The 9th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications
21-23 September, 2017, Bucharest, Romania

Correlation of security events based on the analysis of structures of event types

Andrey Fedorchenko[1,2], Igor Kotenko[1,2] and Didier El Baz[2,3]
[1] St. Petersburg Institute for Informatics and Automation of Russian Academy of Sciences (SPIIRAS),
39, 14th Liniya, St.Petersburg, Russia, {fedorchenko, ivkote}@comsec.spb.ru
[2] St. Petersburg National Research University of Information Technologies, Mechanics and Optics (ITMO University),
49, Kronverkskiy prospekt, Saint-Petersburg, Russia
[3] LAAS-CNRS, Universite de Toulouse, CNRS, Toulouse, France, elbaz@laas.fr

Abstract—The paper studies the process of correlation for SIEM systems based on analyzing the structures of security event types. The approach to automated analysis of security events as input data with dynamic content is proposed. For the automated analysis of events the paper suggests to build a graph of types of events with direct and indirect links between them. Processing of security input data means performing functional and behavioral analysis by computing the frequency-time characteristics of events, their ranking and building of patterns of behavior. The proposed approach allows to use a previously not applied method of rank correlation, alongside with other intelligent methods. The requirements to the normalization of original data are formulated. An example of the analysis of the security event log and the generated graph of event types are provided.

Keywords—security monitoring, data correlation, security events, structural analysis.

I. INTRODUCTION

Currently on the world market there is a great number of different classes of instruments to ensure the security of computer infrastructures. These instruments are aimed at warning, detection and prevention of cyber attacks and malicious activity, as well as monitoring and managing the current level of security. One of the classes of such instruments is SIEM (Security Information and Event Management) systems, which are being developed for over 10 years.

The main task of SIEM systems is to collect certain heterogeneous information and revealing in it of the high level incidents and warnings about the security breach [1,2]. To achieve this task, usually they use methods of normalization, aggregation, filtering, and correlation of events. However, with the development of such threats to the security of computer infrastructures as targeted attacks and attacks on the Internet of things (IoT), currently applied methods and approaches are often unable to provide an adequate level of security. This trend is aggravated by the increasing amount of data, processing of which is becoming more difficult. The process of data correlation in instruments of protection of the SIEM class plays a fundamental role. This process basically aims at defining causal relationships between processed events. It enables the detection of malicious, attacking and abnormal activity, the determination of the source and target of an attack, detecting multi-stage attacks, and depends on the implementation in a specific solution [3]. Despite the diversity of methods and approaches applied in the process of correlation, the most widely used is the rule-oriented method.

In this paper we present an approach to correlation based on analysis of the event types to determine relationships between them. The features of this approach are the use of the generated graph of types of events with direct and indirect links among themselves to meet the functional and behavioral analysis by computing the frequency-time characteristics of events, determining causal relationships, ranking events and building the patterns of behavior.

This paper examines the place, role and general principles of correlation process, the global task is set of development of methods of correlation and private task of automating the analysis of raw data. Individual steps of the proposed approach are studied. The study reveals the process of forming a non-directed graph of relations of types of events, performed on the basis of the analysis of the types of events within one event log taking into account its normalized representation. The peculiarities of the use of the obtained data for making functional and behavioral analysis are presented. As a result of development of the proposed approach for its correct operation, the necessary and sufficient conditions of the use of the original (input) data are formulated. We also describe the experiment for the analysis of the events types for the security log of OS MS Windows. The results of the experiment and evaluation of the proposed approach to the correlation of security events are provided.

II. RELATED WORKS

Correlation of data has been initially applied in the Intrusion Detection Systems (IDS) for identifying relationships between network events with the purpose of their aggregation and subsequent detection of attacks

(including distributed and multi-step ones) [3]. Exactly from the systems of the given class the methods of correlation were adapted to correlation of information in SIEM systems.

In general, the correlation process can be divided into the following stages: (1) normalization; (2) aggregation; (3) filtering; (4) anonymization; (5) prioritization; (6) correlation [3]. Availability and supplementary decomposition of these stages in a particular solution depends on its implementation. From our point of view, each of the stages is necessary for full implementation of the correlation process.

At present there are many methods for correlation of events and information security with their advantages and disadvantages. At a particular stage of the correlation process it is appropriate to apply methods best suited to its task. Methods of implementation of the overall process of correlation in the existing solutions typically combine several methods. All methods can be nominally divided into signature-based and heuristic (behavior analysis). These methods can use different approaches based on similarity analysis, statistical analysis, data mining, etc.

