An Integrated Cyber Security Monitoring System Using Correlation-based Techniques

Qishi Wu, Denise Ferebee, Yunyue Lin, Dipankar Dasgupta
Department of Computer Science
The University of Memphis
Memphis, TN 38152–3240
Email: {qishiwu, dferebee, ylin1, ddasgupt}@memphis.edu

***Abstract*—We propose an adaptive cyber security monitoring system that integrates a number of component techniques to collect time-series situation information, perform intrusion detection, keep track of event evolution, and characterize and identify security events so corresponding defense actions can be taken in a timely and effective manner. Particularly, we employ a decision fusion algorithm with analytically proven performance guarantee for intrusion detection based on local votes from distributed sensors. Different from the traditional rule-based pattern matching technique, security events in the proposed system are represented in a graphical form of correlation networks using random matrix theory and identified through the computation of network similarity measurement. Extensive simulation results on event identification illustrate the efficacy of the proposed system.**

***Index Terms*—Cyber security, decision fusion, event correlation, random matrix theory**

I. Introduction

The successful executions of many commercial, scientific, and military applications require timely, reliable, and accurate information flow in cyber space to support online transactions and remote operations. Developing effective security monitoring mechanisms to provide cyber situation awareness has become an increasingly important focus within the network research and management community. However, providing complete cyber situation awareness based on low-level information abstracted from raw sensor data is extremely challenging primarily because (i) situation information is typically incomplete and imperfect, (ii) security events are constantly evolving over time, space, scale, and function, and (iii) the number and type of cyber attacks are practically immeasurable.

The main objective of our work is to develop a cyber security monitoring (CSM) system that integrates a number of component techniques to collect time-series situation information, perform intrusion detection, keep track of event evolution, and characterize and identify security events so corresponding defense actions can be taken in a timely and effective manner. In particular, we design an intrusion detection component based on a hard fusion algorithm with analytically proven performance guarantee in terms of high hit rate and low false alarm rate without requiring *a priori* knowledge on the probability of intrusion occurrence. We explore the correlations among a set of carefully selected event indicators to characterize and identify different types of security events based on random matrix theory (RMT) and graph matching techniques. Different from the traditional rule-based pattern matching technique, security events are represented in a graphical form of correlation networks and identified through the computation of network similarity measurement to eliminate the need for constructing rule-based user or system profiles. The proposed CSM system attempts to facilitate a better understanding of human analysts' cognitive needs and bridge the gap between the analysts' mental model and the lower level information model. We also conduct extensive experiments on simulation datasets to illustrate the efficacy of the technical approaches in the proposed CSM system.

The rest of the paper is organized as follows. Section II describes the related work. Section III presents the framework of the proposed cyber security monitoring system. The technical details of each component of the proposed CSM system are given in Section IV. The experimental results for performance evaluation are provided in Section V. We conclude our work in Section VI.

II. Related work

The technology of CSM is based on observation, experience, and classification of attacks, vulnerabilities, and countermeasures [1]. There exist a large number of commercial and government off-the-shelf tools and a significant amount of research and development efforts in CSM. A detection method falls into one of two categories using either statistical deviation or pattern matching [2]. The proposed CSM system models security events in a graphical form of correlation networks and applies graph matching techniques for event identification.

Many existing non-model based or model based fusion methodologies are derived from some variants of decision rules such as Voting, Bayes Criterion, Maximum a Posterior Criterion (MAP), and Neyman-Pearson [3], [4]. Data fusion is in general categorized as low-, intermediate-, or high-level fusion, depending on the stage where actual fusion processing takes place. The fusion algorithm we apply to intrusion detection is a model based high-level hard fusion scheme, where a final global decision is reached by integrating local binary decisions made by multiple sensors that detect the same intrusion from different locations.

