operations before reaching critical files or compromising core network operation. Based on such premise, this work identifies the roles of information fusion for cyber security, and focuses on the lacking components – tracking and projection of multistage cyber attacks.

Well administered and secured networks typically utilize firewalls to divide the network into multiple access domains, and have IDSs of various types to monitor malicious activity within and across these domains. The objective of information fusion here is to identify, to track, and to project multistage cyber attacks at their early stages so as to prevent critical data losses and operation interruptions. Multistage cyber attacks, which span over time and across multiple machines and access domains, are considered to be the virtual objects of our fusion processes, with the observables being IDS alerts. Fig. 1 shows the three main fusion tasks and our interpretation of their association to JDL fusion levels and Endsley’s model.

2.1. Sub-object malicious activity detection

Alert messages produced by IDSs via monitoring, e.g., packet and OS kernel activities, are considered as outcomes
of sub-object data association and estimation. This draws correlations with traditional JDL L0 fusion processes such as pixel level filtering and feature detection in image fusion applications. It is sub-object because IDS alerts are observables of malicious activity forming the complex and virtual objects – the multistage cyber attacks. By correlating the observables or alerts, which can be redundant, incomplete, and possibly conflicting, one may identify and trace multistage cyber attack actions.

There is a plethora of research and commercial developments for malicious activity or intrusion detection. Intrusion detections may be classified in terms of whether it monitors network traffic, e.g., [22,23] or host activity, e.g., [24,23]. By technical approach, most solutions are based on signatures of known exploits, e.g., [22,23], or by detecting anomalies in observed activity, e.g., [24]. This work assumes sufficiently accurate and complete intrusion detection, and the IDS alerts are input to our main mission: tracking and projection of attack actions.

2.2. Alert association and tracking

Associating IDS alerts that are hypothesized to be of the same multistage cyber attack reveals traces or tracks of multistage cyber attacks on a network. This alert association task may be considered as a JDL L0/L1 problem or as part of Perception based on Endsley’s model. It is synonymous with target tracking in traditional fusion applications, where position and velocity observations of a physical object is tracked. An error in such process would be alerts of the same multistage attacks being placed into separate tracks resulting in fragmented tracks.

The outcomes of this fusion task are track estimates of attack actions on a network. The attack tracks, with a volume presumably much more manageable than that of alert messages, are the enablers for effective comprehension and assessment of the attacks and their effect to the network. JDL’s L2 situation assessment and Endsley’s Comprehension module both imply reasoning upon the relationships between attack track estimates and their network context. For examples, one may measure the severity of each attack track based on the systems or services compromised, or identify coordinated attacks exhibiting similar characteristics or progressing toward a common goal.

Many alert correlation techniques have been proposed in the past few years, e.g., [6–14]. Though with a common goal of reducing from the abundant alert messages, this set of work adopts different approaches and achieves diverse outcomes. Among which, Ning et al. [8,10] propose to correlate alerts into hyper-alerts based on prerequisites and consequences of cyber attacks. The prerequisites define the conditions in which an alert can happen. The consequences represent the possible outcomes after the prerequisites have been satisfied. Matching already instantiated consequences to prerequisites of new alerts is the operation of correlating alerts, and it generates hyper-alert correlation graphs, each representing a group of alerts belonging to a single attack. This approach is able to detect failed attacks if provided with network configuration information and it also demonstrates a low false alert rate. King et al. [13] modify the Linux kernel to develop a prototype – BackTrack. Using process information from the operating system, BackTrack identifies activities causing IDS alerts reported and dynamically track attacks across multiple hosts. Valeur et al. [9] provide a comprehensive illustration of a 10-step process, from alert normalization to multi-step correlation and to prioritization. In multi-step correlation, they consider two attack scenarios: recon-breakin-escalate and island-hopping. Valeur et al.’s work offers a complete package for intrusion correlation analysis and has shown significant alert reductions.

Most work in correlating alerts requires the use of complex models, representing possible causal and sequential relationships between attack actions on specific networks. An attack action, in general terms, consists of the exploitation method used and the source and target machines involved in the attack. Considering both exploitation methods and network contexts may lead to a large number of models or complex models difficult to use and update. This work proposes to perform alert correlation with minimal network information. As will be described in Section 3, our approach leads to real-time tracking of attack actions with the use of a single attack model, representing generic progression of multistage cyber attacks.

