International Journal of Security and Its Applications
Vol. 10, No. 2 (2016), pp.325-340
http://dx.doi.org/10.14257/ijsia.2016.10.2.29

An Improved Event Scenario Correlation Method for Multi-Source Security Log

Qianyun Wang[1*], Shuyu Chen[2], Hancui Zhang[3], and Tianshu Wu[4]

[1]*College of Software Engineering, Chongqing University, Chongqing, China*
[2]*College of Software Engineering, Chongqing University, Chongqing, China*
[3]*College of Software Engineering, Chongqing University, Chongqing, China*
[4]*College of Computer Science, Chongqing University, Chongqing, China*
[1]*wqy0932@foxmail.com,* [2]*netmobilab@cqu.edu.cn,* [3]*zhc_813@126.com,*
[4]*wutianshu@cqu.edu.cn*

Abstract

Developing computer technologies and a network of persistently growing size put massive hosts and transmission devices in a vast network at increasingly higher risks. Log information of various devices can facilitate the detection of intrusion and attacks. Log information from a single data source is, however, with limitations. The analysis results cannot precisely reflect the current network situation if log information in a single data source is analyzed without correlation to analysis of log information from different data sources. To better demonstrate network situation, this paper proposes an improved event scenario correlation method for multi-source log analysis via researching on numerous existing data fusion methods and event correlation methods as well as integration of conventional event scenario correlation (ESC) method with fuzzy reasoning. Experimental results prove that the proposed method significantly reduces the False Positive rate (FP rate) and False Negative rate (FN rate) of security logs.

Keywords*: security log information, data fusion, event correlation, fuzzy reasoning*

1. Introduction

Analysis of log information on an individual device can no longer reflect actual network security status, considering such a complex network and various cyberattacks. It is imperative to correlate and fuse relatively independent log information on multiple devices in order to fully understand the situation of the current network. Therefore, NSSA (Network Security Situational Awareness) **Error! Reference source not found.** can be better realized.

Different data structures of those multi-source logs, however, make the correlation and fusion processes difficult. In the process of multi-source security log analysis, there are mainly the following issues:

(1) Different data structures of the multi-source-log information: The security logs used for multi-source log analysis, are different not only in data formats, but also in the way of storage. Some of the log information can be easily understood, but others are not. Therefore, understand all the log field clearly to make a comprehensive analysis of these heterogeneous logs is the points for multi-source log analysis.

(2) Large amount of log data for multi-source-log analysis: Logs in each single sensor can be accumulated into a large amount of information over time. When logs from the various sensors gather, the total amount can be tremendous. So, how to extract and process useful logs effectively is also a major issue.

(3) Data alignment: Content of logs from different sensors is produced according to the parameters of the equipment settings, which means, to make a comprehensive analysis of

ISSN: 1738-9976 IJSIA

them is required to put multi-source logs in the same parameter system. Then reducing and merging of the attribute as well as data alignment are the points for multi-source log analysis.

(4) Contradictory information: Functions of sensors vary, so as measured on the same network environment, false information generated by noise may result in conclusions from different sensors in stark contrast to each other. Therefore, how to deal with the false information and fuse the right information effectively is an issue to be solved.

This paper, accordingly, proposes the idea and theoretical research on the application of fuzzy reasoning clustering in log analysis. In this paper, event scenario correlation method combined with fuzzy reasoning is defined together with its algorithm description.

This paper compares the improved event scenario correlation method to both single-log alarm correlation method and conventional event scenario correlation (ESC) method. As the experimental results reveal, event scenario correlation method combined with fuzzy reasoning significantly reduces FN rate and FP rate.

The remainder of this paper is organized as follows. Section 2 introduces related work. Section 3 illustrates the preliminaries. Section 4 proposes the mathematic model of multi-source security log fusion based on fuzzy reasoning. Section 5 presents the improved event scenario correlation method (event scenario correlation method combined with fuzzy reasoning) in detail. Section 6 conducts experiments and presents analyses. Finally, Section 7 gives conclusions and looks into future work.

2. Related Work

This section summarizes research work related to data fusion and event correlation.

2.1 Data Fusion

Multi-source information fusion is also called multi-source data fusion or multi-sensor information fusion. The technology can be concluded as below: Realize the evaluation and decision process via automatic analysis and processing of information in selected multiple sensors in a time sequence with existing computer techniques. It can be described as the processing of multi-source information. The core content (**Error! Reference source not found.**) of multi-source information fusion is how to coordinate the relations between various information types and integrate them.

In the past decade of research practice, domestic and foreign specialists have gradually encapsulated the following representative data fusion methods (**Error! Reference source not found.**-**Error! Reference source not found.**):

(1) Weighted Voting: No training is required for datasets. Weighted value is calculated as per the majority voting rule. This method, however, fails to categorize performance variances of various attack types.

(2) Bayesian Methodology: Multi-sensor data is calculated in training computing via prior probability and conditional probability. Finally, the maximum result of posterior probability is selected as the fusion output.

(3) D-S Evidence Theory Method: This method is mainly applied in data, for which no prior probability nor conditional probability can be established, to convert objects to sets and resolve uncertain problems by establishing correspondence between sets and propositions.

(4) Fuzzy Logical Method: Sample fuzzy attributes in multi-source data. Find similarity via classification analysis and get output.

(5) Neural Network Method: Data is processed concurrently and synchronously. Learning is realized with the self-organization of the system. Sub-neurons not only store and process information but also accept fuzzy, analog, and random information.