RMT was initially proposed by Wigner and Dyson in the 1960s for studying the spectrum of complex nuclei [5] and

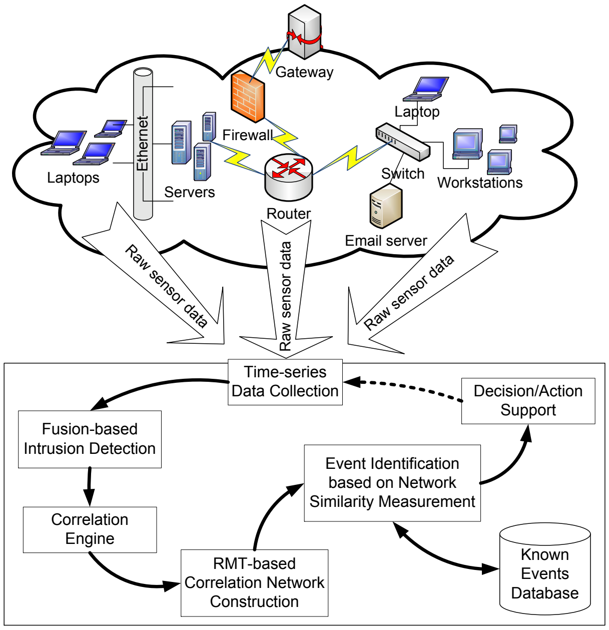


Fig. 1. Framework of an integrated adaptive cyber security monitoring system.

is a powerful approach for identifying and modeling phase transitions associated with disorder and noise in statistical physics and materials science. RMT has been successfully applied to the study of behaviors of complex systems, but its applicability in cyber security remains largely unexplored.

Network characterization and comparison have been studied in various domains, especially biological systems. Most studies of biological networks compare their connectivity properties to theoretical or other types of well-studied graphical systems [6], [7]. There exist a number of approaches to the comparison of biological networks with focus on either the general topological statistics of subgraphs [8] or the statistical prevalence of different types of node connection patterns [9]. The network comparison procedure in [10] is based on the shared-edge ratio. In this paper, we conduct a comparative analysis of security correlation networks to identify security events using the graphical form of situation information data.

III. Monitoring Network Environments

We propose an integrated adaptive cyber security monitoring system to provide cyber situation awareness. The framework of the proposed system is illustrated in Fig. 1.

We use sensors that are distributed in both networks and systems to collect time-series measurements of various event indicators. Each sensor makes a local threshold-based binary decision on the occurrence of an intrusion or security event and sends its decision together with the raw event indicator measurements to a front-end data center. Based on the local votes, the intrusion detector makes a global intrusion detection decision using a hard sensor fusion algorithm. When an alarm signal is raised, the correlation engine is invoked to construct an event indicator correlation matrix from time-series raw situation measurements collected by sensors up to the current time step, which is then processed by the RMT-based component to construct a correlation network of event indicators. Note that the inherent nature of a certain security event is captured in its correlation network that establishes the true relationships between all pairs of event indicators. The graphical representation of the current security event is then compared to those of known events stored in a database to identify the event type based on network similarity measured by graph matching techniques. We consider the following two cases in the process of event identification:

1) If an event is successfully identified with a high matching score, the corresponding defense mechanism is launched to assess and address vulnerability and defeat or mitigate the attack. The identification results are also fed back to the system to redeploy sensors or refine the design of event indicators.
2) If there is no good match to any existing event in the database, the correlation network representing the current security event is added to the known event database for future reference.

This security monitoring process is executed at a certain time interval in an adaptive manner. Senor data is accumulated at more time steps as the event evolves, resulting in more robust and cognitive network representations and therefore more accurate event detection and identification. The system adaptively determines the duration as well as the amount of raw data that has to be collected and processed.

