Approaches to find vulnerabilities and security in the digital production networks

Denis Parfenov, Lubov Zabrodina, Vadim Torchin and Anton Parfenov
Orenburg State University, Orenburg 460018, Russia

E-mail: parfenovdi@mail.ru

Abstract. Currently, the problem of ensuring cybersecurity of the network and information infrastructure of critical facilities is relevant. A special place in this is security automated control systems, which are the basis of industry 4.0. At the same time, the main source of threats to such systems is the external network of the telecommunications service provider, through which the operation of remote objects and systems is monitored. In the framework of this study, an approach was developed to search for abnormal behavior in the networks of telecommunication service providers, based on an analysis of events in the logs of various systems, including those responsible for network security. The proposed approach is the development of an information and event management system.

1. Introduction

Today, the use of computer networks is an integral part of everyday life. However, modern networks not only transmit data between users but also ensure the functioning of critical information infrastructure (CII) objects. At the same time, transit infrastructure is formed by telecommunication service providers, which are a key link in the chain of information interaction. It is their responsibility to ensure the required level of service (SLA), quality of service (QoS) and network security. One of the major global problems affecting the functioning of communication networks is the destructive impact exerted by various violators. If at the beginning of the development of networks the ultimate goal of attacks was ordinary users, today users, and their devices that have access to the network act as tools for conducting attacks on objects of CII. In 2018 alone, 4.3 billion cyberattacks on critical information infrastructure facilities in Russia were detected. In recent years, with the development of digital systems and the transition to industry 4.0, the attack vector itself has changed. If earlier the main share of the attacked objects was banking information systems, then the trend of recent years has been attacks on industrial facilities, namely their automated control systems (ACS). Failure or violation of the set algorithms for the operation of such systems can entail not only heavy financial losses but also man-caused and other negative consequences.

To solve these problems, telecommunication service providers to protect end-users and CII objects from external and internal threats are actively introducing solutions to ensure network security. Today, in this area, there are a lot of technical solutions. However, the basis of any software package is an intrusion detection system (IDS), for the construction of which SIEM technologies (Security Information and Event Management) are used [1]. One of the main tasks of this component is the collection and aggregation of monitoring data for devices and services included in the area of responsibility of a telecommunications service provider. Another equally important task is to analyze
the information received and to identify security incidents or detect abnormal behavior. The work of such security systems, as a rule, is organized according to a proactive principle - the response to incidents is carried out before the situation becomes critical. Therefore, decision-making time is critical for such systems. This is especially true because of the constantly changing attack vectors and methods for their detection. Nevertheless, an analysis of existing solutions shows that most systems today operate in a semi-automated mode. The final decision remains with the network security administrator, who monitors and monitors the operation of this system. Moreover, to make decisions, the network security administrator needs to analyze a lot of heterogeneous factors in a short time. This fact negatively affects the results of the reaction to ongoing attacks. Mistakes made can affect the work of an automated control system almost worse than the attack itself.

Therefore, the construction of intelligent protection and security systems for critical information infrastructure facilities is an urgent area.

2. Related work

A lot of research has been devoted to the development of methods and approaches to detect and detect attacks, including data transmission networks and ACS systems.

Colbert, E. J. M. et al reviewed and systematized the methods and approaches to organizing security systems used for Intrusion Detection in Industrial Control Systems. The authors note the three most effective approaches used when using Network-Based Intrusion Detection Systems. These include the Critical Process Values search method, the Network Packet Reporting Values scanning method, and the Network Traffic Pattern Anomalies method. Currently, methods based on data mining are gaining more and more popularity [2].

The work of researchers from the Peter the Great St. Petersburg Polytechnic University solves the problem of detecting attacks on the data transmission backbone networks. The authors propose a prototype network traffic analysis module that allows you to combine the data received from the traffic stream into time series and carry out further mathematical analysis. The proposed module is based on the hierarchical principle of data aggregation, which significantly reduces the time for data analysis [3].

In work of Khalili A. et. al. the authors of the study developed an application for detecting SysDetect attacks. The proposed approach for determining critical changes in the state of the system uses an iterative data mining algorithm, i.e. Apriori. This allows you to accurately identify all the states of the system [4].

In the study Koh P. et al., the authors search for attacks on SCADA systems by analyzing traffic in an open and closed network at various time intervals. In this case, the main task is to search for anomalies in the behavior of the system by assessing the self-similarity of the SCADA system traffic [5].

An important component for developing an effective system for detecting intrusions on critical infrastructure is data sets characterizing various types of attacks. The authors of the study in their work conduct a deep comparative analysis of the various six data sets obtained by monitoring the operation of Industrial control systems at various levels. Nevertheless, the presented kits do not cover a complete list of threats and, as a result, do not allow to fully train the intelligent intrusion detection system [6].

