10-2006

Evaluating Threat Assessment for Multi-Stage Cyber Attacks

Shanchieh Jay Yang

Rochester Institute of Technology

Jared Holsopple

CUBRC

Moises Sudit

University at Buffalo

Follow this and additional works at: <https://scholarworks.rit.edu/other>

Recommended Citation

S. J. Yang, J. Holsopple and M. Sudit, "Evaluating Threat Assessment for Multi-Stage Cyber Attacks," MILCOM 2006 - 2006 IEEE Military Communications conference, Washington, DC, 2006, pp. 1-7. doi: 10.1109/MILCOM.2006.302216

This Conference Paper is brought to you for free and open access by the Faculty & Staff Scholarship at RIT Scholar Works. It has been accepted for inclusion in Presentations and other scholarship by an authorized administrator of RIT Scholar Works. For more information, please contact ritscholarworks@rit.edu.

EVALUATING THREAT ASSESSMENT FOR MULTI-STAGE CYBER ATTACKS

Shanchieh Jay Yang
Rochester Institute of Technology
Rochester, NY

Jared Holsopple
CUBRC
Buffalo, NY

Moises Sudit
CMIF, University at Buffalo
Buffalo, NY

Abstract—*Current practices to defend against cyber attacks are typically reactive yet passive. Recent research work has been proposed to proactively predict hacker's target entities in the early stage of the attack. With prediction, there comes false alarms and missed attacks. Very little has been reported on how to evaluate a threat assessment algorithm, especially for cyber security. Because of the variety and the constantly changing nature of hacker behavior and network vulnerabilities, a cyber threat assessment algorithm is, perhaps more susceptible than for other application domains. This work sets forth the issues on evaluating cyber threat assessment algorithms, and discusses the validity of various statistical measures. Simulation examples are provided to illustrate the pros and cons of using different metrics under various cyber attack scenarios. Our results show that commonly used false positives and false negatives are necessary but not sufficient to evaluate cyber threat assessment.*

I. INTRODUCTION

Threat assessment is common to our daily activities, as well as critical to military and intelligence operations. The various assessments, though different in nature, are similar in that the process involves information fusion. From making observations, recognizing the situation, to analyzing the potential threat and impact of future events, these activities fall into the JDL fusion model [1], [2]. Research has been conducted to formalize and automate the threat assessment process for application domains ranging from naval combat [3] to school safety [4]. Logically, the objectives of this set of work would be to achieve accurate assessment while reducing information overload. Unfortunately, little has been reported in the public domain, at least to our knowledge, on whether, how, and what to evaluate for a threat assessment scheme.

Due to the emerging use of computer and data networks for almost all business and military operations, combating cyber attacks has become a critical issue, yet different from that of traditional physical attacks. Current practices used for cyber defense are somewhat passive. Analysts or automatic network defense software must react to detected suspicious activities, by which time network entities may be already compromised. An equally, or arguably more serious problem is information overload. Recent research work on situation assessment [5], [6], [7] proposes means to filter out and correlate intrusion detection sensor (IDS) alerts, so as to identify and characterize potentially threatening

attacks. With identified attacks, a threat assessment scheme is proposed in [8] to proactively predict the next moves of cyber hackers.

Predicting cyber attacks is particularly challenging due to the variety and the constantly changing nature of hacker behavior, hacking methods, and network vulnerabilities. This is different from traditional physical attacks, e.g., missile attacks, where weapon capabilities and physical infrastructure are better profiled and unlikely to change on a daily or weekly basis. Consequently, the expectation of false positive and false negative rates for cyber threat assessment may be different from that for assessing threats posed by traditional physical attacks. The question then is how good a cyber threat assessment algorithm one should expect, and in terms of what statistical measures. In searching for the answer, to our surprise, there exists almost no work in the public domain suggesting how and what to evaluate for a threat assessment algorithm for any application domain. Drawing examples from other areas of research, benchmark databases and standard metrics are both available for proposed computer architecture [9] and image processing algorithms [10]. This work will examine in detail the issues involved for evaluating cyber threat assessment, propose a simulation framework, and compare the uses of different statistical measures for such evaluation. Simulation results will demonstrate that false positives and false negatives - two metrics traditionally use in other domains - may not be sufficient for evaluating cyber threat assessments.