IV. Proposed Approaches

A. *Intrusion Detection*

The intrusion detector in the proposed CSM system uses a hard fusion algorithm with analytically proven performance guarantee to make a prompt and reliable decision on the occurrence of an intrusion from a global perspective based on local votes casted by individual sensors [11]. We consider a non-perfect sensor model, which has a hit rate p_{h_i} and a false alarm rate $p_{f_i}, i = 1, 2, \ldots, N$. Sensor i makes an independent binary decision S_i as either 0 or 1. The intrusion detector uses a simple $0/1$ counting rule to collect local decisions and compute S as: $S = \sum_{i=1}^{N} S_i$, which is then compared with a system threshold T to make a final decision. For simplicity, we neglect covariance and assume that sensor measurements are conditionally independent under the hypothesis of an intrusion occurrence. The mean and variance of S are given below under hypothesis H_1 when an intrusion is present:

$$E(S|H_1) = \sum_{i=1}^{N} p_{h_i}, \quad Var(S|H_1) = \sum_{i=1}^{N} p_{h_i}(1 - p_{h_i}). \tag{1}$$

Similarly, the mean and variance of S under hypothesis H_0 when there is no intrusion are defined as:

$$E(S|H_0) = \sum_{i=1}^{N} p_{f_i}, \quad Var(S|H_0) = \sum_{i=1}^{N} p_{f_i}(1 - p_{f_i}). \tag{2}$$

Obviously, the threshold value T is critical to the system detection performance. It is reasonable to provide value bounds for T as $\sum_{\mathrm{i=1}}^{N} p_{f_i} < T < \sum_{\mathrm{i=1}}^{N} p_{h_i}$. Let P_h and P_f denote the hit rate and false alarm rate of the fused system, respectively:

$$\begin{aligned} P_h &= P\{S \geq \mathrm{T}|\mathrm{H}_1\}, \\ P_f &= P\{S \geq \mathrm{T}|\mathrm{H}_0\} = 1 - P\{S < \mathrm{T}|\mathrm{H}_0\}. \end{aligned} \tag{3}$$

We wish to achieve better system detection performance than the weighted averages in terms of higher hit rate and lower false alarm rate which are defined as:

$$\sum_{i=1}^{N} \frac{p_{h_i}}{\sum_{j=1}^{N} p_{h_j}} p_{h_i} = \frac{\sum_{i=1}^{N} p_{h_i}^2}{\sum_{i=1}^{N} p_{h_i}}, \tag{4}$$

$$\sum_{i=1}^{N} \frac{1 - p_{f_i}}{\sum_{j=1}^{N} (1 - p_{f_j})} p_{f_i} = \frac{\sum_{i=1}^{N} (1 - p_{f_i}) p_{f_i}}{\sum_{i=1}^{N} (1 - p_{f_i})}. \tag{5}$$

Thus, the following inequalities should hold:

$$P_h > \frac{\sum_{i=1}^{N} p_{h_i}^2}{\sum_{i=1}^{N} p_{h_i}}, \quad P_f < \frac{\sum_{i=1}^{N} (1-p_{f_i}) p_{f_i}}{\sum_{i=1}^{N} (1-p_{f_i})}. \tag{6}$$

To determine the lower bound on the hit rate of the fused detection system, we have the following:

$$\begin{aligned} P_h &\geq P\{|S - \sum_{\mathrm{i=1}}^{N} p_{h_i}| \leq (\sum_{\mathrm{i=1}}^{N} p_{h_i} - \mathrm{T})|\mathrm{H}_1\} \\ &\geq 1 - \frac{\sigma^2}{\mathrm{k}^2} = 1 - \frac{\sum_{i=1}^{N} p_{h_i}(1-p_{h_i})}{(\sum_{i=1}^{N} p_{h_i} - T)^2}, \end{aligned} \tag{7}$$

where we apply Chebyshev's inequality in the second step as illustrated in Fig. 2 and denote $(\sum_{1}^{\mathrm{N}} p_{h_i} - \mathrm{T})$ by k. Now the inequality of P_h in Eq. 6 can be ensured by the following sufficient condition:

$$1 - \frac{\sum_{i=1}^{N} p_{h_i}(1 - p_{h_i})}{(\sum_{i=1}^{N} p_{h_i} - T)^2} \geq \frac{\sum_{i=1}^{N} p_{h_i}^2}{\sum_{i=1}^{N} p_{h_i}}. \tag{8}$$