2.3. Threat projection and impact assessment

Traces or track estimates or attack actions make it possible for threat projection. Threat projection may be based on network configurations and vulnerabilities or attacker behaviors. This draws analogy in traditional fusion work on, e.g., projecting trajectory of missile attacks, where both the environment (e.g., air or water) and the missile capability determine the estimated trajectory. In cyber space, however, attacker’s intent and capability may be diverse and changing in a very short time scale, as it is for network configurations and vulnerabilities. These characteristics present additional challenges for projecting cyber attack actions.

Threat projection and impact assessment have similar, if not the same, interpretation in the fusion community; they both fall under JDL L3 fusion level and Endsley’s framework of Projection. Yet, we consider two different fusion tasks. Threat projection, as illustrated above, focuses on predicting future attack actions, i.e., where in the network will be attacked and what exploit(s) will be executed? Impact assessment, on the other hand, may be interpreted as estimating damages caused by current and future attack actions. Estimating damages or effects of attack actions on a network clearly depends on network contexts, some of which may not be automatically gathered, such as the criticality of a machine (due to data stored or its operational role) in the network. Damages to be caused by future actions also depend on threat projection. This work focuses
primarily on threat projection, with the system designed to accommodate impact assessment.

Little has been accomplished, to our knowledge, for cyber threat projection. Common approaches in evaluating and projecting courses of actions (not necessary for cyber attacks) include the use of Bayesian Network and Hidden Markov Models. In the cyber domain, much work has been based on modeling the sequences over which system vulnerabilities may be exploited in a network, e.g., [25–27]. Qin and Lee [28] proposed to adapt attack plan model using Bayesian Network as attack actions are detected. Though theoretically sound, one problem of these approaches is the complexity involved for developing and/or maintaining the course of action models, which encompass the diverse exploitation methods, intents, and network configurations. This work proposes to simplify the problem by separating the modeling of network configuration from the modeling of cyber attack methods. As attack actions are detected, predictions are performed independently using the two models, then fused to determine the targeted entities (e.g., machines or services) in the network.

2.4. Other tasks

Tracking and projection of cyber attack actions are two primary fusion tasks, enabling reasoning upon and across individual track estimates. These reasoning tasks, which are out of the scope of this paper, may determine coordinated attacks or uncover missing evidence, including those undetectable due to insider threats or social engineering. In addition, as system, network, and attack technologies evolve, it is essential to refine models or even the fusion processes. It is our hope that this work on tracking and projection will open up new research avenues for both the fusion and the cyber security communities.

3. High level information fusion assisted cyber defense

Identifying the needs for tracking and projection of cyber attack actions, two cyber fusion systems have been developed. INFERD efficiently correlates IDS alerts to identify individual multistage attacks [12] and provides situational measures of the identified attacks [19]. TANDI fuses information extracted from each attack track estimates, to determine threatened entities and to differentiate them by assigning threat scores [20]. This section illustrates the two fusion engines, starting by providing an overview of the integrated system architecture.

3.1. Overall system architecture

INFERD and TANDI were initially designed assuming the separation of JDL’s Level 2 situation assessment and Level 3 threat and impact assessment. A key lesson learned is that the two systems are tightly coupled – recognizing situations involves determining the threats, and determining threats highly depends on how and how well situations may be defined. As a result, an overall system architecture is developed and shown in Fig. 2.

IDS alerts and two reference models are inputs to the overall system. The two models, the Potential Attack Sequence Template and the Information Exposure Sequence Map, are to represent all potential sequences of attack methods a cyber attacker can use and the orders over which vulnerabilities may be exposed due to network and system configurations, respectively. Suspicious activities detected by IDSs will be traced based on the network independent Potential Attack Sequence Template (referred to as the Guidance Template here after). This tracking process aims to provide dynamically changing views of individual attacker behavior. Overlaying the track estimates on the network-specific Information Exposure Sequence

![Fig. 2. An architectural view of the cyber information fusion system, providing real-time tracking and projection for multistage cyber attacks.](image-url)
Map (referred to as the Information Graph hereafter) unmasks network entities and information exposed to each identified multistage attack. Combining projections based on individual attacker behavior with those based on exposed network entities reveals the levels of threat posed on each entity.

The **Guidance Template** is a directed graph intended to model and guide the tracking of all potential courses of known attack actions. Modeling all known courses of actions may be too complex a task to accomplish, if one were to account for specific system vulnerabilities and exploit methods, both of which are diverse and fast changing. Our Guidance Template, therefore, is independent of specific network and system configurations and models from a categorical perspective of the attack methods. This approach significantly reduces model complexity and allows for real-time alert correlation and tracking of attack actions.