The rest of the paper is organized as follows. Section II discusses the challenges of evaluating cyber threat assessment. A set of statistical measures are introduced in Section III. Section IV discusses the simulation framework and results, followed by the conclusion in Section V.

II. THE CHALLENGES OF EVALUATING THREAT ASSESSMENT FOR CYBER ATTACKS

A. Threat assessment for cyber attacks

A definition of threat assessment for cyber security was prepared by the ATIS Committee T1A1 in 1999 [11]. The definition described threat assessment as *a formal description and evaluation of threat to an information system*. Though formal and succinct, this definition offers little insight on what and how a threat assessment may be done for cyber attacks. Intuitively, assessing threats essentially

is to predict and project potential course of actions and consequences of observed attacks. Roy et al. [3] elaborated a similar concept and put forth a general definition for threat assessment.

For the ease of discussion and for comparison purposes, consider threat assessment for missile attacks. Upon detecting a missile, it is imperative to estimate the trajectory of the missile and, consequently, the target of the attack. Evaluation of the criticality of the targeted assets and the effectiveness of defensive and damage control options against the attack shall then determine the actual threat posed by the attack. In other words, assessing threats means estimating the impact that might be caused by an attack, which matches JDL's definition of Level 3 Fusion - Impact Assessment [2]. Note that the two-step process of threat assessment may be done by independent analysts: one predicts missile trajectory and the other estimates the severity of the attack. The projection of missile trajectory depends on the types of missile, the environment, the sensors used, and the laws of physics. It also needs to be done in a real time manner. Analysis of the attack severity, though also done in real time, is based on a priori data, since the population scale, the assets, etc. change in a much larger time scale than that of a missile attack. As can be imagined, an analyst may be able to react to one or few incoming missiles at the same time, but not to a large number of them.

Now consider the case of cyber attacks. Typical practices today have network security analysts monitor IDS alert reports and system logs to make real time decisions based on a priori knowledge of the network and system configuration. The key difference between cyber attacks and missile attacks is that little hacker profiling and no law of physics can be applied to predict the exact course of actions. Cyber hackers may have a variety of intentions, including hacking-for-fun, while missile attacks will almost certainly result in physical damages if no intervention is in place. Moreover, cyber attacks typically consist of multiple stages: reconnaissance of the system existence and configuration, followed by some form of intrusion and/or privilege escalation, and ended with implanting malicious software or stealing critical information. In any stage of a cyber attack, hackers may decide not to proceed or discover previously unknown opportunities. This stop of action or change of course makes it extremely difficult to predict accurately what the hacker will do next, not to mention estimating the impact of his actions.

Recognizing the potential complexity and unreliability involved for building cyber attack models, a threat assessment algorithm has been proposed [8]. The algorithm, called TANDI, separately models the hacker's capability and the opportunity he discovered during the attack, and

fuses the real time observations based on the two models to predict the hacker's next target in the network. The algorithm, though in its infancy and is under revision, shows promising results. More importantly, seemingly as the first attempt of its kind, analyzing TANDI has led to many open and interesting questions. Some of these questions will be discussed and addressed in this paper.

B. Accurate threat prediction?

A rarely discussed issue in the literature is what and how to evaluate a threat assessment algorithm. Logical choices of performance metrics are the commonly used "false positives" and "false negatives." Caution, however, should be placed on the definition and interpretation of them, especially for the case of cyber attacks.

Consider again the contrasting example of a missile attack. A false positive in assessing the threat posed by a missile attack may be due to either a mis-detection of the missile, a mis-calculation of the missile trajectory, or a mis-estimation of the consequence of the attack. Whatever the reason is, a false positive for missile attack assessment means a flaw in the detection and assessment system. The same holds for false negatives.

Evaluating threat assessment for cyber attack, however, is not as straight forward as that for missile attacks. Since a cyber hacker may arbitrarily decide to stop his attack or to change course in the middle of an attack, it may not be reasonable to attribute solely as a system flaw when an attack action that was predicted likely to happen but never happens. Likewise, it is possible that an ignorant hacker attacks a network entity that was not considered valuable by the analyst. This may result in a false negative, but is unclear if the blame should be placed on the threat assessment system. Perhaps more problematic are cyber attacks based on insider threats and social engineering. Insufficiency at the low level sensor capability makes it impossible for high level fusion scheme to function properly. In short, it is unclear whether the common perception of "false positive" and "false negative" can be directly applied to the cyber domain. Note that the argument here is not that false positives or false negatives are not suitable for the cyber domain; instead, it is that one needs to cautiously interpret false positives and false negatives for cyber threat assessment. Section III will discuss our interpretation of false positives and false negatives along with other statistical measures that may be used for assessing cyber attacks.