Following that, an upper bound on T can be derived from Eq. 8 as follows:

$$T \leq \sum_{i=1}^{N} p_{h_i} - \sqrt{\sum_{i=1}^{N} p_{h_i}}. \tag{9}$$

The upper bound on the false alarm rate of the fused detection system can be derived in a similar way. The final range of T that provides fusion performance guarantee on both hit rate and false alarm rate as follows:

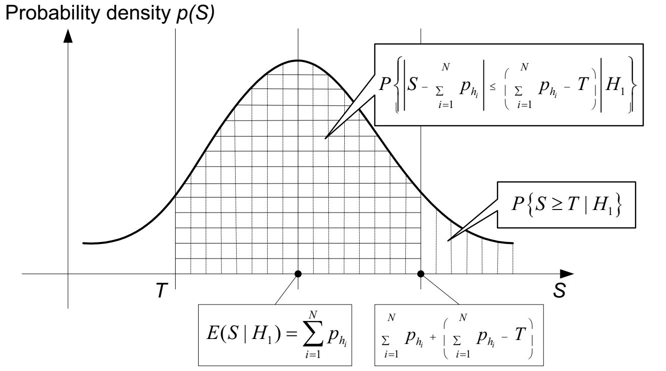


Fig. 2. Application of Chebyshevs inequality to obtain performance bounds.

$$\left[\sum_{i=1}^{N} p_{f_i} + \sqrt{\sum_{i=1}^{N} (1-p_{f_i})}, \sum_{i=1}^{N} p_{h_i} - \sqrt{\sum_{i=1}^{N} p_{h_i}}\right]. \tag{10}$$

B. Correlation Engine

We design a correlation engine based on the Pearson's correlation coefficient where the input table containing time-series event indicator measurements is transformed into a correlation matrix with each element calculated as:

$$\rho = \frac{SP}{\sqrt{SS_x SS_y}}, \tag{11}$$

where $SP = \sum XY - \frac{\sum X \sum Y}{n}$, $SS_x = \sum X^2 - \frac{(\sum X)^2}{n}$, $SS_y = \sum Y^2 - \frac{(\sum Y)^2}{n}$, n is the number of time steps, x and y are a pair of event indicators, and X and Y are the time-series measurements (vectors) of event indicators x and y, respectively. The correlation matrix establishes the relationship between each pair of event indicators under the cyber situation up to the most recent time step. Since a security event is constantly evolving, the number of time steps sampled so far may not be sufficient to cover the entire period of the event, resulting in incomplete measurement data. Furthermore, the measurement data is generally imperfect due to the inappropriateness of event indicator selection, inaccurate measurements, and delay effects. Therefore, the correlation matrix contains noise or random components that must be filtered out to reflect the true correlations among event indicators under the current cyber situation.

C. Correlation Network Construction

The lack of comprehensive and accurate system knowledge makes it hard to determine an appropriate threshold to differentiate true correlation from random noise. Random matrix theory (RMT), which has been widely and successfully used in physics, is a powerful approach to distinguish system-specific, non-random properties embedded in complex systems from random noise. There are many different classical random

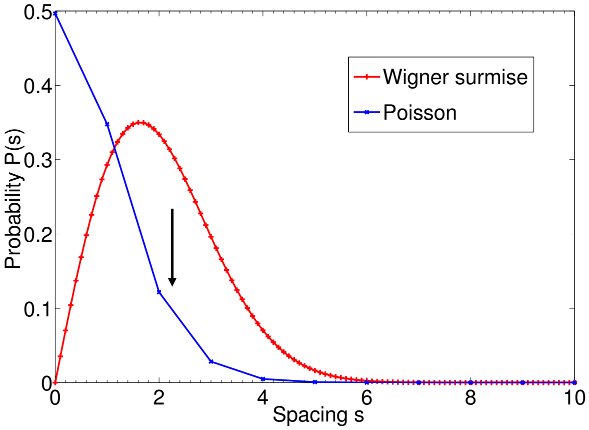


Fig. 3. Transition from GOE distribution to Poisson distribution in random matrix theory.

