Monitoring Security Events Using Integrated Correlation-based Techniques

Qishi Wu, Denise Ferebee, Yunyue Lin, Dipankar Dasgupta
University of Memphis
209 Dunn Hall
Memphis, TN 38152-3240
{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, and characterize and identify security events so corresponding defense actions can be taken in a timely and effective manner. We employ a decision fusion algorithm with analytically proven performance guarantee for intrusion detection based on local votes from distributed sensors. The security events in the proposed system are represented as forms 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.

General Terms
Security, Performance, Theory

Keywords
Cyber security, event correlation, random matrix theory

1. 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, and characterize and identify security events so corresponding defense actions can be taken in a timely and effective manner. We design an intrusion detection component based on a hard fusion algorithm with analytically proven performance guarantee. We explore the correlations among a set of event indicators to characterize security events based on random matrix theory (RMT) and identify the event type using graph matching techniques. Different from the traditional rule-based pattern matching technique, security events in the proposed system 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, which often involves subjective human judgement and interpretation. The proposed CSM system attempts to facilitate a better understanding of human analysts' cognitive needs and bridge the gap between human 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.

2. RELATED WORK
There exist a large number of commercial and government off-the-shelf tools and a significant amount of research and development efforts in cyber security monitoring, most of which are focused on intrusion detection. A detection method falls into one of two categories using either statistical deviation or pattern matching [4]. The fusion algorithm we apply to intrusion detection is a model based high-level hard fusion scheme, where a final global decision is made by integrating local binary decisions made by multiple sensors detecting the same intrusion from different locations.

RMT was initially proposed by Wigner and Dyson in the 1960s for studying the spectrum of complex nuclei [6] and 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 [2]. The network comparison procedure in [1] is
based on the shared-edge ratio. We conduct a comparative analysis of security correlation networks to identify security events using graphical forms of situation information data.

3. MONITORING NETWORK

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 frontend 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. The graphical representation of the current event is then compared to the known events stored in a database to identify the event type based on network similarity measured by graph matching techniques. This security monitoring process is executed at a certain time interval in an adaptive manner. In the case of a security event occurrence, sensor data are 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 also adaptively determines the duration as well as the amount of raw data that must be collected and processed.

4. PROPOSED APPROACHES

4.1 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 [7]. We consider a non-perfect sensor model, which has a hit rate $p_h$, and a false alarm rate $p_f$, $i = 1, 2, \ldots, N$. 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:

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

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

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

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

$$P_h = P\{S \geq T|H_1\}, \quad P_f = P\{S \geq T|H_0\} = 1 - P\{S < T|H_0\}. \quad (3)$$

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:

$$\sum_{i=1}^{N} p_{h_i} \frac{p_{h_i}}{\sum_{i=1}^{N} p_{h_i}} = \sum_{i=1}^{N} p_{h_i} \frac{p_{f_i}}{\sum_{i=1}^{N} p_{h_i}}, \quad (4)$$

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

Thus, the following inequalities should hold:

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

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

$$P_h \geq P\{S - \sum_{i=1}^{N} p_{h_i}| \leq \left( \sum_{i=1}^{N} p_{h_i} - T \right)|H_1\} \geq 1 - \frac{\sigma^2}{2\pi} \left( \frac{1}{(\sum_{i=1}^{N} p_{h_i} - T)^2} \right), \quad (7)$$

where we apply Chebychev’s inequality in the second step and denote $(\sum_{i=1}^{N} p_{h_i} - T)$ by $k$. If we use the inequality in Eq. 7 to ensure the inequality of $P_h$ in Eq. 6, an upper bound on $T$ can be derived as follows: $T \leq \sum_{i=1}^{N} p_{h_i} - \sqrt{\sum_{i=1}^{N} p_{h_i}}$. 

![Figure 1: Framework of an integrated adaptive cyber security monitoring system.](image-url)
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:

$$
\sum_{i=1}^{N} x_{pi} + \sum_{i=1}^{N} (1-p_{i}) \cdot \sum_{i=1}^{N} p_{i} - \sum_{i=1}^{N} p_{i} = (8)
$$

4.2 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: $\rho = \frac{SP}{\sqrt{SS_x SS_y}}$, where $SP = \sum XY - \frac{1}{N} \sum X \sum Y$, $SS_x = \sum X^2 - \frac{1}{N} (\sum X)^2$, $SS_y = \sum Y^2 - \frac{1}{N} (\sum Y)^2$, $n$ 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. The correlation matrix contains noise or random components, which must be filtered out to reflect the true correlations among event indicators under the current cyber situation.

4.3 Correlation Network Construction

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. 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 [3]:

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.

The transition between these two distributions can potentially serve as a reference point and be used as a threshold to construct an event indicator correlation network.

The detailed RMT procedure to determine the threshold is similar to the one used in [5]. 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 zero 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 $\chi^2$ 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.

$$
\rho = \frac{SP}{\sqrt{SS_x SS_y}}
$$

4.4 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. As shown in Fig. 5, 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 $V_{non-shared}$ in the current network and $V_{non-shared}$ in the known network. We have $V = V_{shared} + V_{non-shared}$ in both networks. 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 $e_{1.2}$ 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 $e_{1.2}$ 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 $e_{3.8}$, $e_{3.9}$, 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 $E_{SI} = E_{SI}^k$. The similarity $s$ between two networks is determined by the following measurement:

$$
s = \omega_{SI} \cdot \sum_{\epsilon \in E_{SI}} (1 - |\rho(e_c) - \rho(e_k)|) + \omega_{BR} \cdot \sum_{\epsilon \in E_{BR}} |\rho(e_c)| + \sum_{\epsilon \in E_{BR}} |\rho(e_k)| + \sum_{\epsilon \in E_{NI}} \rho(e_c) + \sum_{\epsilon \in E_{NI}} \rho(e_k) - \omega_{NI} \cdot \sum_{\epsilon \in E_{NI}} |\rho(e_c)| - \sum_{\epsilon \in E_{NI}} \rho(e_c) - \sum_{\epsilon \in E_{NI}} \rho(e_k)
$$
to construct correlation networks from correlation matrices.

5. PERFORMANCE EVALUATION

These technical components are implemented and integrated into a proof-of-concept system. We build a database of correlation networks for 100 different types of known security events based on 12 carefully selected event indicators, which are used for comparison with testing datasets.

5.1 Correlation Network Construction

RMT technique is used to select appropriate cutoff values to construct correlation networks from correlation matrices. Fig. 2 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.

5.2 Performance Measurements

5.2.1 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. The identification performance in response to the number of event indicators increases with the number of event indicators, 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%.

5.2.2 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. 4. 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%.

6. CONCLUSION

We proposed 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, RMT-based network representation of security events, and event identification based on graph matching and network similarity measurement. The simulation results show that the proposed system exhibits promising performance in security monitoring and event identification.

7. REFERENCES