Fig. 3 shows a diagrammatic representation of the Guidance Template, at its highest level with each node representing a category of attack methods. Each template node is associated with a set of IDS alerts that are indicative to the attack, i.e., the observables. This categorical approach, along with its network independence, eliminates the need for multiple templates in guiding the tracking process for different network configurations and diverse exploitation methods. The directed and attributed arcs connecting the template nodes are to represent the possible causal and attributed relationships between the categories of attack actions. Without favoring any specific course of action, the template does not require conditional probabilities assigned for the directed arcs, which is typically needed.

---

Fig. 3. The Potential Attack Sequence Template (a.k.a. Guidance Template) and the categories of attack methods used by INFERD and TANDI.
with Bayesian approaches, e.g., [25,29,28,27]. Note the ‘External’ versus the ‘Internal’ network rings. This is to reflect the typical attack sequence where external systems are compromised and become stepping stones for attacking internal systems. Additional network rings may be added if multiple levels of internal networks exist.

As malicious activities are detected, IDS alerts will be associated with existing or form new track estimates based on the Guidance Template. Network independent alert correlation, though efficient, lacks information for situation assessment and attack projection. The Information Graph, a directed graph modeling the sequence over which network entities and information may be exposed, provides network information. By network entities and information, we mean the existence of host or servers, access privileges, and critical data files. The model does not intend to capture the exact service and OS vulnerabilities as existing work does, e.g., [25,26], which may present a significant overhead in model development and update. The creation of the Information Graph will be based on the logical connectivity information of the network and a per-node information graph template, which detail will be discussed in Section 3.3.

Overlaying track estimates on the network specific Information Graph provides contextual information for situation assessment. In addition, it reveals the threatened entities in the network due to their exposure or privileges gained at each step of traced multistage attacks. The use of two models, one intended for attacker behavior and the other intended for network entity exposed, is the premise of our approach. In the cyber domain, the skills and the motivations of hackers may vary significantly. What the hacker may do can be very different from what the analysts expect based on their understanding of the network vulnerabilities and the criticality of different entities. The use of the two models not only simplifies the modeling process, but also enables combination of projection estimates independently from the red team’s and the blue team’s perspectives.

3.2. INFERD for tracking

INFERD aims to efficiently correlate and associate alerts in forming individual multistage attack tracks across temporal and virtual network space. A performance goal for this tracking process is to perform at real-time speeds, to handle streamlined incoming IDS alerts. IDS alerts and the Guidance Template (recall Fig. 3) are the inputs to INFERD’s tracking process. The overall tracking process can be broken down into four subtasks: (1) Data Alignment, (2) Connotation Elicitation, (3) Data Association, and (4) Track Update.

A ‘feature tree’ is associated with each template node in the Guidance Template. The feature trees prescribe the semantic and syntactic bridge between heterogeneous IDS alert messages and the categorical concept presented by the template node. The Data Alignment process homogenizes alert messages to provide common referencing. This process requires a priori definitions of alerts associated with each template node, and may be referenced through the Common Vulnerabilities and Exposures (CVE) dictionary [30] or the National Vulnerability Database (NVD) [31]. The Connotation Elicitation process is operated in real-time, and uses the feature tree to map detected malicious actions with one of the attack categories defined in the Guidance Template. The Data Association process then associates incoming IDS alerts with existing attack tracks or uses them to form new track estimates. Attack tracks then are updated and provided with measurements indicative of their fidelity and/or severity. Consider the following IDS alerts:

1. ICMP Ping NMAP (62.34.46.54 → 45.34.12.1)
2. SCAN nmap fingerprint attempt (38.244.61.9 → 45.34.12.2)
3. x86 mountd overflow (62.34.46.54 → 45.34.12.1)
4. gobbles SSH overflow (62.34.46.54 → 45.34.12.1)
5. SCAN cybercop os SFU12 probe (38.244.61.9 → 45.34.12.2)
6. WEB-MISC windmail.exe access (38.244.61.9 → 45.34.12.2)
7. ICMP Ping NMap (45.34.12.1 → 45.34.13.1)
8. EXPLOIT RADIUS MSID overflow attempt (45.34.12.2 → 45.34.12.2)
9. chown command attempt (62.34.46.54 → 45.34.12.1)
10. MS-SQL:PROCEDURE-DUMP (45.34.12.2 → 45.34.12.2)

Two multistage attacks are embedded within these messages and are correctly traced by INFERD. Fig. 4 shows the two attack tracks. The first attack discovers an external server (45.34.12.1) on the destination network then unsuccessfully tries to exploit it with a ‘mountd overflow’. The unsuccessful attempt is then followed by a successful ‘SSH overflow’ attack and proceeds to discover a server (45.34.13.1) on the internal network. The second attack targets another external server (45.34.12.2) and intrudes as user then root, to obtain a SQL database.