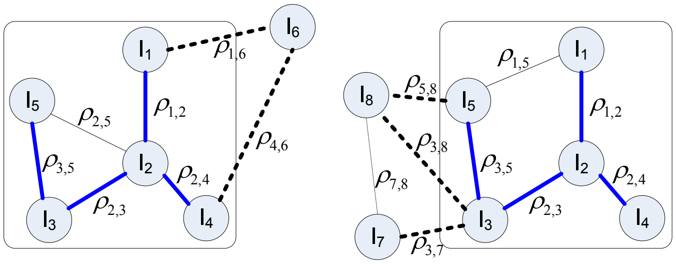
matrix ensembles including Gaussian, Jacobi, circular, and Frontier [12].

We hypothesize that the universal properties of RMT are also applicable to the sensor data in cyber space and the correlation threshold can be determined by characterizing the correlation matrix of network profiles using RMT. We develop an approach based on RMT to denoise the correlation matrix by considering the two main properties in reference to symmetric matrices [12]:

1) if a correlation between nearest-neighbor eigenvalues exist, the nearest neighbor spacing distribution (NNSD) of eigenvalues follows Wigner surmise of Gaussian Orthogonal Ensembles (GOE);
2) if there is no such correlation, the NNSD conforms to a Poisson distribution.

As illustrated in Fig. 3, the transition between these two distributions can potentially serve as a reference point and be used as a threshold to automatically construct an event indicator correlation network. The nodes in an event indicator correlation network represent event indicators and the edges represent correlations between all pairs of event indicators with weights equal to correlation coefficients. Once the threshold is determined, a correlation network is constructed from the original correlation matrix by keeping those edges with weights or correlation coefficients higher than the threshold and eliminating all others below the threshold. Such a correlation network is the graphical representation of a security event under the current cyber situation.

The detailed RMT procedure to determine the threshold is similar to the one used in [13]. For a given Pearson correlation matrix, we construct a series of new correlation matrices using different cutoff values. Any element in the original correlation matrix that has an absolute value less than the selected cutoff is set to 0 in the new matrices. We calculate the eigenvalues of each correlation matrix using direct diagonalization of the matrix. Standard spectral unfolding techniques are applied to have a constant density of eigenvalues and subsequently the nearest neighbor spacing distribution, which is employed to describe the fluctuation of eigenvalues of the correlation matrix. We use χ^2 test to determine two critical threshold values that define the transition range from GOE to the Poisson distribution at a certain confidence level, and the value at which the reference point starts to follow the Poisson distribution will be used as the threshold or pruning value.



(a) Current correlation network

(b) Known correlation network

Fig. 4. Similarity measurement between (a) current and (b) known correlation networks.

D. Event Identification

Event identification compares the current correlation network to those stored in the known event database and finds the closest one as a winner. Obviously, a good network similarity measurement technique is crucial to the overall performance of the proposed CSM system.

As shown in Fig. 4, given a pair of current and known correlation networks: $G^c = (V^c, E^c)$ and $G^k = (V^k, E^k)$ for comparison, we first identify the shared subgraphs that contain the same set V_{shared} of event indicator nodes. The set of non-shared nodes is denoted as $V^c_{non-shared}$ in the current network and $V^k_{non-shared}$ in the known network. We have $V = V_{shared} + V_{non-shared}$ in both networks. Note that there may not exist a one-to-one node correspondence in these two networks if the number of event indicators changes, considering the adaptive nature of the system. Based on the shared subgraphs, we characterize each network by dividing the set E of edges into four subsets:

1) E_{SI}: this shared internal subset contains edges that are shared in both networks and connect pairs of nodes in V_{shared}, such as edges $e_{1,2}$, $e_{2,3}$, $e_{2,4}$, and $e_{3,5}$ in both networks;
2) E_{NI}: this non-shared internal subset contains edges that are not shared but connect pairs of nodes in V_{shared}, such as edge $e_{2,5}$ in the current network G^c and edge $e_{1,5}$ in the known network G^k;
3) E_{BR}: this bridging subset contains edges that connect pairs of nodes from V_{shared} to $V_{non-shared}$, such as edges $e_{1,6}$ and $e_{4,6}$ in the current network G^c and edges $e_{5,8}$, $e_{3,8}$, and $e_{3,7}$ in the known network G^k;
4) E_{EX}: this external subset contains edges that connect pairs of nodes in $V_{non-shared}$, such as edge $e_{7,8}$ in the known network G^k.