Modern production systems in their work use a variety of sensors that collect information about the condition of the equipment and the ongoing technological processes. In the study of Morales L. V. V. et al., the authors propose a methodology to develop a systematic approach to analyzing a data set to detect traffic anomalies in the network of such IoTs [7].

In the study of Khalili A. et al. explores attacks on cyberphysical systems (CPS). The authors proposed the SIDS method, aimed at detecting anomalies in the behavior of CPS. The proposed method is based on the analysis of CPS states and their frequency of change. In a study, the authors note that the SIDS method can effectively detect cyberattacks on large I / O CPSs [8].

A different approach to searching for vulnerabilities in critical infrastructure networks is considered in the study. The authors propose using a fairly popular method based on the use of the Honeynet system. The obtained data can be used to enrich the data sets used in training the intelligent attack detection system [9].
Thus, a review of research in the field of security of critical information infrastructure showed that at present there is no comprehensive solution allowing detecting attacks at an early stage due to the lack of effective methods for their detection.

3. Problem formulation
To effectively identify attacks, it is necessary to collect data on the state of vulnerable components and build a typical profile of their behavior. In the framework of this study, we consider the statement of the problem of monitoring the flow of events and security incidents in a network of telecommunication service providers. In general, the task of network monitoring can be described as follows. Since the events occurring in the network of telecommunication service providers are random in nature, mathematically probabilistic models of queuing theory — the theory of Markov chains — are most suitable for their analysis. Most often, the network model of a telecommunications service provider is a tree topology with additional redundant links. An integral component of the network of any telecom operator is an event monitoring system that collects information about the status of devices and the circulating traffic flows. Most often, one of the following Zabbix / Nagios / Cacti systems is used as a monitoring service. The monitoring system tasks include not only collecting data but also its preliminary processing, as well as launching alert scripts about the triggering of certain types of events. Also, under current law, providers are required to implement a hardware and software system on their network to collect, accumulate and store information about subscribers of telecom operators, as well as provide storage of users' passing traffic in a specialized data warehouse. Thus, these systems form a single information platform whose main purpose is to search and identify network attacks.

4. Model of the security network monitoring system
At the initial stage of modeling, it is necessary to determine the input data. As states of the network monitoring system under consideration, it is proposed to consider events that occur in the provider's network in terms of traffic passing through telecommunication equipment. Network events are stored in the event log. Consider the source data, namely, the list of records passing through the network traffic.

Within the framework of the study, among the many characteristics that describe network connections, such characteristics as the IP addresses of the sender and receiver, their corresponding ports and the protocol over which data is transmitted were selected. When choosing these characteristics, the general view of the record can be represented:

\[
< \text{dt}, \text{src}_\text{ip}, \text{dst}_\text{ip}, \text{src}_\text{port}, \text{dst}_\text{port}, \text{proto}, s >
\]

(1)

where \( \text{dt} \) – package date and time (\( x_1 \)); \( \text{src}_\text{ip} \) – source IP-address of the packet (\( x_2 \in [0;2^{32} - 1] \)); \( \text{dst}_\text{ip} \) – destination IP-address of the packet (\( x_3 \in [0;2^{32} - 1] \)); \( \text{src}_\text{port} \) – source port (\( x_4 \in [0;65535] \)); \( \text{dst}_\text{port} \) – destination port (\( x_5 \in [0;65535] \)); \( \text{proto} \) – network connection protocol (\( x_6 \)); \( s \) – package size (\( x_7 \)).

We represent each such line as a vector

\[
x_k = \{ x_{k1}, x_{k2}, x_{k3}, x_{k4}, x_{k5}, x_{k6}, x_{k7} \}
\]

(2)

The list of package entries is presented in the form of a set of the following form:

\[
X = \{ x_k \}, \quad k = 1, n
\]

(3)

where \( n \) – log list length.

In order to analyze the state of nodes in the process of transmitting data on the network, a Zabbix server is installed, which provides data from \( Z \) – the monitoring system about network devices in the following format:

\[
< \text{CPU}, \text{ping}, \text{ram}, \text{cap}_\text{av}, \text{port}_\text{s}, \text{port}_\text{pln} >,
\]

(4)
where CPU - CPU load % (\(z^i_{1}\)); ping - response time (\(z^i_{2}\)); ram - load memory % (\(z^i_{3}\)); s_cap - channel bandwidth Mbps (\(z^i_{4}\)); av_port - port availability (true / false) (\(z^i_{5}\)); s_port - port speed Mbps (\(z^i_{6}\)); pln - packet loss (\(z^i_{7}\)).