(6) Cluster Analysis Method: It is a process of similar objects to get together and each type of similar objects forms a cluster. This analysis method is to analyze differences and similarity between data types.

(7) Synthetical Average Method: It is to get the output by calculating the average weighted of multi-source data.

See Table 1 for summarization and comparison of various data fusion methods.

Table 1. Various Data Fusion Methods Are Summarized and Compared

Data fusion method	Operating environment	Information type	Presentation of information	Uncertainty	Fusion Technique	Applicable Scope
Weighted average	Dynamic	Redundant	Original recording value	Weighted average	Weighted average	Fundamental data fusion
Kalman filtering	Dynamic	Redundant	Probability distribution	Gaussian noise	System simulation filtering	Fundamental data fusion
Bayesian estimation	Static	Redundant	Probability distribution	Gaussian noise	Bayesian estimation	Fundamental data fusion
Statistical decision theory	Static	Redundant and complementary	Probability distribution	Cumulative noise	Extremum decision	High-level data fusion
Evidence reasoning method	Static	Redundant and complementary	Proposition		Logical reasoning	High-level data fusion
Fuzzy reasoning	Static	Redundant and complementary	Proposition	Membership	Logical reasoning	High-level data fusion
Neural network	Static /dynamic	Redundant and complementary	Neuron input	Learning error	Neural network	Fundamental/high-level Data fusion
Production rule	Static	Redundant and complementary	Proposition	Confidence factor	Logical reasoning	High-level data fusion

2.2 Event Correlation Methods

Currently, massive log information analysis researches have been continuously proposed various event correlation methods, among which common ones are divided into three categories (**Error! Reference source not found.-Error! Reference source not found.**):

(1)Establish a database according to specialist knowledge. Match and then conduct correlation analysis of alarms according to previous attack sequences in the database.

(2)Define attack similarity for various attack types to identify if attacks are of the same type.

(3)Establish correlation by judging the possible prerequisites and subsequent results for the generation of an attack event.

The first method is efficient but with more limitations. This is because the attack sequences used in matching are determined by specialist knowledge in the database. In case of persistent new attack types, this method turns inefficient. There will be a large number of missing correlation results.

*Qianyun Wang

The second method calculates similarity only according to attack definition. It can generate fusion results for alarms but may cause large errors due to value selection in the similarity calculation.

The third method is much more flexible than the first one. It correlates alarms according to prerequisite sets and result sets, without prior input of specialist knowledge with all attack sequences. This method, of course, has its limitations. Correlation results are subject to the definition of prerequisite sets and result sets.

This paper aims at algorithm improvement based on the third method after comparison between the three above mentioned methods. Based on the third method, new definitions are added to reduce edge attributes and improve efficiency. In addition, heterogeneous multi-source log data is fused in formalized expression with fuzzy reasoning. Event correlation efficiency and accuracy are improved together with the fusion of security log information.

3. Preliminaries

3.1 Definition of Multi-source Security Log Fusion

According to descriptions of multi-sensor information fusion in various literature, this paper defines multi-source information fusion with security log as its research object as below: Multi-source security log information fusion indicates the method to decide and assess network situation via integrated analysis and processing of heterogeneous log information in various sensors, screening and filtering of various attributes in various logs, as well as integration of comprehensive information. In this method, network situation can be judged and threats can be evaluated rapidly and completely.

3.2 Basic Ideas of Fuzzy Reasoning

The mathematical description of a fuzzy set is as follows: For a known discourse domain X, if there is a fuzzy set N, then any $x \in X$. There must be a [0, 1] number $\rho N(x)$ to describe the membership of fuzzy set N in the discourse domain. It can be represented as $\rho N: X \rightarrow [0, 1]$, $x \rightarrow \rho N(x)$.The value of $\rho N(x)$ determines if x belongs to set N. The closer to 1 the value is, the more possible x belongs to set N (**Error! Reference source not found.**).

According to fuzzy set theory overview, uncertain object attributes can be expressed as fuzzy sets in the variable domain. Therefore, correlation degree in the correlation of events can be calculated according to the distribution probability of various attributes in the value domain. Form fuzzy relation in discourse domain with rules formed by prerequisites or results in linguistic variables. Define an operator μ as a part of the fuzzy relation for expression. In formalized expression, it will be:

$$R: X \times Y \rightarrow [0, 1] : (x, y) \rightarrow \mu(W(x), M(y)), \mu(x, y) \in X \times Y$$

For example, a rule in the fuzzy relation between discourse domains X and Y is: if Q is W, then Z is M.Q and Z in the rule are variables in discourse domains X and Y. W and M are fuzzy sets in the discourse domains. Fuzzy relation R can be regarded as the possibility that Z is y when Q is x.

3.3 Definitions of Event Scenario Correlation

The following definitions are required in the correlation of attack alarms in logs:

Definition 1: Define a triple *TRI = (work, pre, res)*, in which work represents the set of all attribute names in a log event. Different attribute names are correspondent to different values, while *pre* and *res* respectively represent the prerequisite set for an event to happen

and the result set of the event to happen. To analyze an attack event, bringing the event into the triple. Event is correspondent to work. If the attack event succeeds, its prerequisite *pre* must be true. If the event is judged to be successful, the result event *res* can be true.

For example, launch an SYN FLOOD attack against the host device and send massive connection requests of false source addresses to the host to consume host resources and result in DoS (Denial of Service) by host server or the network. This attack in formalized language is :

Synflood-Dosa=({IP, Port}, synfloodattack(IP)∧Exist(IP), {Dosa})

In the language, *IP* and *Port* respectively represent IP address and port number of host being attacked. If the attack succeeds, the prerequisite is that there is a host with IP address *IP* under the SYN FLOOD attack. If the attack succeeds, a possible result is the occurrence of server DoS.