The complexity of assessment of quality of used methods for data correlation is that the manufacturers of SIEM systems in order to protect intellectual property do not disclose the features of the technological solutions used in their systems. Besides, even when buying a SIEM system, the study of correlation module is hampered by the fact that its tuning is mainly in the formation of new (additional) rules and exceptions.

However, along with paid solutions of SIEM systems, there are also open source projects, as well as many scientific publications on methods and approaches to correlation of events and security information.

The most popular and easiest in implementation is the rule-based method [4-6] based on the fixed correlation of events with each other under certain conditions. These conditions may contain logical operations on the data, their properties and the calculated indicators. The main drawback of this method is the complexity and duration of the compilation of rules by the security administrator. Performing of correlation by the rule-oriented method also directly depends on the skill of the implementation specialist.

Many methods such as template-based (scenario-based) [4], graph-based [7,8], based on finite state machines [7,9], based on similarity [10,11] and others inherently have different models for representation of events and their relationships, but ultimately, they can also be expressed in the form of rules.

Modern direction of development of events correlation methods is the application of self-learning approaches to data mining such as Bayesian networks [4,7,12], immune networks [7,12], artificial neural networks [7,12-14] and others. The advantage of these approaches lies in the possibility of independent (unconditional) correlation of events with the minimization of manual settings. However, building of learning models requires a preliminary analysis of the data, which is not always possible to be automated. In addition, the application of intelligent approaches imposes a requirement for assessing the adequacy and quality of the models and the original training data should be fairly complete.

III. Approach to events correlation

A. Process of correlation

For formulation of research task, initially, we must determine the place and role of the correlation process in SIEM systems. It is believed that the correlation process is aimed at (1) determining the relationship between events and security information, (2) grouping of low-level events into higher-level events, and (3) detection of incidents and security alerts. Thus, the implementation of the correlation process starts with the collection of data from disparate sources and ends at the stage of formation of the report on the current state of security of the analyzed infrastructures. It should also be noted that the correlation process is continuous and should be designed to run in real time.

The global task is *to develop methods of automated correlation of heterogeneous security information.* To achieve this it is proposed to use the results of the structural, functional, behavioral and evolutionary analysis of protected objects. Suggested division of the task is defined by the relevant aspects of the complexity of computer infrastructures as complex dynamical systems. In the current research, the private task is *to develop the approach to correlation based on the analysis of event types*. The novelty of suggested approach is in the way to automate the search of causal relationships between disparate events to perform the correlation process. This approach would ensure a smooth addition to the model of correlation of events types that were not previously known, but only after conversion to the normalized representation.

Various sources of information (internal and external) may serve as input data of the correlation process: sensors of measurement, agents for data collecting, event logs, configuration of infrastructure objects and many others. The general scheme of the use of different data sources is shown in Figure 1, in which the platforms bases are the stores of information about identification characteristics of the installed hardware and software in the analyzed infrastructure. In this scheme, the input (raw) data is represented inside information with dynamic content and internal and external information with the semi-static content.

This separation is necessary because of the complexity of the correlation of information from different categories in a single process, the main distinction of which is the reference for the time scale (for dynamic content). It is also worth noting that at this

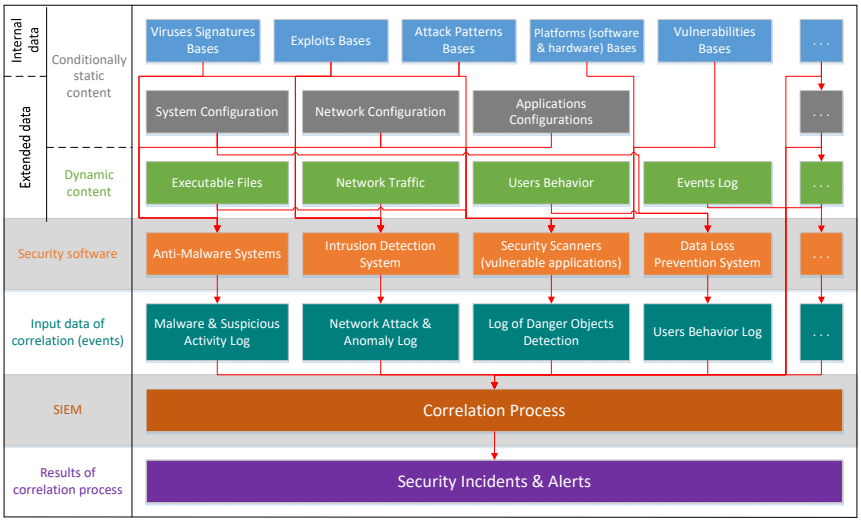

Figure 1. The general scheme of using the input data for the correlation process in SIEM systems