Then the list of records from the monitoring system about network devices can be represented as a set of the following form:

\[
Z^i = \{z^i_k\}, \quad k = \overline{1,m}
\]

\[
z^i_k = \{z^i_{21}, z^i_{22}, z^i_{23}, z^i_{24}, z^i_{25}, z^i_{26}, z^i_{27}\}
\]

where \(m\) - number of network devices.

Data \(Z^i\) from the monitoring system on the state of the information system that supports the operation of critical objects have a different format:

\[
<\text{time\_work}, \text{CPU}, \text{storage}, \text{stat\_in}>,
\]

where \(\text{time\_work}\) - execution time ms. (\(z^i_{1}\)); \(\text{storage}\) - load ram % (\(z^i_{2}\)); \(\text{CPU}\) - load CPU % (\(z^i_{3}\)); \(\text{stat\_in}\) - statistics of entry to the IP from IP addresses (contains records of the form: (IP address, average daily number of entries)).

The list of system status records in this case is:

\[
Z^i = \{z^i_k\}, \quad k = \overline{1,p}
\]

\[
z^i_k = \{z^i_{k1}, z^i_{k2}, z^i_{k3}, z^i_{k4}\}
\]

where \(p\) - number of systems.

To get a real opportunity to operate with available information for incident investigation, we introduce a data structure of the following form:

\[
<\text{id, name, adj\_matrix, id\_ch, cond, pid}>,
\]

where \(\text{id}\) - subsystem identifier entered to be able to refer to the subsystem; \(\text{name}\) - subsystem name; \(\text{adj\_matrix}\) - adjacency matrix of elements of a system / subsystem; \(\text{id\_ch}\) - vector of identifiers of the components of the system / subsystem, if the node is considered, the adjacency matrix is not turned on, and this parameter will be the ip address; \(\text{cond}\) - vector of characteristics describing the current state of the subsystem; \(\text{pid}\) - subsystem ancestor identifier.

Based on the constructed model, within the framework of this study, an approach has been developed that allows scaling data from two interconnected systems: monitoring network conditions and intrusion detection. To do this, we describe the many possible states: \(S_0\) – lack of malfunctions; \(S_1\) – network congestion; \(S_2\) – bandwidth reduction; \(S_3\) – port unavailability; \(S_4\) – physical device unavailability; \(S_5\) – packet fragmentation; \(S_6\) – complete system failure.

Note that each attack vector is characterized by its graph of transitions from the \(S_i\) state to the \(S_j\) state. Also, the state transition itself is characterized by a certain frequency and distributed over time. Another equally important characteristic is the set of events generated by the elements of the information system that is part of the critical information infrastructure object. This data set also allows you to characterize the behavior of the system at a given point in time. These and other characteristics make it possible to form a data set that allows one to construct a profile of the typical behavior of each component of the system and, as a result, to identify attacks on target systems.

The proposed solution allows you to detect abnormal behavior of critical information infrastructure objects and respond to them ahead of time.
5. Experimental results

As part of the study, a virtual stand was emulated that emulated the work of the provider's network, providing a connection to the ACS. For this, three network segments are deployed in an experimental network deployed based on the OpenNebula cloud platform. The first network segment consists of 6 computing nodes that emulate the operation of an ACS system. The second segment consists of 8 computing nodes that emulate the operation of the main components of the telecommunications service provider network, including security systems, intrusion detection, and incident data storage. The third segment includes 16 computing nodes that emulate the work of legitimate users and cybercriminals. For the purity of the experiments, the roles of nodes in the third network segment were randomly assigned. It is also worth noting that the second network segment is a transit link between the first and third links.

Apache Spark was chosen as the base system for organizing intelligent data processing. As a data warehouse used Cassandra. To generate attack mechanisms, public Datasets were selected as the initial data sets: KDD99 CUP, NSL-KDD, UNSW-NB15. To study the listed data sets, the following attack vectors were selected as target threats: Fuzzers, Analysis, Backdoors, DoS, Exploits, Generic, Reconnaissance, Shellcode and Worms. For classification according to the principle of attack-normal traffic, the MLLib machine learning library, which is part of Apache Spark, was used.

The computational experiment consisted of two stages. At the first stage, traffic flows were generated on the target ACS system according to the selected attack vector. At the same time, data on events occurring both in the network of the first segment and directly in the ACS were collected in the ACS system under study. The received data was aggregated by the monitoring system and loaded into the data warehouse for further processing. After normalizing and processing this data, an IS_LOG dataset was generated that characterizes each of the attack vectors. Based on the data obtained, hybrid data sets <name_dataset> + IS_LOG is generated.