We have $E = E_{SI} + E_{NI} + E_{BR} + E_{EX}$ in both networks and $E^c_{SI} = E^k_{SI}$. The similarity s between two networks is determined by the following measurement:

$$\begin{aligned} s \quad &= \omega_{SI} \cdot \sum_{e \in E_{SI}} \left(1 - \left|\rho(e^c) - \rho(e^k)\right|\right) \\ &+ \omega_{BR} \cdot \left(\frac{|E^c_{BR}|}{|E^c|} \sum_{e \in E^c_{BR}} \rho(e) + \frac{|E^k_{BR}|}{|E^k|} \sum_{e \in E^k_{BR}} \rho(e) \right) \\ &- \omega_{NI} \cdot \left(\frac{|E^c_{NI}|}{|E^c|} \sum_{e \in E^c_{NI}} \rho(e) + \frac{|E^k_{NI}|}{|E^k|} \sum_{e \in E^k_{NI}} \rho(e) \right), \end{aligned} \tag{12}$$

where ω_{SI}, ω_{BR}, and ω_{NI} are weight coefficients for three different subsets of edges, and $|E|$ represents the number of edges in E.

The first term on the right side of Eq. 12 depicts the similarity between two overlapped subgraphs. The second term depicts the relations between the shared and non-shared nodes, which are considered as a positive factor because high correlations with other nodes indicate the significance of the shared node. The third term depicts the relations between the shared nodes that are only identified in one network. We consider the last term as a negative factor because a true correlation is expected to be correctly captured for the same event type if these two indicators exist in both networks. Such unmatched correlations in the shared subgraphs could be caused by noise or inaccurate measurements. Since the edges in the subsets E^c_{EX} and E^k_{EX} do not have any shared nodes, the correlation information carried in these subsets should not affect the shared subgraph similarity as much as other subsets. Therefore, we do not consider the external edge subsets E^c_{EX} and E^k_{EX} in our similarity measurement.

In practice, the weight coefficients ω_{SI}, ω_{BR}, and ω_{NI} can be determined based on empirical study. The guideline for choosing appropriate values for them is that the first term should be a dominating factor in similarity measurement compared to the last two terms. Obviously, a high similarity measurement value indicates a good match between the current and known event types.

V. Performance Evaluation

A. Experiment Settings

We implement these technical components and integrate them into a proof-of-concept system for cyber security monitoring. This system is tested on a large number of simulation datasets to evaluate the performance of event identification.

We first build a database that stores 100 different types of known security events based on 12 carefully selected event indicators as follows:

1) For each known security event type, we define a characteristic correlation profile by specifying a different set of correlations between indicators.
2) Based on the profile, we generate simulated time-series raw sensor data of all event indicators for 100 time steps, which results in a 12×100 raw data table. A small percentage of the measurement is added to every data point specified by a pair of (event indicator, time step) to simulate randomness caused by measurement noise, environment dynamics, and delay effects.

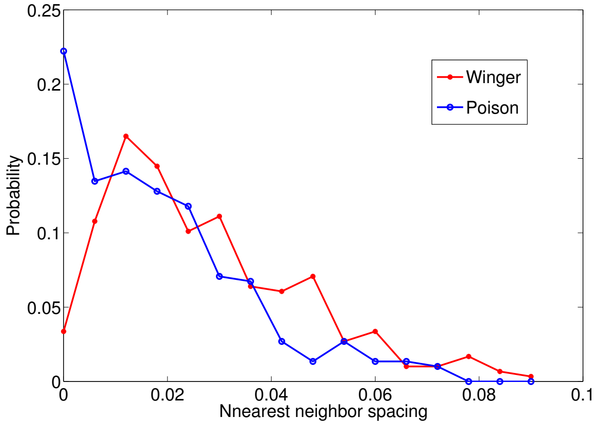


Fig. 5. Transition from Wigner surmise to Poisson distribution in RMT.