C. Cyber attack generation

In addition to the lack of statistical measures that can be used to evaluate cyber threats, it is also a challenge to find representative databases for network topologies,

system configurations, and cyber attacks. Due to liability, loss of reputation, and competition issues, cyber attack data does not seem to be available in the public domain [12], at least not to our knowledge. Note that a representative database needs to be updated quite frequently because the constantly changing network and system vulnerabilities and attack methods. Interestingly, the lack of validation data and, surprisingly, the lack of reported effort in validation is not uncommon in the research community conducting threat and impact assessment. Often times, a framework and an algorithm is proposed but little validation and comparison effort is put forth in the published work. Again, this can be due to various confidentiality and security issues. Given such limitations, an alternative is to create artificial cyber attacks and networks based on subject matter experiences.

A critical component in creating data for experiments is to introduce randomness, so as to reduce potential bias of the creator of the data. While a more rigorous Monte-Carlo simulation needs to be investigated, a preliminary attempt in generating attack data is put forth. Consider an arbitrary network topology with a set of IP and subnet addresses and services and operating systems running on each machine. Based on this information, a directed information entity graph is created to represent possible network entities. To create a multi-stage attack sequence, a network entity is randomly chosen among those who are exposed to the Internet. An IDS alert or a system log message is then randomly selected among those that are indicative to the compromise of that entity. A next network entity then is chosen based on the directed information graph. This process repeats until “no-action” is chosen to complete the attack. All random selections are uniform, i.e., no preference will be given to one selection versus another.

Note that a criticism of the above attack generation is that it follows the “presumed” model of attack and may be biased. This is a valid criticism, and, indeed, this procedure is not intended to replace real data set. Instead, it is intended for conducting “controlled” simulation so as to evaluate a threat assessment algorithm under specific situations. For example, we have purposely created coordinated attacks and insider threat attacks where the sequences of the attacks do not follow the attack model. This allows one to explicitly expose the flaws of a proposed model and to investigate revisions needed for it. It is our belief that this is a critical step in advancing research on high level information fusion, especially for assessing cyber threats.

III. STATISTICAL MEASURES FOR EVALUATING THREAT ASSESSMENT

Given the challenges posed by the nature of cyber attacks, this section discusses a few statistical measures that may be

used to evaluate threat assessment algorithms. Though the target application domain of these metrics is cyber security, they could be applicable to evaluating threat assessment for other multi-stage attacks.

Consider a threat assessment algorithm that assigns a threat score, $0 \leq t_i(e_j^A) \leq 1$, to the i th entity upon the occurrence of the j th event of an attack A . The threat scores may qualitatively describe the threat level of an entity as compromised if $t_i(e_j^A) = 1$, threatened if $0 < t_i(e_j^A) < 1$, and unthreatened if $t_i(e_j^A) = 0$.

A. Normalized Compromising Score

An ideal threat assessment algorithm with perfect models should generate threat scores that accurately depict the sequence of attack events. In other words, the compromised entity should have the highest threat score with respect to the other threatened entities one step before it is compromised. Based on this intuition, a *normalized compromising score*, $\hat{t}_i^*(A)$, is defined as the normalized threat score for entity i one event prior to it being compromised by attack A . That is,

$$\hat{t}_i^*(A) = \left\{ \frac{t_i(e_j^A)}{\max_{k \in C_j^A} t_k(e_j^A)} \mid t_i(e_{j+1}^A) = 1, t_i(e_j^A) < 1 \right\}, \quad (1)$$

where C_j^A indicates the set of entities that have not been compromised prior to and including event j of attack A . The average of the normalized threat scores of entities compromised by an attack A is represented as $\bar{t}^*(A)$. For a “perfect” assessment, $\bar{t}^*(A)$ would be 1, indicating that each compromised entity has the highest threat score just before it is compromised. Averaging $\bar{t}^*(A)$ for different attacks, denoted as $\bar{t}^*$, could then be indicative to the accuracy of a threat assessment algorithm. Note that, however, this metric can be misleading if many entities share the same highest threat score. For example, a naive algorithm that assigns the same threat score to all network entities in every step will have $\bar{t}^* = 1$.