The alert descriptions, e.g., ‘x86 mountd overflow’ and ‘gobbles SSH overflow’, are referred to as the feature (leaf) nodes of the feature tree associated with the template node ‘External Intrusion Root’. The Connotation Elicitation process identifies the categorical template node for the incoming alerts. The Data Association process then determines whether the incoming alert will (1) instantiate a new template node, (2) instantiate a new feature node for an existing template node, or (3) amend to an existing feature node of an existing template node. In the example above, Alerts 1, 2, 3, 6, 7, 8, and 9 will instantiate a new template node, and Alerts 4, 5, and 10 will instantiate a new feature node to the existing template node. While the example shows a MAX function to aggregate the feature nodes, INFERD adopts Yager’s Generalized Ordered Weighted Average (GOWA) function [32,33] for a general class of
aggregation. This allows for aggregation of probability values or scores in assessing the confidence or severity of each attack from the alert messages.

If a new template node is instantiated, it will either (1) start a new attack track or (2) be associated with an existing attack track. This decision is based on the directed arcs and their attributes defined in the Guidance Template. Continuing from the above example, Alert 3 instantiates a new template node ‘External Intrusion Root’, which is one of the assessor nodes of ‘External Scanning’ in the Guidance Template. In addition, the source and target IPs provided by the first and the third alerts are the same. These conditions together determine that the new ‘External Intrusion Root’ node will follow the existing ‘External Scanning’ node. On the other hand, since Alert 2 has a different source and target IPs than Alert 1 does, it is determined that a new attack track will be instantiated. Note also that Alert 7, which has a different target IP from all others, is considered as a stepping stone attack and is part of the first attack because its source IP is the same as the target IP of Alerts 1, 3, and 4.

Distinguishing external versus internal IPs, where there can be multiple internal IP domains, is the only network dependent information required for INFERD’s tracking process. Once the associations of one or more IDS alerts with new or existing tracks are determined, INFERD will update the attack tracks and calculate various situation measurements associated with the detected attack tracks. A set of track measurements have been implemented, but, more importantly, INFERD is designed to allow for customized measurement plug-ins [19].

The strength of INFERD lies in the use of the categorical and network-independent attack Guidance Template. This and the streamlined tracking process allows for timely, efficient, and accurate alert correlation and attack identification. Tests have shown that the fusion core of INFERD can process over 3000 alerts per second. A conservative estimate of the implementation that includes I/O with JMS and writing results to a MySQL database, which are not optimized, reveals about 100 alerts per second. Both numbers exceed significantly the demands of processing 24,000 alerts per day for typical enterprise networks. More test results will be shown in Section 4.

3.3. TANDI for projection

Given the tracked attack actions, TANDI aims to determine the next likely targets for each detected multistage attack. The track estimates based on the Guidance Template reveals attack methods taken by each attack, but not network specific information. An Information Exposure Sequence Map will be created a priori for each enterprise network to model the order over which virtual network entities can be compromised. Overlaying a track estimate onto the Information Map reveals which machines, privileges, and databases are susceptible to the attack, depending on which others are compromised. Combining projections based on both models is key to TANDI’s approach in cyber attack projection.

An attack track is a subgraph of the Guidance Template and contains information about not only which reconnaissance and exploitation methods that have been used but also the sequence and statistics of these actions. By interpreting the Guidance Template as an expert model of hacker behavior, the asserted nodes and the successors of the asserted actions on each attack track will comprise the set of next possible actions of that attack. Assigning conditional probabilities on the Guidance Template transforms the problem into a classical Bayesian analysis. In the case where no consensus can be reached for assigning the conditional probabilities, a weighting scheme can be applied to distinguish between, e.g., the asserted action and the successors of the asserted actions. The current implementation of TANDI adopts this weighting approach.

The network specific Information Map is created based on the logical connectivity of the network and a per-node
information graph template. By logical connectivity, we mean the order in which a hacker can compromise the computers in the network. A Logical Topology is required to be developed a priori, by SMEs and based on information gathered via automated network vulnerability scanners, such as Nessus [34] and NMap [35]. Each node in the Logical Topology represents one or more computers with similar configurations, and each directed arc means that compromising the source computer(s) will expose the destination computer(s).

Fig. 5 shows an example Information Graph based on a simple Logical Topology and the information graph template. The information graph template represents the inherent three level hierarchy of network entities, i.e., the existence of the machine, the privileges and the file and system configurations. Each machine in the Logical Topology will instantiate a clone of the information graph template and become part of the overall Information Graph. The information graph template is inspired by Phillips and Swiler [25] and Liu and Man [27]. The Logical Topology in this figure represents a computer network with a web server external to the firewall, an internal subnet, and an internal server. Once developed, each node in the Information Graph will be associated with a subset of nodes in the Guidance Template, which is the set of attack methods that can compromise that network entity.