3) The data table is passed as input to the CSM system, which computes the Pearson correlation matrix and uses RMT to find an appropriate cutoff value for constructing a correlation network.

We generate a correlation network for each of 100 known security event type and store these correlation networks in a database of known event types. This database will be used later for comparison with testing datasets.

B. RMT-based Correlation Network Construction

We implement RMT technique to select appropriate cutoff values to construct correlation networks from correlation matrices. Fig. 5 illustrates a transition from Wigner surmise to Poisson distribution of the nearest neighbor spacing of eigenvalues computed from a correlation matrix with total 300 indicators. The transition occurs within a range of cutoff values $[0.96, 0.985]$ and the maximum value is selected for correlation network construction. This clear transition between two different distributions justifies the validity of our RMT technique in removing system- and measurement-related noise in the sensor data.

C. System Testing and Performance Measurements

We conduct two sets of experiments to study the effects of the number of event indicators and the number of time steps on the event identification performance, respectively. These performance measurements provide us with valuable insight into how the raw data collection process should be constructed and how the CSM system would respond at various time points with different numbers of event indicators.

1) Effect of the number of event indicators: By applying the predefined event profiles for building the event database, we create 100 testing events with a different number of event indicators (ranging from 3 to 12) based on the raw data collected during the first 40, 60, 80, and 100 time steps, respectively. We construct a correlation network for each testing event using the same procedure as for known events and perform network similarity comparison with all known events in the database. A testing event is labeled as “correctly identified” if the corresponding event in the database with the same event profile is selected as the winner (with the highest score) based on the similarity measure defined in Eq. 12; otherwise, it is considered as an incorrect identification. The identification performance in response to the number of event indicators for various time steps is plotted in Fig. 6. We notice that only about 10% of the testing events are successfully identified when 3 event indicators are used in data collection. As the number of event indicators increases, which means that a more comprehensive measure of the event’s impact is considered at each time step, we observe an obvious increasing trend in event identification performance. When the number of event indicators is close to that we used for building the database, the identification rate is approaching 100%, which demonstrates the effectiveness of our approach.

2) Effect of the number of time steps: Similarly, by applying the predefined event profiles for building the event database, we create 100 testing events with 4, ,6, 8, and 10 event indicators, respectively, based on the raw data collected for a different number of time steps (ranging from 10 to 100). We construct a correlation network for each testing event and perform network similarity comparison with all known events in the database. The event identification performance using different numbers of event indicators in response to the number of time steps is plotted in Fig. 7. We notice that a small portion of the testing events are successfully identified based on the raw sensor data we collect during the first 10 time steps. As the number of time steps increases, which means that more temporal contextual information is gathered about the current event, we observe an obvious increasing trend in event identification performance. When the number of time steps reaches 80, the identification rate using 8 or 10 event indicators is approaching 100%. These performance curves strongly indicates that even with a subset of event indicators that are used to build the database, after reaching a certain time point, the collected information would be sufficient to correctly identify the current event.

VI. Conclusion

We investigated an adaptive cyber security monitoring system that integrates a number of component techniques including intrusion detection based on decision fusion, correlation computation of event indicators, network representation of security events based on RMT, and event identification based on graph matching and network similarity measurement, in a unified framework. The simulation results show that the proposed system exhibits promising performance in security monitoring and event identification.