Definition 2: Define an instance *w* (sub-set of work) of work in the specified triple. Each group in the triple has a period with a start time *s_t* and an end time *e_t.* For the example in Definition 1, the SYN FLOOD attack against the host can be expressed in a formalized way as :

Synflood-Dosa=({10.250.95.46,2336}, synfloodattack(10.250.95.46)∧Exist(10.250.95.46) , {Dosa(10.250.95.46) }).

Above mentioned information in formalized expression can be understood as the correlation logical combination of an attack event. According to analysis of a part of the logs, the recorded information is an individual attack event. However, such independent attack events can be categorized into a series of events with *s_t* and *e_t* added.

Definition 3: Represent the verb set of the prerequisite set and result set in the triple *TRI* with *P(TRI)(R(TRI))*.For the example in Definition 1, it can be expressed as

P(Synflood-Dosa)={synfloodattack(10.250.95.46)∧Exist(10.250.95.46)} , C(Synflood-Dosa)={Dosa(10.250.95.46)}.

In the processing of the two instances *w1* and *w2*, $r \in R$ for all results in between the start and end time, If conditions $p \in P(w2)$ and *R(w1)* can be met when *p* can be launched via the set of verbs in *R*, then *w1* can be prepared for *w2*. An example can be defined as below:

Synflood-Dosa=({10.250.95.46,2336}, synfloodattack(10.250.95.46)∧Exist(10.250.95.46) , {Dosa(10.250.95.46)})

is an event. Besides,

ping =({10.250.95.46, 2336}, pingattack(10.250.95.46)∧Exist(10.250.95.46) , {shut down(10.250.95.46)})

is another event. When the time of a ping attack event encounters the time of a SYN FLOOD attack event, it can be found in the verb combination that there is an item about the attack against the same IP address in both events. The two events can be, in a specified period, judged as two correlated events.

*Qianyun Wang

4. Mathematic Model of Multi-source Security Log Fusion based on Fuzzy Reasoning

Multi-source fuzzy fusion is to introduce alarms collected from various sensors to the membership function in set theory according to their attributes and influence weight, and divide unclear boundaries between false positive, false negative and useful information using the concepts of membership function and fuzzy relation matrix in set theory, in order to reduce alarm (to be fused) quantity and improve the fusion accuracy of multi-source log information. This paper, with reference to existing fuzzy reasoning clustering methods, proposes the mathematic model of multi-source log fusion based on fuzzy reasoning clustering.

The raw data matrix is established, with log attributes as the characterization index of samples, as well as historical log alarm information template in the database and sample alarm information in the world of experience as sample space. With the setting of m number of samples and n number of attribute indexes, the raw data matrix is as follows:

$$\begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix} \quad (1)$$

In the formula, a_{ij} ($1 \le i \le m$, $1 \le j \le n$) represents the membership function value of the i^{th} sample under the j^{th} index. Calculate the mean $\bar{Y}_j$ and standard deviation σ_j of each index. The formula is as follows:

$$\bar{Y}_j = \frac{1}{m}\sum_{i=1}^{m} a_{ij} \; ; \; \sigma_j = \sqrt{\frac{1}{m-1}\sum_{i=1}^{m}(a_{ij} - \bar{Y}_j)^2} \; (1 \le j \le n)$$

Merge the original data matrices to get the fuzzy sets $A=(a'_{ij})_{m\times n}$ of the sample space at the n number of attribute indexes:

$$a'_{ij} = \frac{a_{ij} - \bar{Y}_j}{\sigma_j} \quad (2)$$

Then the fuzzy similarity matrix $R=(r_{ij})_{m\times n}$ at discourse domain A can be determined:

$$r_{ij} = 1 - d(x_i, x_j) \quad (3)$$

In the formula, $d(x_i, x_j)$ is the Euclidean distance between x_i and x_j. The calculation formula is as follows:

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{m}(x_{ik} - x_{jk})^2}$$

In the formula, X_i is the fuzzy set of the i^{th} sample at the n^{th} index.

Next, determine the weight factor β. Reliability will be questioned if the weight is determined only according to experience or qualitative setting. Therefore, coincident matrix method is adopted to compare all factors in a paired comparison method but not in a direct comparison of all factors. Relative scale is adopted in comparison between different attributes to reduce difficulty and improve accuracy.

Calculation method of weight factors: Assume that sample x contains n number of attributes c_1, c_2, c_3, …, c_n. Compare the attributes in a paired comparison method according to the fusion rules as per the formula below to get weighted values B= (β_1, β_2, β_3, …, β_n) :

$$\beta_i = \frac{\sum_{j=1}^{n} x_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{n} x_{ij}}$$

(4)

Finally，It can be represented with the weighted average model with fuzzy evaluation that can cover all factors:

$$V = R \times B$$

(5)

$$v_j = \sum_{i=1}^{n} \beta_i \times r_{ij} \ (\ j = 1, 2, 3, ..., n)$$

(6)

5. Event scenario Correlation Method Combined with Fuzzy Reasoning

5.1 Calculation of Correlation Similarity

This paper selects the following five attributes for the calculation of correlation similarity: timestamp, source/destination IP address, port, protocol and event type. The weights of five attributes are calculated according to formula (1-6), respectively represented by B_1, B_2, B_3, B_4, and B_5.