stage the developed approach focuses mainly on input data with dynamic content, since any change in the conditionally static information may also be represented as an event. However, this fact does not exclude accounting of the data with conditionally static content in the analysis of the security state. The scheme also includes protection tools, carrying out intermediate processing of input information and generating higher-level events. However, the connection of data source with the given tools is not fixed, that is, the use of a source by a particular tool depends on its implementation. Thus, the original data to perform the correlation process are heterogeneous and multi-level security events that must be taken into account when solving the global task.

B. *Analysis of events logs*

The event is understood as result of the action (completed, denied, failed) or attempt to commit the action generated by either source of action, or by its processing system having a predefined description format, understandable by processing system, and also having specific properties that describe the action itself.

The events of different types within a single log are the initial data for our research and can be expressed as follows:

$$\{e_1, e_2, \ldots, e_k\} \in E^L, \quad \{t_1, t_2, \ldots, t_n\} \in T^L, \quad (1)$$

where E is a set of events of the log L, and T is a set of events types of the log L.

Analysis of the event types is proposed to be held on real input data (events logs). In this case we eliminate the possibility of errors associated with changes in the format of types, and in the presence of such change, such events will be appropriately marked. On the basis of log analysis we make the formation of structures of events types with certain properties:

$$\{p_1, p_2, \ldots, p_m\} \in P^T, \quad (2)$$

where P is the set of properties of the set of types T.

Events properties can be nominally divided into the following groups:

- the identification properties whose values for each event are unique within a set of events of a single log (a group of logs) or the system. For example, the identifiers of event records;
- properties of membership, whose values indicate the content of events in certain sets, such as the type, provider, host;
- temporal properties that reflect the time value of creation, recording, start, finish, and other temporal characteristics of actions;
- properties of the audit, determining the result of action that describes the action as successful, forbidden, failed, etc.;
- information properties reflecting the specific characteristics of the actions described in the event (it is the most extensive group of attributes).

Thus, the logs analysis aimed to identify structures of types and their properties can be represented in the form of mapping of set of events to set of types and properties of events types:

$$E \rightarrow T, P. \quad (3)$$

C. *Correlation based on analysis of event types*

As identified types of events consist of properties characterizing the action described in the event, the relations between types of events by analyzing their structures are formed by relations between their properties. To determine the place of structural analysis in the problem of determining the relations, we must introduce the classification of *the relations between the properties of event types*. So the relations on *equivalent* and *nonequivalent* properties are separated.

Equivalent property p is the same property of two different event types t_1 and t_2:

$$\forall p \in P^T : \quad p \in P^{t_1},\ p \in P^{t_2};\quad t_1, t_2 \in T\ . \qquad (4)$$

In their turn the relationships *on nonequivalent properties* are divided into the same type and intertype ones. *The same type nonequivalent properties p_1 and p_2* are properties that are equivalent in content type:

$$p_1, p_2 \in P^T : \quad p_1 \sim p_2\ . \qquad (5)$$

Intertype nonequivalent properties are properties that are equivalent in content values with the apparent difference between the content types.

In addition, a single event type t can contain several same type and intertype nonequivalent properties p in its structure:

$$\forall \{p_1, p_2, \ldots, p_s\} \in P^t : \quad p_1 \sim p_2 \sim \cdots \sim p_s\ , \qquad (6)$$

where s is the number of the same type or intertype properties.

On subsequent analysis the presence of equivalent properties for different types of events will be treated as *direct relation* between the properties of events types, and the presence of the same type and intertype nonequivalent properties – as the *same type and intertype indirect relations*, respectively. However, within the framework of the structural analysis only direct connections between events types are considered, while the functional and behavioral analyses involve determining the indirect same type and intertype relations, respectively.

For example, when comparing structures of two types of security events of OS MS Windows: "creation of process"(4688) and "completing the process"(4689), one of the equivalent properties of both types is the "ProcessId" (initiator process), which is a direct relation between these types of events.