For a given attack track, each feature node (i.e., an alert message) has a source IP and a target IP. These IPs correspond to machines in the Logical Topology, and consequently the entities deduced from these machines in the Information Graph. This set of entities may be narrowed down to the set of compromised entities by finding those associated with asserted attack method nodes in the attack track. The successor nodes of these compromised entities in the Information Graph are then the potential next targets of the multistage attack.

Many factors can be used to differentiate the threat levels of potential next targets of a detected attack. They include hacker’s intention, the value of the target, the skill level of the hacker, vulnerabilities exposed, etc. Theses factors may produce similar or contradicting predictions. TANDI adopts Dempster–Shafer rule [36] to combine these estimates – assuming each entity $i$ will be assigned a threat score $t_i$: $[0,1]$. A threat assessment algorithm is implemented to investigate the benefit of estimation based on attack methods used versus machines compromised.

The threat assessment algorithm assigns weights to already executed attack actions ($w_{a\text{Current}}$), the successor nodes of already asserted attack actions in the Guidance Template ($w_{a\text{Next}}$), already under attacked machines ($w_{m\text{Current}}$), and the successor nodes of these machines in the Logical Topology ($w_{m\text{Next}}$). These weights are combined to determine the threat score of each entity defined in the Information Graph, by examining the attack graph nodes and the Logical Topology nodes associated with the entity. Note that there are various ways to combine the weights. TANDI employs an arbitrary function that simply sums up the weights associated with each entity, yet only account for each weight type at most once. In other words, the maximum threat score for an uncompro-
mised entity will have a threat score of \( t_i = w_{a\text{Current}} + w_{a\text{Next}} + w_{m\text{Current}} + w_{m\text{Next}} < 1.0 \), where all compromised entities will have a threat score of 1.0.

Continue from the example attack track shown on the top of Fig. 4 and consider the Logical Topology and Information Graph shown in Fig. 5. Let the IP for the external web server be 45.34.12.1, 45.34.13.1 for the internal server, and 45.34.13.0/24 be the subnet address for all other hosts in the ‘Subnet’ node. Consider a weight assignment of \( w_{a\text{Current}} = 0.3, \ w_{a\text{Next}} = 0.1, \ w_{m\text{Current}} = 0.4, \) and \( w_{m\text{Next}} = 0.15 \). Fig. 6 shows the threat scores next to each network entity in the information graph.

Notice that TANDI identified ‘Server Root’ to be the next most likely entity to be compromised. This is because the IDS alerts indicating a root intrusion of the Internal Server fall into both the categories of current attacks and next attacks. The Internal server was already attacked, so its threat score is further increased by \( w_{m\text{Current}} \). A user-level intrusion was identified to be the second most likely next step. This is due to the fact that a lesser weight is assigned for the attack methods than for the attacked machines.

The threat assessment algorithm described above is by no means optimized, yet the idea is to examine the benefit of prediction based on attack methods versus compromised machines. Simulation results of TANDI have shown good prediction accuracy even with this simple algorithm. TANDI, however, fails to predict coordinated attacks (multiple attack tracks are actually executed by the same hacker or hacker group) or insider threats (no evidence observed). These shortcomings are expected, and, more importantly, sheds light on how one may combat the commonly perceived ‘hard-to-detect’ cyber attacks. The key idea is to take advantage of bad predictions with obvious signatures, such as:

- A few badly predicted attacks may be a single coordinated attack if a substantially better prediction accuracy can be achieved when they are assessed together.
- A false negative prediction showing unauthorized access of critical data with a privilege not likely to be exposed to any external personnel is likely an insider threat or social engineering attack.

Clearly, not all the coordinated attacks, insider threats, social engineering attacks, etc. can be detected based on the observations of bad predictions. In fact, IDS failures can lead to bad predictions. Nevertheless, this approach shall help enhance the awareness of these hard-to-detect attacks when it happens and is the foundation of the adaptiveness of the high level information fusion system described in Section 3.1.

4. System evaluation

Formal methodology for testing a high level information fusion system is a largely uncharted area of research. There have been discussions on the effort of building evaluation systems for general high level information fusion systems [37,38], but they do not address the specific problems associated with the cyber environment. Meanwhile, in responding to DARPA’s grand challenge, MIT Lincoln Laboratory has produced cyber attack data sets [39] in 1998, 1999, and 2000. While being largely used for experiments in early 2000’s, these data sets quickly become outdated since network and system vulnerabilities have changed significantly since then. In fact, obtaining up-to-date attack data is a key challenge in evaluating cyber fusion system. Our evaluation uses newly generated attack data with background traffic on virtual networks. The specific data sets used for testing INFERD and TANDI will be discussed separately in the next two subsections.