Other statistical measures similar to the average normalizing compromising scores can be used. For example, one can record the percentile rank of the compromised entities one event before they are compromised. One can also examine the frequency of a top threat score entities being compromised in the next event of an attack¹. These metrics are also indicative to the performance of a threat assessment algorithm, but suffer the same problem as the average normalized compromising score described above.

B. Percent False Positives and Percent False Negatives

Because of the diverse cyber hacker motivation, behavior and skills, estimating the exact network entity a hacker

¹Readers may refer to [8] for a detailed discussion of these metrics.

targets on is unlikely. Consequently, a threat assessment algorithm for cyber attacks will likely give high threat scores to more than one entity. Let β be a minimum threshold indicating a normalized threat score high enough for an analyst to be concerned about. For an attack A , a false positive shall occur when $\max_k \hat{t}_i(e_k^A) \geq \beta$ but the entity i was never compromised. A false negative shall occur when $\hat{t}_i^*(A) < \beta$ for some entity i , meaning an entity with a low threat score becomes compromised.

Percent false positives and percent false negatives can then be formally defined as follows. Let $N_A(X)$ be the number of entities that satisfy the condition X . Let C_A represent the set of compromised entities due to an attack A . The formal definitions of percent false positives (%FP) and percent false negatives (%FN) are given in (2).

$$\begin{aligned} \%FP &= \frac{N_A((\max_k \hat{t}_i(e_k^A) \geq \beta) \text{ and } (i \notin C_A))}{|C_A|} \\ \%FN &= \frac{N_A(\hat{t}_i^*(A) < \beta)}{|C_A|} \end{aligned} \quad (2)$$

The use of %FP avoids the misleading result of a high $\bar{t}^*$ when too many network entities are categorized as highly threatened. When too many entities are above the threshold β , %FP will be high, indicating inaccurate threat predictions. Meanwhile, too few entities above β may lead to misleadingly low %FP. For example, a conservative cyber threat assessment algorithm may only predict the threatened entity when it is absolutely sure about its occurrence. This will lead to a misleadingly low %FP while the %FN will be high - many hacker actions are missed. Examining both %FP and %FN seems to be essential to determine the accuracy of a threat assessment algorithm.

Note that the use of β is not restricted to comparing against the normalized threat scores. It can also be used to compare against the percentile and other statistics. In Section IV where simulation results are discussed, β is compared against the normalized threat scores.

C. Percent Assessee Reduction

In addition to examining the accuracy of a threat assessment algorithm, it is also essential to evaluate the benefit of using an automatic tool in assisting a human analyst. Recall that a key reason for this work is to present manageable information to the network security analyst. Therefore, a metric called percent assessee reduction (%AR) is used to measure the percentage of uncompromised entities whose threat scores are below the threshold β at each step of an attack. The average value across attacks will be indicative of the information filtering provided for the analyst.

D. Percent Abnormal

We further define a metric to track the cases when the cyber attacks deviate from the a priori developed models. Consider the network entities that are compromised but have their compromising score, $t_i^*(A) = 0$. Each instance of such is referred to as an abnormality. "Percent abnormal" is the percentage of abnormalities over the set of compromised entities. Tracking this statistic allows to help determine possible root causes of high %FP and %FN.

In regards to TANDI and cyber attacks, an abnormality may occur due to the following reasons:

- Sensor readings: the abnormality could be due to a false positive or undetected event by the IDS sensors.
- Event correlator: the correlator responsible for level 2 fusion could have falsely correlated (or uncorrelated) events to an attack. A mis-correlation at this level could be due to a stealthy coordinated attack.
- Threat assessment model: Incomplete or inaccurate models may lead to unexpected hacker actions.
- Insider threat: A threat assessment algorithm is susceptible to not detecting insider threats if it assumes that attacks will originate from outside of the organization.
- Social Engineering: Social engineering attacks attempt to gain information about the network by interacting with employees and not the computer network. These attacks will therefore not be detected by network defense software or hardware.

IV. ANALYSIS VIA SIMULATION

To illustrate and to address the challenges of evaluating a threat assessment algorithm for multi-stage cyber attacks, we consider the previously proposed algorithm TANDI [8]. Note that a small set of results have been published and discussed in [8]. The focus here is to examine the issues involved in interpreting and evaluating a threat assessment algorithm, but not promoting TANDI. These discussions may also help steer the direction of designing a better threat assessment algorithm for cyber attacks.