When analyzing the structure of the event type "Start process"(4688) in addition to properties "ProcessId" there were identified properties "NewProcessId" and "Execution ProcessId". All three properties characterize the process identifier, only in the first case – initiator process ("parent"), in the second – child process (the heir) and in the third – process-source (agent) of the event.

This relation is indirect same type on the content type (the type is "processId"), therefore it allows us to trace additional functional relations between events of different types and in this case to identify the events of working sessions of processes and their inheritance hierarchy.

Also the event type "Start process"(4688) contains a property named "ProcessName". When considering the properties "ProcessId" and "ProcessName", the types of their contents are clearly different: "ProcessId" in the first case, and "Location (executable file) in the file system" in the second one. However, both properties describe the identifying characteristics of the process. In the first case, this characteristic is tied to the time scale: the identifier is assigned by the system to each created process and has a unique random value within the session process (from creation to completion). In the second case, the identification feature is more static and has no reference to time scale. As the result of calculating the frequency-time characteristics between the values specified between the nonequivalent properties of different type indirect intertype relation can be determined.

Further, the analysis of structures of events types an undirected graph G is formed; this stage completes the structural analysis of events types:

$$G = (T, P, \varphi), \quad \varphi : P \to T \times T\ . \qquad (7)$$

In the future research we are planning to evolve the proposed approach to correlation by functional and behavioral analysis. These stages are to be implemented through frequency, time-frequency and pattern analysis of events in logs. The purpose of these stages is to compute the directions of links in the graph, generate data for holding the behavioral analysis, rank events and build patterns of behavior.

One way to rank events for conducting behavioral analysis is to determine the strength of the links between the event types, as well as between the event instances themselves. Thus, there are: (1) *the specific weights* of direct, indirect same type and indirect different type relations between types of events, given by the number of equivalent, unequal, and unequal varietal properties, respectively (in structural analysis); (2) *the relative weights* of the links between instances of events, determined by the ratio of the number of coinciding values of the properties to the corresponding specific weights. As a result of analyzing the selected time window within the analyzed log, a set of pairs of values will be generated: the relative weight and the time interval. It is assumed that the frequency analysis of the resulting sets between the types of events, as well as the use of rank correlation methods, will allow us to determine the cause-effect relations between the types of events and between specific instances of events.

It is also possible to use intelligent approaches to correlation. However, due to the probable presence of cycles in a directed graph of relations of types, the use of Bayesian networks will be significantly hampered. This fact is due to that the cyclical relations of the elements of a graph theoretically cannot be resolved:

- removing of unlikely cyclical relations may lead to distortion of the result of correlation and to omission of abnormal event groups;
- simplification is impossible due to usage of low-level and indivisible (elementary) events.

D. Requirements for input data

The proposed approach has a number of constraints on the input data. It is supposed that before detecting patterns of event types within the same model the format of events is normalized. Normalization of structures is mainly reflected in the following condition: the structure of one event type t must not have equivalent properties p_1 and p_2:

$$\forall \{p_1, p_2\} \in P^t : p_1 \neq p_2, \{p_1, p_2\} \in P^T, t \in T\ . \qquad (8)$$

This restriction is necessary to avoid looping on a single event in the course of using the proposed approach. However, we should adhere to the normalized (single-valued) format for records of properties of events of different types.

It should be noted that the initial data must also satisfy the necessary condition for the completeness of various types of events within the frame of the discussed model and sufficient condition for the completeness of the number of different types of events to perform the time-frequency and behavioral analyses. In addition, in connection with the sensitivity and binding of the proposed approach to real time, the timing properties of events in the same model should be synchronized. Thus, for the correct application of the proposed approach to the log, system, segment, or infrastructure, the timing parameters of the events must be synchronized within the system log, segment and infrastructure, respectively.

Also, the definition of the time window of the event log for performing the behavioral analysis should be made taking into account the requirement of representativeness of the sample. Thus, the size of the sample is proposed to be determined on the basis of: (1) a time-frequency analysis of input event types, the use of which is determined by the frequency of the execution of a number of processes and tasks in different systems; (2) the dynamics (frequency) of the changes in the values of equivalent, unequal same type and unequal different type properties, which is caused both by the accident and by the periodicity of the actions described in the logs.

IV. Experiments and Discussion

Within the performed research the security event log of OS MS Windows 8 of the office computer, not included in the local domain, was used as the source data.