In addition to the challenge of generating up-to-date attack data, a rarely discussed issue in the literature is how to evaluate the outcomes of situation, threat and impact assessments. First, when alerts are correlated and grouped into individual attack tracks, it is unclear how to match the detected attack tracks to the ground truth attacks. Do the alerts grouped to an attack track need to be an exact match to the set of alerts belonging to a ground truth attack? Second, it is questionable how to define the performance metrics that accurately reflect the quality of tracking and projection of cyber attacks. Finding appropriate metrics is challenging because, unlike traditional physical attacks, cyber attacks are much more diverse and changing. A cyber attack can stop in the middle of a perfect execution simply because the hacker loses interest or is ignorant of the vulnerabilities exposed to him. Furthermore, current cyber sensing technology is far from sufficient to catch all stealthy cyber attack activities.
especially with social engineering and insider threats. Without reliable sensor data, one needs to cautiously interpret traditionally used metrics, such as false positives and false negatives. We will attempt to address some of these challenges in the next two subsections, where we discuss the tests performed for INFERD and TANDI.

4.1. Blind tests for INFERD

The core competency of INFERD is the use of a network-independent attack Guidance Template to create dynamic attack tracks. IDS alerts are fused to form a series of Attack Tracks ranked by various measurements, presenting a situational view to the network security analyst and allowing him to focus on which tracks/alerts are the most threatening to the network.

To test the performance of INFERD’s fusion algorithm in identifying attacks, a series of tests were conducted between the US Air Force Research Laboratory (AFRL) and the development team of INFERD. The tests calculated various metrics developed by AFRL researchers and program managers. Part of the results are shown in Fig. 7.

The tests consisted of five unlabeled attack scenarios of various sizes and types generated by Skaion Corporation on a VMware network with actual scripted attacks and artificially created background traffic. The term ‘unlabeled’ refers to the fact that the INFERD development team was given no indication of the type of attacks contained within the scenario, only the network configuration (of which only a small subset of information is needed) and deployed sensors utilized during the data set generation. Ground truth for the test scenarios consisted of Snort alerts, Dragon alerts, and IIS/Apache web log alerts grouped based on the attacker and labeled with attack types such as background scanner, client traffic, and attacker.

Fig. 7 contains four charts showing the four metrics, Recall, Precision, Fragmentation, and Mis-Association, scored by INFERD using one of the five unlabeled data sets. We show the results for one tested data set because the overall conclusion is the same, though the exact scores gathered for different data sets are different. The x-axis indicates the ‘threshold’ value used for determining the corresponding metric scores. The use of a threshold comes from the need of matching detected attack tracks to the actual occurred attacks in ground truth. We adopted AFRL’s methodology, which uses Jonker and Volgenent’s algorithm to solve this linear assignment problem. The threshold value is the minimum percentage of the alerts associated with a detected attack track in order for the attack track to be associated with an attack in ground truth. The higher the threshold value is, the more stringent, yet more definite it is for the attack tracks to match those in ground truth.

Each chart in Fig. 7 contains two sets of scores, to show the improvements of INFERD over the development phases. An initial thought of the scores is that they should...
provide definite, quantitative answers to the accuracy of INFERD in grouping alerts and identifying attacks. The experience, however, reveals the lack of certainty even in the evaluation of identified cyber attacks. Salerno et al. [44] and Tadda et al. [43] from AFRL have recommended various metrics to measure the “confidence,” the “purity,” the “cost utility,” and the “timeliness” of a high level information fusion system. The four metrics discussed here are to measure the confidence, i.e., how well INFERD detects the true attacks. According to Tadda et al. [43], the four metrics are defined as follows:

Recall measures the percentage of tracks detected in relation to the “total known” tracks (in the ground truth). Precision is the percentage of correct tracks detected in relation to the number of detections or proposed results. Fragmentation is defined as the percentage of tracks reported as multiple tracks that should have been reported as a single track and Mis-Associations are those known tracks that were incorrectly identified by the proposed tracks. Simply put, the sum of Precision, Fragmentation and Mis-Association should be 100%.