Define e_0 to represent the event information in the current period and $E=\{e_1,e_2,e_3,...,e_n\}$ to represent historical event information. $e_i.time$ represents the time when an event happens, $e_i.sip$ and $e_i.dip$ represent the source and destination IP addresses when an event happens, $e_i.port$ represents the port number of an event, $e_i.protocol$ represents the protocol of an event, $e_i.type$ represents the type of event.

(1) Time Similarity

There is an interval between two events. When two events (e_0, e_1) happened in an interval shorter than the specified minimum interval (as $|e_i.time-e_0.time|<T_{min}$), time similarity a_{i1} gets the value 1. When two events happened in an interval longer than the specified minimum interval, $|e_i.time-e_0.time|>T_{max}$ judges that the two events do not correlate with each other. Similarity a_{i1} gets the value 0. When two events happened in an interval in between the specified minimum interval and maximum interval(as $T_{min}<|e_i.time-e_0.time|<T_{max)}$, similarity a_{i1} gets the value $(T_{max}-|e_i.time-e_0.time|)/(T_{max}-T_{min})$.

*Qianyun Wang

(2) Source/Destination IP Address Similarity

The similarity of source/destination IP address between two events is represented with a_{i2}. The calculation is as follows:

$$a_{i2} = \frac{\varphi}{32}$$

In the formula, φ means that $e_0.sip$ & $e_i.sip$ have the same left-to-right subnet mask mantissa digits as $e_0.dip$ & $e_i.dip$.

(3) Port Similarity

Port similarity is represented with a_{i3}. When the port number of the current event information is identical to that of the event in the historical template, the similarity value is 1. Otherwise, the similarity value is 0.

(4) Protocol Similarity

Protocol similarity is represented with a_{i4}. When the protocol of the current event information is identical to that adopted in the information in the historical template, the similarity value is 1. Otherwise, the similarity value is 0. Protocols include common ones such as TCP, UDP, ICMP, HTTP, FTP, and so on.

(5) Event Type Similarity

Event type similarity is represented with a_{i5}. When the type of event information is identical to the type of information in the historical template, the similarity value is 1. Otherwise, the similarity value is 0. Main event types include various alarms and so on.

Introduce membership functions in the five similarity types to the matrix (1) in Section 4 to get membership function matrix, and then get fuzzy similarity matrix via merging and vector distance calculation of formulae (2) and (3). Get the final quantitative results via formulae (4) and (5) according to the assigned weight factors $v_j=\sum_{i=1}^{n} \beta_i \cdot r_{ij} (j=1, 2, \cdots, n)$. According to the maximum membership degree principle, assume that $v_k=max\{v_1, v_2, v_3,..., v_k\}(1\leq k\leq n)$. Compare current events with those in historical event template one by one. If the probability of e_0 and e_k belong to the same type of alarms is the largest, merging e_0 and e_k into the same alarm type.

5.2 Log Data Fusion based on Fuzzy Clustering

After a series of preprocessing and rule library information matching of original log data, conduct fuzzy clustering of alarm events. See Figure 1 for the detailed process.

Input: new alarm event $A=\{a_1,a_2,a_3,...,a_n\}$, historical event template $E=\{e_1,e_2,e_3,...,en\}$
Output: fusion event $Y=\{y_1,y_2,y_3,...,y_n\}$

01 $C \leftarrow \Phi$ Make set C null.
02 $C \leftarrow E$ assign historical event template to set C.
03 For each $a_i \in A$ Do
04 $v_i \leftarrow V_a(a_i)$ Calculate the membership of event a_i to the historical event template E.
05 If $v_i \geq e_{hatm}$ Then
06 $C = C$
07 Else $C = C \cup \{a_i\}$
08 End if
09 End for
10 $Y \leftarrow C$
11 Return Y

Figure 1. Detailed Process to Conduct Fuzzy Clustering of Alarm Events

Thresholds can be set flexibly according to the actual alarms in the on-site environment and the actually detected FP rate and FN rate. Once the threshold of the current event is larger than the preset one, it can be identified that the current event is greatly correlated to an event in the historical event template and can be merged into the same alarm type as that template. Meanwhile, the current alarm event will be updated in the database. Different information will be recorded in information sheets of correlated alarms. Once the threshold of the current event is smaller than the preset one, it can be identified that the current event does not have alarm events in the historical event template that meet relevant membership, i.e. the current event is a new malicious event. New alarm and alarm quantity will be added to the database. New alarm information will be updated in the historical event template. A feedback will be returned and a new alarm will be generated.

To better filter factors of various sets, destination sets are divided into multiple fuzzy levels. Data with fuzzy level divided of membership degree [0, 1] can fully match the value conditions of the fuzzy system and form a good statistical support to the fuzzy system.

5.3 Improvement of Correlation Algorithm with Fuzzy Clustering

The conventional log event correlation calculation seeks the correlation of different events in correlation methods and judges whether events belong to the same attack/alarm sequence according to correlation degree. A threshold for correlation degree will be set before the start of correlation. If the correlation degree of two events that are correlated in a correlation method is higher than the threshold, then they belong to the same attack/alarm sequence. If the correlation degree is lower than the threshold, new events will be defined and recorded. The expected final result is that the similarity of events from the same attack/alarm sequence is above the threshold and that of events not from the sequence is below the threshold. However, similarity of completely irrelevant events may be judged as above the threshold due to errors in correlation. This paper aims at reducing the uncertainty of edge events by introducing fuzzy reasoning to event correlation. It will better reduce alarm redundancy and add limitations into the judgment of event correlation.