A. Simulated Network Topologies

Two topologies with different system and network configurations were manually created for simulation. The first one (T1) is shown in Figure 1, and has four subnets, each with an external server connecting to the Internet. The four external servers are fully connected to each other; Therefore, a hacker compromising one external server can use it as a stepping stone to attack another external server. Note that the workstations in the same subnet have the same configuration and allow the same set of network users to log in, and, hence, are grouped as a single node in the logical topology representation. The second topology (T2), shown

Fig. 1. T1: a smaller network with only 2 layers of hierarchy.

in Figure 2, represents a larger network with server farms. Because, however, the servers in the server farms have the same configurations and are vulnerable to the same attacks, they are grouped as a single node with TANDI's scalable logical topology modeling approach. Note that T2 has more hierarchical levels than T1 does.

Fig. 2. T2: a larger network with more layers of hierarchy.

B. Simulated Attacks

Four sets of attack data were created based on the procedure described in Section II-C. Set 1 represents attacks that do not deviate from the a priori developed models. Sets 2 and 3 contain abnormalities, and Set 4 includes carefully crafted attacks representing coordinated attacks. Note that attack generation is topology specific. Sets 1, 2, and 3 are generated only for T1.

- *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 and ten attacks, each with 6 to 21 events, are randomly generated for T1 and T2, respectively.
- *Set 2* - A set of attacks that contain abnormalities due to various reasons, except insider threat. Ten attacks (4 to 12 events) are generated for T1.
- *Set 3* - A set of insider attacks. These attacks originate from one of the internal machines, a more extreme

version of Set 2. Each of the five generated attacks has between 3 to 6 events for T1.

- *Set 4* - A set of attacks that are fragmented from a regular attack to represent a coordinated attack for T1.

C. The Misleading Compromising Score, or Is It?

The first set of simulation results is run on the two topologies with the attack Set 1. Two different TANDI configurations with weight settings² ($w(A^*)$, $w(A(A^*))$, $w(M^*)$, $w(M(M^*))$) equals to (0.3, 0.1, 0.4, 0.15) (W1) and (0.1, 0.3, 0.15, 0.4) (W2) are tested to illustrate the insights provided by the different statistical measures.

Figure 3 shows four statistical measures, $\bar{t}^*$, %AR, %FP, and %FN. For the last three, the threshold β is set to 0.6 and compared against the normalized threat scores.

Fig. 3. Statistical measures ($\bar{t}^*$, %AR, %FP, %FN) collected by simulating TANDI on T1 and T2 with setting W1 and W2.

Observation 1: Notice first that, in terms of the average normalized compromising score ($\bar{t}^*$), shown in the first group of 4 bars in Figure 3, using W2 seems to predict more accurately than using W1 for either topology. This, however, could be misleading. Observe %AR and %FP, i.e., the 2nd and the 3rd set of 4 bars. Note that the lower %AR implies that more entities are categorized as highly threatened when using W2 than that when using W1. Moreover, the much higher %FP indicates that many entities are mis-categorized as highly threatened and are not compromised when using W2.

Observation 2: Based on Observation 1, %FP seems to be the metric for determining the performance of a threat assessment algorithm. The results in %FN for the same tests (the last set of 4 bars in Figure 3), however, suggest otherwise. Note that using W1 actually gives more false negatives, i.e., more entities are mis-categorized as less threatened, especially for T2. This conflicting observations bring forth “the” question: *Which statistical measure should a cyber threat assessment aim to optimize for?* Perhaps

²The readers may refer to [8] for the use of the weights.

more puzzling is that, will the answer depend on the network the threat assessment is running on?

Observation 3: Note the relatively small difference on $\%FN$ for T1, but the large difference for T2. With a closer examination of the two topologies, we conjecture that T2 has more “hierarchical levels” and, hence, is more sensitive to the choice of weights in the current TANDI implementation. Note that the use of W1 gives higher emphasis of already compromised machine and already used attack method, while W2 emphasizes more on the machines and attack methods that are next in line according to the models developed by SMEs and network analysts. In the case of the larger network T2, four levels of hierarchy, instead of two for T1, presents more possible courses of actions for a hacker to take. This, in turn, leads to the shortfalls of using W1, which emphasizes less on the possibility of a hacker moving toward the deeper level of the network.