Through the development and testing phases, INFERD has improved in terms of all four metrics, particularly by introducing a defragmentation module. In the case of threshold = 75%, INFERD recalls over 70% ground truth attacks, out of which 80% were accurate (Precision), close to zero fragmentation, and less than 20% Mis-Association. INFERD was regarded as providing good “confidence” in identifying cyber attacks.

4.2. Simulated evaluation for TANDI

With attacks and their courses of actions identified, TANDI aims to project the next moves for each identified attack. The effectiveness of TANDI has been tested via simulation with randomly generated attacks on artificially created networks. TANDI was not tested using Skaiogen data sets because they contain insufficient network configuration information at the time. Two networks are created to reflect realistic network and system configurations with the associated operating systems and running services. The logical topologies of these two networks are shown in Fig. 8. Note that the systems that are susceptible to the same attack are grouped to reduce the complexity of the logical topologies. Network 1 is designed to represent the case with more segmented departments that can individually access the Internet, while Network 2 is to represent the case where the internal workstations hide behind levels of more tightly controlled servers.

Three sets of attacks are created for each network. Set 1 represents attacks that do not deviate from the a priori developed Guidance Template and Logical Topology. Sets 2 and 3 contain attack actions that do not follow the a priori designed reference models. These actions, referred to as abnormalities, are tested to examine how TANDI may react to unknown exploits, 0-day attacks and insider threats.

- **Set 1**: A set of cyber attacks originate from the Internet. Each of these attacks first compromises at least one external computer and may penetrate into one or more internal computers. Fifteen attacks, each with 6–21 events, are randomly generated for each network.
- **Set 2**: A set of attacks that contain abnormalities due to various reasons, except insider threat. Ten attacks, each with 4–12 events, are generated for each network.
- **Set 3**: A set of insider attacks. These attacks originate from one of the internal machines – a more extreme version of Set 2. Five attacks are generated, each of which has between three and six events, for each network.

The attack actions are associated with IDS alerts. TANDI assumes that the IDS alerts are grouped into attack tracks, such as the output from INFERD. Taking these grouped IDS alerts, TANDI performs fusion using the methodology described in Section 3.3 with the Guidance Template, the Information Graph, and the Logical Topology. For Set 1, attack tracks are randomly chosen among feasible paths on the directed Information Graph. For each action in the chosen attack track, an IDS alert will be randomly selected from the set of feasible alerts associated with the OS and/or running services of the corresponding system in the Logical Topology. Verifying IDS alerts against the vulnerability information contained in the Logical Topology allows us to avoid generating false positive IDS alerts and to focus on evaluating attack projection. For Attack Sets 2 and 3, the attack paths are chosen carefully to mimic potential unknown attacks and insider threats. The selection of IDS alerts remains random.

It is not unreasonable that a threat assessment algorithm is susceptible to flaws in the model developed a priori by SMEs. This is especially true for cyber attacks, where the attacker profiles not only exhibit large diversity, but are also constantly changing. The question, then, is whether TANDI cannot only predict well for known attacks, but also distinguish attacks that do not fall under the reference models from those that do. Fig. 9a and b show five statistical measures collected when simulating Attack Sets 1, 2, and 3 for the two test networks. The five statistics are defined as follows:

- **Average Normalized Compromising Score ($\hat{r}$)**: Given the assessed threat scores of network entities at each stage of an attack, the normalized compromising score for an entity is the normalized threat score (between 0 and 1) just before the entity is compromised. Averaging over all compromised entities by the attacks gives the Average Normalized Compromising Score.
- **Percent Assessee Reduction (%AR)**: The Percent Assessee Reduction is the percentage of uncompromised entities whose threat scores are below a threshold $\beta$ at each
step of an attack. The threshold $\beta$ represents the minimum normalized threat score for an analyst to be concerned about.

- **Percent Abnormal (%Abnorm):** Consider the network entities that are compromised but have their compromising score equal to zero. Each instance of such is referred to as an abnormality. Percent Abnormal is the percentage of abnormalities over the set of compromised entities.

- **Percent False Positives (%FP):** Considering the threshold $\beta$ and for a given attack, the Percent False Positives is the fraction of uncompromised entities that have normalized threat score larger than $\beta$.

- **Percent False Negatives (%FN):** Considering the threshold $\beta$ and for a given attack, the Percent False Negatives is the fraction of compromised entities that have normalized threat score smaller than $\beta$.