There are numerous events in log information. However, not all events can be correlated. Some of them are independent events. So, force correlation is not recommended, and limitation measures should be introduced to make the model more effective. In model design, correlated or not is a judgment entry. The newly added

*Qianyun Wang

definition determines that whether this event is in correlation processing with another one or not.

New Definition: Add a judgment time difference T between events. When two different events w_1 and w_2 happened, record their event time T_1 and T_2. The time difference concept is used to judge whether to correlate the events or not. If correlation is required, the time difference $/T_1\text{-}T_2/$ between events w_1 and w_2 must be smaller than the preset judgment time difference T.

Introduce the fuzzy set concept:

In the correlation of events recorded in log information, some correlation conditions cannot be met in direct correlation of all events. This will influence the results. Thus, this paper adopts fuzzy reasoning to weaken set boundaries. See Figure 2 for the details of the improved algorithm with fuzzy reasoning integrated into correlation.

Required databases: WorkL, RulesL and LogL.
Required inputs: time interval *TI*, start time *s_t*, and end time *e_t*.
Required results: event correlation results *WCR*.

(1) Pre-process collected original data according to the required standard format.
(2) Store multi-source log data in log database LogL.
(3) Categorize and number various log events.
(4) Find all events in the scope of *s_t* and *e_t* by search.
(5) Number the earliest event *s_id* and the last event *e_id*.
(6) Have fuzzy clustering of each group of attributes of log data with the fuzzy clustering algorithm, divide them into several fuzzy sets, and set thresholds according to the network environment.
(7) Set TypeID according to the preset thresholds.
(8) Initialize input data and libraries: $Log\_id \leftarrow s\_id$. $LogL \leftarrow \phi$. $WCR \leftarrow \phi$.
(9) $Judge \leftarrow Flase$. (*Judge* is the flag bit to judge that the information with event ID *w_id* may be an independent event and should be added to another separate JudgeL library.)
(10) For $Log2\_id \leftarrow Log1\_id + 1$ to log_id and $Log_{Log2\_id}.workTime - Log_{Log1\_id}.workTime \leq TI$
(11) Number TypeID's and WorkL's of various origins in data sources with WorkTypeID's.
(12) For each rule in rule library R R_i($i \leftarrow 1$ to r_num)
  If ($R_i.pre\_id = W_{W1\_id}.workTypeID$ and $Ri.res\_id = W_{W2\_id}.workTypeID$) then
    { ADD ordered pair <*W1_id*, *W2_id*>
    , ADD *E* to ResL
    , and *Judge*←True
    , break }
  If(*Judge*=False) then
    ADD W_{W_id} To JudgeLab.
(13) W_id++,
  if(W_id<=w_id) then
    goto(3)
  Else
    goto (5);
(14) The final output is the alarm attack correlation result (WCR) formed together by result correlation database ResL and isolated event library JudgeL.

Figure 2. Detailed Process of the Improved Algorithm

6. Experiments and Analysis

This section conducts experiments on VAST2013 dataset and presents analyses.

6.1. Dataset Description

This paper selects heterogeneous information, such as host logs, snort logs, and switch logs, as data sources as well as select datasets in the international standard dataset VAST2013 as validation sets and test sets, in order to realize a more accurate evaluation of network security situation.

In an established testing environment, attacks against hosts and switches in the testing network are simulated from both external network and LAN. Attacks consist of the following types:

SYNFlood Attack: Launch attacks against hosts respectively from the internal LAN in the experimental environment and the external network.

ARP Attack: Launch attacks against switches from the internal LAN in the experimental environment.

UDPFlood Attack: Launch attacks respectively against hosts and switches from the internal LAN in the experimental environment and the external network.

TCP Concurrent Attack: Launch attacks against switches inside the internal LAN in the experimental environment.

ICMPFlood Attack: Launch attacks respectively against hosts from the internal LAN in the experimental environment and the external network.

GET Request Attack: Launch attacks respectively against hosts from the internal LAN in the experimental environment and the external network.

6.2. Experimental Results and Analysis

The simulation attack experiment lasts for five days. In the period, 102,541 pieces of logs are collected. Among these logs, there are 48,356 pieces of logs collected by the Snort intrusion detection system. 17,249 pieces are Linux host logs. 32,526 pieces are Windows host logs. 4,410 pieces are switch logs. See Figure 3 for the details. Y-axis of Figure 3 is the number of attacks.