The experimental dataset has the following characteristics:

- the log size was 4 GB (7 GB in the XML format);
- the duration of the log recording was 1 month;
- the processing time of the log was equal to 50 minutes;
- the number of log events was equal to 6 700 000;
- the number of identified types of events was 80 of 418 stated in the documentation [15] for the given version of this OS (the event types of the previous version was not taken into account; the number of instances of events of this version was no more than 20);
- the number of identified properties was equal to 158, 14 of which were common (found in all types of events), 53 were unique (found only in one type of events); 89 were adjacent (occurred in more than one type of events).

As the result of the analysis of the log there was formed the graph of direct links of events types (Figure 2).

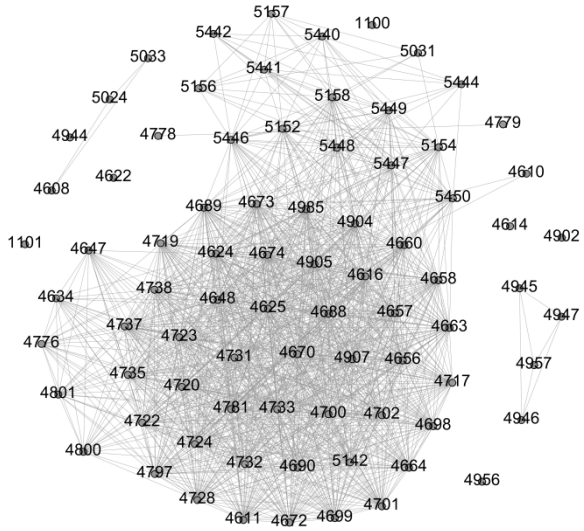

Figure 2. Example of the graph of events types connectivity

The presented figure shows that most of the identified events types have a great number of direct links. The entire graph contains 1309 nodes. There are also types of events that have no direct links to any other type.

Figure 3 shows a fragment of the graph of MS Windows 8 event types, taking into account the calculated specific weights of the links between them.

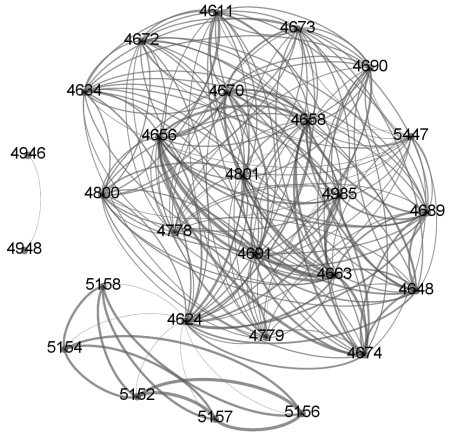

Figure 3. Example of graph of event types with the calculated specific weights of the relations between them

In the refined graph of event types, the strength of the link is displayed ин the thickness of the arcs.

Table 1 presents the most common properties of adjacent event types.

The proposed approach based on the analysis of event types is an integral part of a common methodology for correlation of security events. Generally, it may be applied for drawing up the framework of the model of the relationships (correlation) of events, which will be refined at each subsequent step of the suggested technique. The process of analysis of event types may be also regarded as the final stage of the normalization process.

TABLE I. ANALYSIS OF THE PROPERTIES OF EVENT TYPES

Property name	*Number of types*	*Property name*	*Number of types*
SubjectDomainName	28	ProviderKey	9
SubjectLogonId	28	ProviderName	9
SubjectUserName	28	ObjectServer	8
SubjectUserSid	28	TargetUserName	8
ProcessId	25	HandleId	7
ProcessName	16	TargetDomainName	7
LayerName	9	Application	6

In its turn, the events, which in the result of the technique became linked by particular properties of their types, will allow to generate events of higher levels and complex types that will contain the composite properties.

To ensure the work in real time with the necessary requirement for the completeness of the source data and considering the theoretically unlimited size of the analyzed infrastructure, the implementation of the techniques will require the application of technologies of parallel and multithreaded processing of big data. However, these technologies already exist, they continue to evolve and have been proved themselves on the positive side for different tasks. The developed technique will be mainly targeted at solving modern problems in the domain of security, such as monitoring security of cyber-physical infrastructures; detection of targeted attacks (based on the automated identification and pseudo-classification of anomalies), as well as automated security assessment for infrastructures of unlimited size.