Remarks: Based on the above observations, we believe that $\%FP$ and $\%AR$ are the primary metrics for evaluating a threat assessment algorithm for cyber attacks. A good cyber threat assessment algorithm should reduce information presented to a network analyst significantly so that the human analyst can focus on determining the impact of the few highly threatening attack and take necessary actions. Reducing false positives is a logical choice because high false alarm rate typically leads to deficiency, or even interrupts, of network operations.

The $\%FN$ and $\bar{t}^*$ are secondary, but still important because they offer different insights about not only the threat assessment algorithm but also the properties of the attacks. The metric $\%FN$, together with the knowledge of the threat assessment algorithm and the network, may be indicative to the hacker behavior: one knows what he wants and progress quickly deep into the network, versus one scans the network thoroughly to obtain complete information. The metric $\bar{t}^*$ provides a more detailed information about the severity of the attack actions detected, whereas the $\%FP$ alone does not distinguish between compromised entities that are most threatened vs. highly threatened. In the next section, we will discuss a specific case where $\bar{t}^*$ and $\%FN$ are used as indicators for abnormal and coordinated attacks.

D. Abnormalities, Insider Threats, and Coordinated Attacks

Any threat assessment algorithm is susceptible to flaws in the model developed by SMEs. This is especially true for cyber attacks, where the hacker profiles not only exhibit large diversity, but are also constantly changing. Attack Sets 2 and 3 are therefore created for Topology 1, to examine how the statistical measures may (or may not) exhibit the abnormalities. Based on the observations made in the previous section, the weight configuration W1 is used

for these tests since it exhibits a better $\%FP$ for T1.

Figure 4 shows five statistical measures collected when simulating the attack Sets 1, 2, and 3 on T1. In addition to the four statistics discussed in the previous section, the Percent Abnormal ($\%Abnorm$, recall Section III-D) is also presented in Figure 4.

Fig. 4. Statistical measures ($\%Abnorm$, $\bar{t}^*$, $\%AR$, $\%FP$, $\%FN$) collected when simulating TANDI with attack S1, S2, and S3 on T1.

Based on the way the abnormal attacks are created, Set 3, which contains insider attacks, is expected to have more abnormal events than Set 2 and there should be no abnormality for Set 1. This is exhibited by the $\%Abnorm$ in Figure 4. By contrast, there is little difference in terms of $\%AR$ and $\%FP$. This suggests that the two primary metrics used for assessing “normal” attacks may not be indicative of those that do not follow pre-defined 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 categorized as highly threatened.

The metrics, $\bar{t}^*$ and $\%FN$, on the other hand, exhibit trends when more abnormalities are in the attack sequence. It should not be hard to understand that $\%FN$ increases when more attack activities deviate from the prediction models. As for $\bar{t}^*$, it reflects the severity of the compromised entities, and, thus, will be affected significantly when averaged with cases where the compromising score is zero.

It seems that the metrics $\bar{t}^*$ and $\%FN$ are good indicators for abnormalities. One type of abnormality stems from coordinated attacks - seemingly different attacks actually have the same goal. This can be due to multiple hackers coordinated with each other, or several *zombie clients* being controlled by one hacker. A cyber threat assessment algorithm will likely fall short with such disguises.

Consider coordinated attacks as fragments of a single attack. These fragments, when evaluated by TANDI, would yield abnormalities - hence poor $\bar{t}^*$ and $\%FN$. If, however, the fragmented attacks were combined into a single attack, some or all of the abnormalities may be no longer abnormal, thus improve the overall $\bar{t}^*$ and $\%FN$.

Figure 5 shows statistics collected when TANDI assesses two coordinated attacks separately or as a single attack. In this simple example, the hacker uses two zombie machines to execute the attack. The first zombie client is used to footprint the network (Recon) and to transfer the password files (File Transfer) after the external computers are compromised by the second zombie client (Priv Escalation).

Fig. 5. Statistical measures ($\%Abnorm$, $\bar{t}^*$, $\%AR$, $\%FP$, $\%FN$) collected when simulating TANDI with coordinated attack S4 on T1.