First notice the increasing %Abnorm for the three sets of attacks tested. This verifies the intuition where Set 3, which contains insider attacks, shall have more abnormal events than Set 2 and there should be no abnormality for Set 1. The $\bar{c}$ reflects the declining prediction accuracy of TANDI with more abnormalities in the attack set. Note that under a perfect threat assessment algorithm, all compromised entities should have the highest threat score among all uncompromised entities just before it is compromised, which then will lead to $\bar{c} = 1$. Caution, however, is needed in using the metric $\bar{c}$. An over-conservative scheme may give high threat scores to many entities, and will see a high $\bar{c}$ value. This type of scheme, which offer little benefit to the analysts, will exhibit low %AR and high %FP. As seen in Fig. 9a and b, this is not the case for TANDI.

The metrics of %AR and %FP seem to be good performance indicators of threat assessment. This may be true for Set 1 – the attacks that follow the reference models. Note, however, the little variation exhibited in terms of these two metrics when abnormal attack actions are introduced. Meanwhile, observe the obvious decreasing trend of $\bar{c}$ and the increasing trend of %FN as abnormalities increases from Set 1 to Set 2 and Set 3. This suggests that $\bar{c}$ and %FN, but not %AR and %FP, are indicative to the presence of attack events that deviate from a priori developed models. The reason for this is that more abnormalities increases the false negatives, but not false positives and neither does it change much for the number of entities catego-
rized as highly threatened. As for $\tau^*$, it reflects the severity of the compromised entities, and, thus, will be affected significantly when averaged with cases where the compromising score is zero.

An idea resulting from the above observation is to use $\tau^*$ and $\%$FN to detect coordinated attacks, where seemingly unrelated IDS alerts grouped into different attack tracks are actually results of either one attacker executing attacks from different machines or different attackers coordinating through means undetectable by IDSs. The realization of this idea is out of the scope of this paper. A preliminary discussion of such concept can be found in [21].

We now turn our attention to compare the performance of TANDI for the two different networks. Notice that, TANDI performs better for Network 1 than it does for Network 2 in terms of $\%$AR and $\%$FP, while it does worse for Network 1 in terms $\tau^*$. This is due to the logical connectivity exhibited by Network 2. Network 2 is designed in a way that, as a node in the Logical Topology is compromised, one can hardly distinguish between the potential next moves. For example, the external servers, though connected to each other in both networks, do not connect to workstations directly in Network 2. The internal servers in Network 2, on the other hand, connect directly to workstations but not interconnect to each other. Because of this structure, TANDI with its current unoptimized fusion algorithm, cannot distinguish and give different threat scores to similar entities that are directly connected to compromised entities and also susceptible to similar attack methods. Consequently, a low $\%$AR is exhibited and many entities will have similar high threat scores. The $\tau^*$ is higher for Network 2 because the actual moves of the attacks are still predicted by TANDI with high threat scores.

The results shown above exhibit promising performance with TANDI’s approach to model separately the attacker behavior and the network vulnerabilities. This modeling approach makes cyber threat assessment less dependent on a priori knowledge. We have also purposely shown the limitations of TANDI, to advocate the needs for optimized fusion algorithm to be built upon the novel modeling approach.

5. Conclusion

Defending against multistage cyber attacks has become a top priority for government, business, as well as individuals. Unlike physical spaces, which is governed by the law of physics and problems within have been investigated for generations, cyber space presents numerous new constraints and challenges along with the old. Built upon experiences in tackling traditional problem domains, information fusion fills the void of current cyber defense practices, where network security analysts are overwhelmed with data but little information can be extracted. The diverse and constant changing attacks and vulnerabilities present an even greater challenge for the analyst to uncover accurate assessment. This work proposes to track and project cyber attack actions, enabling a proactive mechanism in combating cyber attacks.

This paper explores this new territory of high level information fusion for cyber security, and offers our solution approach in providing real-time tracking and projection of attack actions. The success of our approach lies in two aspects. First, alert correlation and tracking adopt a network independent and categorical Guidance Template. The categorical abstraction allows for efficient alert associated to attack tracks. Once created, the track estimates are mapped to network specific information exposure graph, enabling contextual reasoning and situation assessment. Projections based on network independent attack sequences (the red team) and that based on network specific exposures (the blue team) are fused to provide combined estimates of future attack actions. Results from the testings of INFERD (for tracking) and TANDI (for projection) reveal advantages and limitations of the proposed cyber fusion system. We anticipate much work ahead leveraging accurate and efficient attack tracks and projections for a better cyber defense system.

Acknowledgements

This work is partially sponsored by the United State Air Force Research Laboratory (AFRL) and through the National Center for Multisource Information Fusion
(NCMIF) grant under the technical supervision of AFRL/IEFA. The authors would like to thank John Salerno, Michael Himan, and George Tadda for their comments toward the developments of INFERD and TANDI. The authors would also like to thank Michael Holender and Jason Kistner for their participation in the initial discussion of the paper.

References