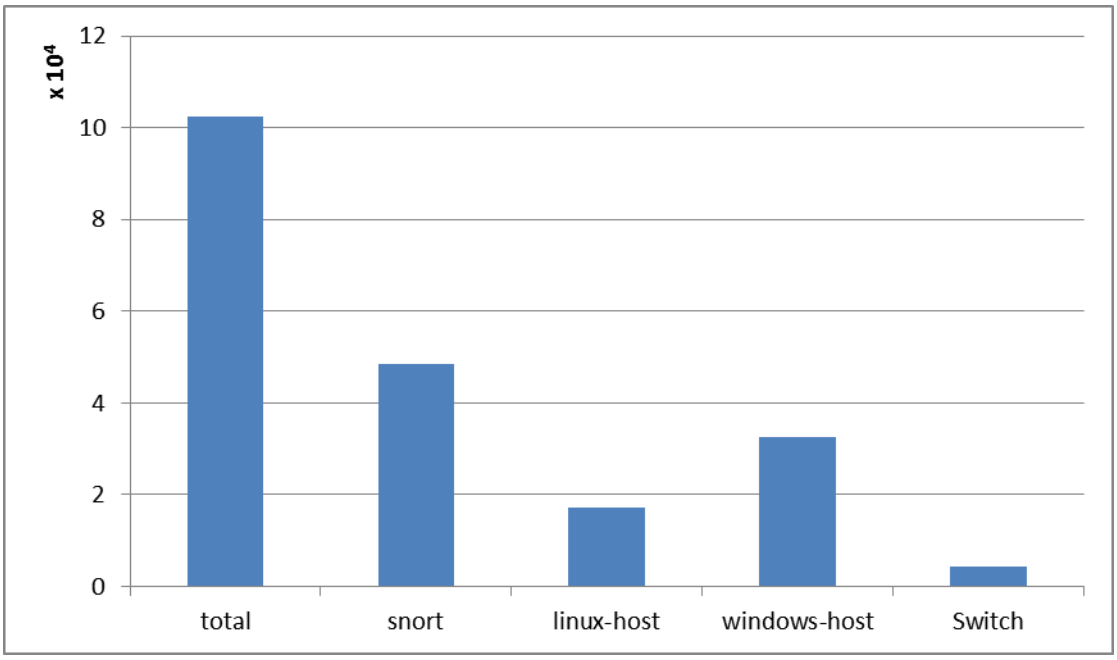


Figure 3. Quantity Statistics of the Log Information

*Qianyun Wang

The statistical analysis of all traffic data covers from addresses of each device to their input and output packet data, as well as their input and output byte number. See Table 2 for the details.

Table 2. Traffic Statistics of Different IP Addresses

Device IP address	Input packets	Output packets	Input bytes	Output bytes
10.250.95.18	13865	18172	1358401	8679035
10.250.95.46	17429	23356	1803326	12428017
10.250.95.77	14658	16489	1092365	15087641
192.168.24.11	24388	26138	5208265	39212754
192.168.24.85	21904	23746	4639134	24603719
192.168.24.96	27565	25371	6530026	27142526

After format unification as well as attribute reduction and filtering of collected original data, Figure 4 illustrates the results of direct and separate analysis of log information from various sensors without the proposed event correlation method. Y-axis of Figure 4 is the percentage of FP rate and FN rate.

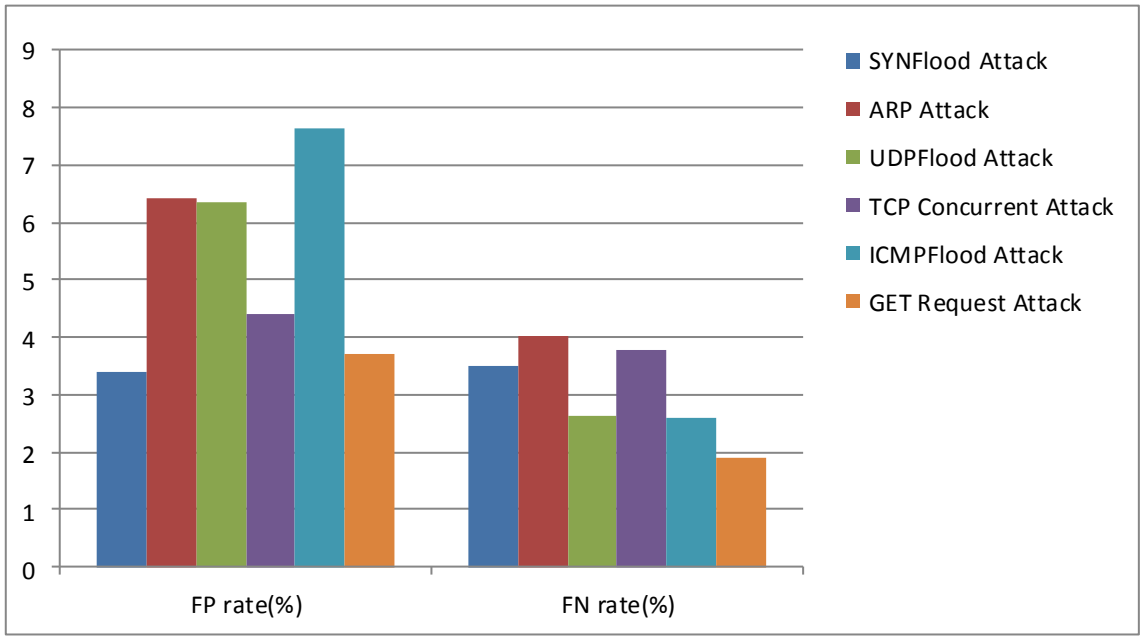


Figure 4. Results of Single Sensor

Figure 5 illustrates the analysis results of data using event scenario correlation combined with fuzzy reasoning. Y-axis of Figure 5 is the percentage of FP rate and FN rate.