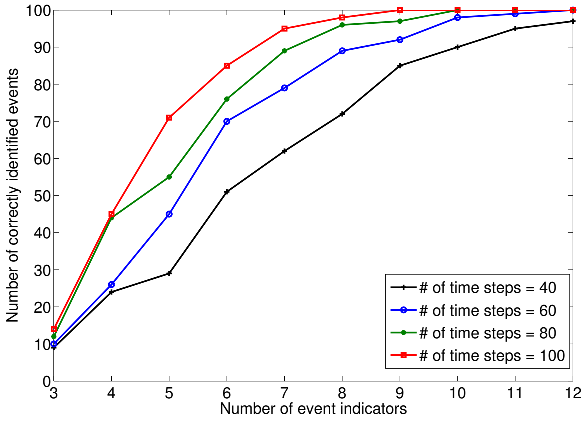


Fig. 6. Event identification performance vs. the number of event indicators in data collection.

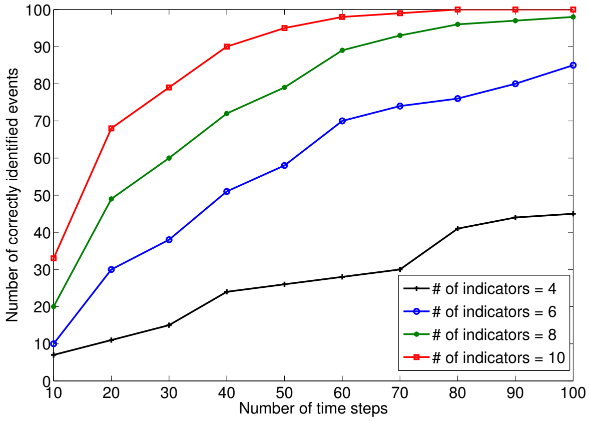


Fig. 7. Event identification performance vs. the number of time steps in data collection.

References

[1] L. Lapadula, “State of the art in cybersecurity monitoring,” 1999. http://www.mitre.org.

[2] L. Lapadula, “A compendium of commercial and government tools and government research projects,” 2000. http://www.mitre.org.

[3] R. R. Tenney and N. R. Sandell, “Detection with distributed sensors,” *IEEE Trans. Aerosp. Electron. Syst.*, vol. AES, pp. 501–510, July 1981.

[4] F. Sadjadi, “Hypothesis testing in a distributed environment,” *IEEE Trans. Aerosp. Electron. Syst.*, vol. AES, pp. 134–137, March 1986.

[5] E. Wigner, “Random matrices in physics,” *SIAM Review*, vol. 9, pp. 1–23, 1967.

[6] A. Barabasi and R. Albert, “Emergence of scaling in random networks,” *Science*, vol. 286, p. 509512, 1999.

[7] R. Albert, H. Jeong, and A. Barabasi, “Error and attack tolerance of complex networks,” *Nature*, vol. 406, p. 378382, 2000.

[8] H. Yu, X. Zhu, D. Greenbaum, J. Karro, and M. Gerstein, “Topnet: a tool for comparing biological sub-networks, correlating protein properties with topological statistics,” *Nucl. Acids Res*, vol. 32, p. 328337, 2004.

[9] R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, and I. Ayzenshtat, “Superfamilies of evolved and designed networks,” *Science*, vol. 303, p. 15381542, 2004.

[10] Pairing subgrapHs Using NetworK Environment Equivalence. http://www.sbg.bio.ic.ac.uk/ phunkee/.

[11] M. Zhu, R. Brooks, Q. Wu, N. Rao, S. Ding, and S. Iyengar, “Fusion of threshold rules for target detection in self-organizing sensor networks,” in *Proc. of the 9th ONR/GTRI Workshop on Target Tracking and Sensor Fusion*, (Gatlinburg, TN), June 22-23 2006.

[12] A. Edelman and N. Rao, “Random matrix theory,” *Acta Numerica*, vol. 14, pp. 233–297, 2005.

[13] F. Luo, Y. Yang, J. Zhong, H. Gao, L. Khan, D. Thompson, and J. Zhou, “Constructing gene co-expression networks and predicting functions of unknown genes by random matrix theory,” *BMC Bioinformatics*, vol. 8, pp. 299+, August 2007.