At the second stage of the experimental study, each initial and hybrid data set is divided into two fragments: a test set and a training set. At the same time, the volume of each test set is 10% of the total data volume under investigation. Logistic regression was chosen as a classifier in the second stage of the experimental study, since this method is quite effective for large-scale tasks. Logistic regression is based on a statistical model used to predict the probability of events by fitting data to a logistic curve. To evaluate the results of the experiment, we apply the ROC analysis since it is based on True Positive Rate and False Positive Rate.

| Dataset               | Accuracy | TP      | TN      | FP      | FN     |
|-----------------------|----------|---------|---------|---------|--------|
| KDD99 CUP             | 92%      | 95%     | 92%     | 3.0%    | 6.0%   |
| KDD99 CUP + IS_LOG    | 94%      | 97.2%   | 93%     | 2.0%    | 5.8%   |
| NSL-KDD               | 95%      | 94%     | 92%     | 2.0%    | 5.7%   |
| NSL-KDD + IS_LOG      | 96.6%    | 95.1%   | 93.8%   | 1.8%    | 4.3%   |
| UNSW-NB15             | 96%      | 95%     | 94%     | 1.0%    | 4.3%   |
| UNSW-NB15 + IS_LOG    | 97.9%    | 97.51%  | 99.48%  | 0.51%   | 2.48%  |

Thus, the study showed that the proposed approach to identifying attacks based on a hybrid data set expanded by analyzing events in the logs of various information systems allows one to more accurately determine intrusions in the networks of telecommunication service providers.

6. Conclusion

During the study, the following tasks were solved:

1) The main security threats and types of attacks relevant to the critical information infrastructure facilities of industry 4.0 are considered.
2) A model for monitoring events and detecting network attacks based on data mining methods has been developed.

3) Analyzed network traffic data sets suitable for modeling the traffic of an industrial network of enterprises: KDD99 CUP, NSL-KDD, UNSW-NB15 for the task of detecting network attacks, and also proposed the extension of existing data sets by analyzing events in the logs of various information systems.

4) A computational experiment was conducted to build an attack classifier based on a hybrid data set based on logistic regression. The effectiveness of the proposed solution is proved.

Acknowledgment
The research work was funded by the Russian Foundation for Basic Research, according to the research project No. 18-07-01446, 18-37-00400, the President of the Russian Federation within the grant for state support of young Russian scientists (MK-860.2019.9) and with support of the Ministry of Education of the Orenburg region (project No. 12 on 14 August 2019).

References
[1] Bolodurina I, Parfenov D, Torchin V and Zabrodina L 2018 Development and research of model for ensuring reliability of operation of network infrastructure objects in cyber-physical system located in the cloud platform Proc. of the 2018 Global Smart Industry Conference (GloSIC) (IEEE) pp 1-7
[2] Colbert E and Hutchinson S 2016 Intrusion detection in industrial control systems Cyber-security of SCADA and other industrial control systems 66 209-37
[3] Poltavtseva M, Zegzhda P and Pankov I 2018 The hierarchial data aggregation method in backbone traffic streaming analyzing to ensure digital systems information security Proc. of the 2018 Eleventh International Conference” Management of large-scale system development”(MLSD) (IEEE) pp 1-5
[4] Khalili A and Ashkan S 2015 SysDetect: a systematic approach to critical state determination for industrial intrusion detection systems using apriori algorithm J. of Process Control 32 154-60
[5] Pauline K, Hwa-Jae C, Se-Ryong K, Hyuk-Min K and Huy Kang K 2012 Intrusion detection methodology for scada system environment based on traffic self-similarity property J. of the Korea Institute of Information Security and Cryptology 22 267-81
[6] Seungoh C, Jeong-Han Y and Sin-Kyu K 2018 A comparison of ICS datasets for security research based on attack paths In: Proc. of the International Conference on Critical Information Infrastructures Security (Springer, Cham) pp 154-66
[7] Morales L, Lopez-Vizcaino M and Iglesias D 2019 J. Anomaly detection in IoT: methods, techniques and tools J. Multidisciplinary Digital Publishing Institute Proceedings 21(1) 4
[8] Khalili A, Sami A and Pouresmaeeli S 2018 SIDS: State-based intrusion detection for stage-based cyber physical systems J. International Journal of Critical Infrastructure Protection 22 113-24
[9] Ullah I and Mahmoud Q 2017 A hybrid model for anomaly-based intrusion detection in SCADA networks In: Proc. of the 2017 IEEE International Conference on Big Data (Big Data) (IEEE) pp 2160-7