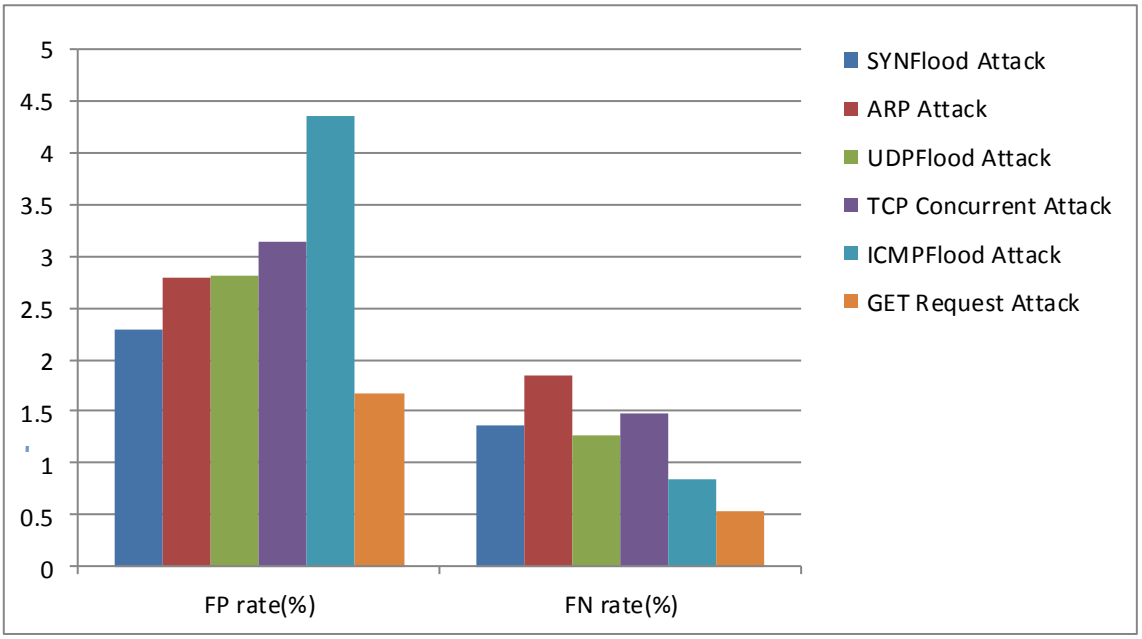


Figure 5. Results of Using the Proposed Event Correlation Method

See Table 3 for the comparison between experimental results in conventional event scenario correlation method and the event scenario correlation method combined with fuzzy reasoning.

Table 3. Comparison between Experimental Results in the Conventional Correlation Method and the Improved Correlation Method

Attacks	Improved correlation method		General correlation method	
	FP rate	FN rate	FP rate	FN rate
SYNFlood	2.286%	1.355%	3.325%	1.652%
ARP	2.793%	1.852%	3.973%	2.178%
UDPFlood	2.817%	1.264%	4.168%	1.442%
TCP Concurrent	3.148%	1.471%	3.840%	1.586%
ICMPFlood	4.351%	0.837%	6.007%	1.305%
GET Request	1.668%	0.538%	2.461%	0.693%

According to the experimental results, FP rate and FN rate are significantly improved in the event scenario correlation method combined with fuzzy reasoning. The reason is as follows: Compared to single log information source, fuzzy clustering fusion based on event correlation analysis can integrate features of multiple logs, interpret security events in a more complete way, and unearth the connection hidden behind events. In addition, event scenario correlation method combined with fuzzy reasoning has, compared to conventional event scenario alarming correlation methods, better boundary judgment of events with pairwise correlation and fusion. It also reduces uncertainty of edge events. Thus, it is more effective to distinguish correlatable events and non-correlatable events as well as repeatable events and independent events.

7. Conclusion and Future Work

This paper studies the categorization of multi-source log fusion methods and compares existing fusion methods. According to log features, this paper proposes the event scenario

*Qianyun Wang

correlation method combined with fuzzy reasoning by integrating fuzzy clustering in the conventional event scenario correlation method. According to tests in an experimental environment, it is verified that the proposed method can reduce the False Positive rate and False Negative rate (FN rate).

In future research, it is imperative to test and verify the method in a larger-scale network since the established testing network environment is different from the large-scale network in the world of experience.

Acknowledgments

We are grateful to the editors and anonymous reviewers for their valuable comments on this paper.

The work of this paper is supported by National Natural Science Foundation of China (Grant No. 61272399 and No. 61572090) and Research Fund for the Doctoral Program of Higher Education of China (Grant No. 20110191110038).