Note that the metrics $\bar{t}^*$ and $\%FN$ clearly show improvement when the fragmented attacks are combined, and neither $\%AR$ nor $\%FP$ provides any indication. This example suggests that the use of $\bar{t}^*$ and $\%FN$ may help identify coordinated attacks. Note that the use of $\%Abnorm$ will not work in this example, because the detected privilege escalation presents no abnormality according to the attack model - it is not uncommon to have a hacker start with an valid user account and privilege escalation will be the first detected malicious activity. The question then, is whether the suggested approach using $\bar{t}^*$ and $\%FN$ will work in general. More specifically, will evaluating a combined attack always give higher $\bar{t}^*$ and lower $\%FN$ as compared to evaluating fragmented attacks? The answer is “no” at the moment, and requires further investigation. Nevertheless, the results demonstrated here have asserted that the use of various statistical measures helps to strengthen the capability of a cyber threat assessment algorithm.

V. CONCLUSION

Combating cyber attacks has become a top priority in many business, intelligence, and military operations. Much work has been devoted to developing more secured networks and systems, better detection mechanisms for suspicious network activities, and more accurate predictions of future attacks. Researchers in the areas of cryptography, networking, and information fusion are coming together to tighten security measures against the largely unprofiled and constantly changing cyber attacks. Performance guarantees

of these security measures means the need of formal and thorough evaluation of proposed schemes.

In the case of cyber threat assessment, evaluation is particularly challenging, yet often overlooked. The lack of publicly available databases and commonly acceptable evaluation framework and metrics is, perhaps, part of the reason why little can be found on evaluating cyber threat assessment. The work presented here aims to take the initiative, and open up avenues for the high level information fusion community to examine more carefully on proposed fusion schemes. Traditionally recognized false positives and false negatives have been examined and shown to be insufficient. It is suggested that different mixtures of statistical measures are necessary for “normal” vs. “abnormal” attacks. Much more work is needed before a cyber threat assessment can be trusted with quantitative performance guarantees!

REFERENCES

- [1] D. L. Hall and J. Llinas, “An introduction to multisensor data fusion,” in *Proceedings of the IEEE*, vol. 85, no. 1, pp. 6–23.
- [2] J. Llinas, C. Bowman, G. Rogova, A. Steinberg, E. Waltz, and F. White, “Revisions and extensions to the JDL data fusion model II,” in *Proceedings of The 7th International Conference on Information Fusion*, June 2004, pp. 1218–1230.
- [3] J. Roy, S. Paradis, and M. Allouche, “Threat evaluation for impact assessment in situation analysis systems,” in *Proceedings of SPIE, Defense and Security Symposium, Signal Processing, Sensor Fusion, and Target Recognition XI Conference*, vol. 4729, 2002.
- [4] R. Fein, B. Vossekuil, W. Pollack, R. Borum, W. Modzeleski, and M. Reddy, “Threat assessment in schools: A guide to managing threatening situations and to creating safe school climates,” *United States Secret Service and United States Department of Education: Washington, D.C.*, 2003.
- [5] S. Noel, E. Robertson, and S. Jajodia, “Correlating intrusion events and building attack scenarios through attack graph distances,” in *Proceedings of 20th Annual Computer Security Applications Conference*, December 2004.
- [6] M. Sudit, A. Stotz, and M. Holender, “Situational awareness of a coordinated cyber attack,” in *Proceedings of SPIE, Defense and Security Symposium*, March 2005, pp. 114–129.
- [7] O. Dain and R. K. Cunningham, “Fusing a heterogeneous alert stream into scenarios,” in *Proceedings of ACM Workshop on Data Mining and Security*, December 2001.
- [8] J. Holsopple, S. J. Yang, and M. Sudit, “Tandi: Threat assessment for networked data and information,” in *Proceedings of SPIE, Defense and Security Symposium*, vol. 6242, April 2006.
- [9] Netlib, “The performance database server,” <http://www.netlib.org/performance/html/PDStop.html>.
- [10] R. Gross, “Face databases,” in *Handbook of Face Recognition*, A. S. Li, Ed. New York: Springer, February 2005, accessible via <http://www.face-rec.org/databases/>.
- [11] ATIS Committee T1A1: Performance and Signal Processing, “Atis telecom glossary 2000,” February 2001, <http://www.atis.org/tg2k/tlg2k.html>.
- [12] C. Taylor, A. Krings, and J. Alves-Foss, “Risk analysis and probabilistic survivability assessment (RAPSA): An assessment approach for power substation hardening,” in *Proceedings of ACM Workshop on Scientific Aspects of Cyber Terrorism*, Washington, D.C., November 2002.