Currently, the analysis of structures of different event types is used, for example, by the Splunk tool of analytical analysis of security data [16]. In this solution, the properties of the types of events are used to normalize the data after its download and subsequent indexing to perform processing tasks. In this case, direct links, obtained from the analysis of the source data, are used in processing requests, which, in their turn, are based on expert knowledge and manual adjustments. To obtain indirect links we should also conduct additional frequency analysis of similar content. In other words, in this solution direct links only imply their use, while in the proposed approach direct links are the foundation of model of relationships between events.

V. CONCLUSION

Conducted research in the domain of correlation of security events for SIEM systems showed the necessity of development of the correlation technique for solving existing problems in this domain. As a result of the evaluation of the overall correlation process and the formulation of the problem, the paper suggested an approach to correlation, based on analysis of event types. This approach is proposed to be applied at the initial stage of the common correlation technique for solving global research task. In the future we plan to continue development of the common correlation technique, based on the determination of functional relations between events and building the behavior patterns of the analyzed infrastructures.

ACKNOWLEDGMENT

This research is being supported by the grants of the RFBR (15-07-07451, 16-37-00338, 16-29-09482), partial support of budgetary subjects 0073-2015-0004 and 0073-2015-0007, and Grant 074-U01.

REFERENCES

[1] I. Kotenko and A. Chechulin, “A Cyber Attack Modeling and Impact Assessment Framework,” *Proceedings of 5th International Conference on Cyber Conflict 2013*. 2013. pp. 119–142.

[2] I. Kotenko, O. Polubelova, and Igor Saenko, “The Ontological Approach for SIEM Data Repository Implementation,” *2012 IEEE Intern. Conference on Green Computing and Communications, Conference on Internet of Things, and Conference on Cyber, Physical and Social Computing*. 2012. pp. 761–766.

[3] C. Kruegel, F. Valeur, G. Vigna, “Intrusion Detection and Correlation: Challenges and Solutions,” *Advances in Information Security*, Vol.14. Springer. 2005. 118 p.

[4] R. Sadoddin, A. Ghorbani, “Alert Correlation Survey: Framework and Techniques,” *Proceedings of 2006 International Conference on Privacy, Security and Trust: Bridge the Gap Between PST Technologies and Business Services*. 2006. Article no. 37.

[5] A. Hanemann, P. Marcu, “Algorithm Design and Application of Service-Oriented Event Correlation,” *Proceedings of Conference BDIM 2008, 3rd IEEE/IFIP International Workshop on Business-Driven IT Management*. 2008. pp. 61–70.

[6] T. Limmer, F. Dressler, “Survey of event correlation techniques for attack detection in early warning systems,” *Tech report. University of Erlangen, Dept. of Computer Science,* 2008. 37 p.

[7] A. Muller, “Event Correlation Engine,” *Master`s Thesis. Swiss Federal Institute of Technology Zurich*. 2009. 165 p.

[8] P. Ning, D. Xu, “Correlation analysis of intrusion alerts,” *Intrusion Detection Systems: series Advances in Information Security*, Vol. 38. Springer. 2008. pp. 65–92.

[9] A.A. Ghorbani, W. Lu, M. Tavallaee, “Network Intrusion Detection and Prevention,” Springer. 2010. 224 p.

[10] M. A. Hasan, “Conceptual framework for network management event correlation and filtering systems,” *Proceedings of the Sixth IFIP/IEEE International Symposium on Integrated Network Management*, 1999. pp. 233–246.

[11] U. Zurutuza, R. Uribeetxeberria, “Intrusion Detection Alarm Correlation: A Survey,” *Proceedings of IADAT International Conference on Telecommunications and computer Networks,* 2004. pp. 1–3.

[12] D.W. Guerer, I. Khan, R. Ogler, R. Keffer, “An artificial intelligence approach to network fault management,” SRI International. 1996. 10 p.

[13] M. Tiffany, “A survey of event correlation techniques and related topics,” [Internet resource]. – Access link: http://www.tiffman.com/netman/netman.html.

[14] H.T. Elshoush, I.M. Osman, “Alert correlation in collaborative intelligent intrusion detection systems — A survey,” Applied Soft Computing. 2011. pp. 4349–4365.

[15] Windows Security Log Events. [Internet resource]. – Access link: https://www.ultimatewindowssecurity.com/securitylog/encyclopedia/Default.aspx.

[16] Splunk Security. [Internet resource]. – Access link: https://www.splunk.com/en_us/solutions/solution-areas/security-and-fraud.html.