References

[1] J. Ahmad, "Network Security Situational Awareness", The International Journal of Computer Science and Communication Security (IJCSCS), vol. 8, no. 3, **(2013)**, pp. 61-67.
[2] Z. Fangfang, S. Ronghua, Zhao Ying, Huang Yezi, and Liang Xing, "A visualization system for network security situational awareness", Proceedings of 5th International Symposium on Cyberspace Safety and Security, Zhangjiajie, China, **(2007)** November 13-15.
[3] W.Yong, L.Yifeng, and F.Dengguo, "A Network Security Situational Awareness Model Based on Information Fusion", Journal of Computer Research, vol. 846-847, no. 3, **(2013)**, pp. 1632-1635.
[4] Wang Juan, Qin ZhiGuang, and Ye Li, "Research on prediction technique of network situation awareness", Proceedings of the IEEE Conference on Cybernetics and Intelligent Systems, Chengdu, China, **(2008)** September 21-24.
[5] Huang Yihai, Hu Jun, "Design and implementation of log audit system", Computer Engineering, vol. 32, no. 22, **(2006)**, pp. 67-68.
[6] Makarau Aliaksei, Palubinskas Gintautas, and Reinartz Peter, "Factor graph models for multisensory data fusion: From low-level features to high level interpretation", Proceedings of the IEEE International Conference on Geoscience and Remote Sensing Symposium (IGARSS), Munich, Germany, **(2012)**, July 22-27.
[7] Bass Tim, "Intrusion detection systems and multisensory data fusion", Communications of the ACM, vol. 43, no. 4, **(2000)**, pp. 99-105.
[8] Tmazirte Nourdine Ait, Najjar Maan E. EI, Smaili Cherif, and Pomorski Denis, "Multi-sensor data fusion based on information theory. Application to GNSS positionning and integrity monitoring", Proceedings of the 15th IEEE International Conference on Information Fusion (FUSION), Singapore, Singapore, **(2012)**, September 7-12.
[9] Potdar Akshay A, Longstaff Andre P, Fletcher Simon, and Mian Naeems S, "Application of multi sensor data fusion based on Principal Component Analysis and Artificial Neural Network for machine tool thermal monitoring", Proceedings of 11th International Conference and Exhibition on Laser Metrology, Coordinate Measuring Machine and Machine Tool Performance, Queensgate, West Yorkshire, United kingdom, **(2015)** March 17-18.
[10] Piella G, "A general framework for multiresolution image fusion: from pixels to regions", Information Fusion, vol. 4, no. 4, **(2003)**, pp. 259-280.
[11] Llinas, James, Hall, and David L, "An introduction to multi-sensor data fusion", Proceedings of the IEEE International Symposium on Circuits and Systems, Monterey, USA, **(1998)** May 31-June 3.
[12] Thomas Hilker, Michael A. Wulder, and Nicholas C. Coops, Julia Linke, Greg McDermid, Jeffrey G. Masek, Feng Gao, Joanne C. White, "A new data fusion model for high spatial- and temporal-resolution mapping of forest disturbance based on Landsat and MODIS", Remote Sensing of Environment, vol. 113, no. 8, **(2009)**, pp. 1613–1627.
[13] Durrant-Whyte Hugh, "Data Fusion in Sensor Networks", Proceedings of the 10th IEEE International Conference on Video and Signal Based Surveillance, Sydney, NSW, Australia, **(2006)** November 22-24.
[14] Renzo Marco Di, Imbriglio Laura, Graziosi Fabio, and Santucci Fortunato, "Distributed data fusion over correlated log-normal sensing and reporting channels: Application to cognitive radio networks", IEEE Transactions on Wireless Communications, vol. 8, no. 12, **(2009)**, pp. 5813-5821.
[15] Bonnie W. Morris, Virginia Franke Kleist, Richard B. Dull, and Cynthia D. Tanner, "Secure information market: A model to support information sharing, data fusion, privacy, and decisions", Journal of Information Systems, vol. 28, no. 1, **(2014)**, pp. 269-285.

[16] Tabar A M, Keshavarz A, and Aghajan H, "Smart home care network using sensor fusion and distributed vision-based reasoning", Proceedings of the 4th ACM international workshop on Video surveillance and sensor networks, Santa Barbara, United states, **(2007)** October 27-27.
[17] Ticha M B B, and Ranchin T, "A case based reasoning data fusion scheme: application to offshore wind energy resource mapping", Proceedings of the 9th International Conference on Information Fusion(FUSION), Florence, Italy, **(2006)** July 10-13.
[18] Zhang Shuying, Gao Yue, Zhang Mengqun, and Wang Shuangli, "The Study of Network Security Event Correlation Analysis Based on Similar Degree of the Attributes", Proceedings of the 4th International Conference on Digital Manufacturing and Automation (ICDMA), Qindao, Shandong, China, **(2013)** June 29-30.
[19] Lee D H, Kim J G, and Kim K J, "A study on abnormal event correlation analysis for convergence security monitor", Cluster computing, vol. 16, no. 2, **(2013)**, pp. 219-227.
[20] Liu Jing, Gu Lize, Xu Guosheng, and Niu Xinxin, "A correlation analysis method of network security events based on rough set theory", Proceedings of the 3rd IEEE International Conference on Network Infrastructure and Digital Content (IC-NIDC), Beijing, China, **(2012)** September 21-23.
[21] Noda Masaru, Higuchi Fumitaka, Takai Tsutomu, and Nishitani Hirokazu, "Event correlation analysis for alarm system rationalization", Proceedings of Process Systems Engineering (PSE), Asia, **(2011)** May-June.
[22] Pedrycz W, and Rai P, "Collaborative Fuzzy Clustering with the use of Fuzzy C-Means and its Quantification", Fuzzy Sets and System, vol. 159, no. 18, **(2008)**, pp. 2399-2427.
[23] Takai Tsutomu, Noda Masaru, and Higuchi Fumitaka, "Identification of nuisance alarms in operation log data of ethylene plant by event correlation analysis", Kagaku Kogaku Ronbunshu, vol. 38, no. 2, (**2012**), pp. 110-116.

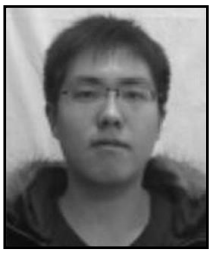
Authors

Qianyun Wang, He received her B.S. degree in Chongqing University, P. R. China, at 2013. Currently she is a full-time M.S. candidate in College of Software Engineering, Chongqing University. Her research interests include distributed systems and large-scale data mining.

Shuyu Chen, He received his Ph.D. degree in Chongqing University, P. R. China, at 2001. Currently, he is a professor of College of Software Engineering at Chongqing University. His research interests include distributed systems, cloud computing, and embedded Linux system. He has published over 120 journal and conference papers in related research areas during recent years.

Hancui Zhang, He received her M.S. degree in Chongqing University of Posts and Telecommunications, P. R. China, at 2011. Currently, she is a Ph.D. candidate in College of Software Engineering, at Chongqing University. Her current interests include cloud computing, large-scale data mining, flash memory, and fault detection.

*Qianyun Wang



Tianshu Wu, He received his B.S. degree in Chongqing University of Posts and Telecommunications, P. R. China, at 2011. Currently, he is a Ph.D. candidate in College of Computer Science, at Chongqing University. His current interests include cloud computing, large-scale data mining and fault detection.