Preliminary Processing of Data From Information and Communication Systems for the Task of Preventing Cyberattacks

Mir-Khusan Kadirov[1], Zoxidjon Tulyaganov[1], Dilbar Karimova[1]
[1]*Tashkent State Technical University, 2 Universitet Street, Tashkent city, Uzbekistan*
mirhusank@rambler.ru, zoxidjontulyaganov@gmail.com, taqriz@mail.ru
https://doi.org/10.5281/zenodo.10471779

Keywords: cyberattack, data processing, critical systems, two-stage data aggregation, traffic, IoT, DdoS, Information security, Analysis, Forecasting, Attack detection, Prevention of cyber-attacks, TCP SYN attack.

Abstract: The article describes pre-processing of data from endpoints and network devices, aimed at reducing the amount of processed data. An approach to detecting distributed denial of service attacks in information and communication systems based on two-stage data aggregation is considered. The flow of control commands must be regularly monitored and controlled, since the control device is a desirable target for an attacker. Circulating sent traffic flows also need to be monitored and analysed to detect cyber-attacks on information and communication systems. Research has shown that in order to prevent computer attacks on information and communication systems, it is necessary to detect anomalies in the operation of system components as early as possible with high accuracy. Due to the speed and specificity, it is difficult to determine that a system is under attack, since conventional tools do not detect the type of denial of service threats. To ensure the security and stable operation of network resources, it is necessary to implement an approach based on detecting DDOS attacks in information and communication systems. Since distributed denial of service attacks have caused significant financial losses worldwide in recent years. This article is aimed at solving problems related to ensuring information security in information and communication systems and preventing DDOS attacks in them.

1 INTRODUCTION

Modern information and communication systems (ICS) are at the stage of active development. The development of ICS has a significant impact on the transformation of all branches of human activity, including critical ones [1].

The energy industry is actively developing both in Uzbekistan and abroad. One of the priority areas for the development and digitalization of the energy sector is the creation of intelligent energy supply networks Smart Grid - a new type of cyber-physical systems that implement distributed energy exchange using “smart” devices [2].

In recent years, the vast majority of industrialized countries have begun to implement programs and projects in the direction of Smart Grid, covering a wide range of problems and tasks. The most large-scale programs and projects have been developed and are being implemented in the USA, Canada and EU countries, as well as in China, South Korea and Japan.

Among the ICS of the critical transport industry, one can distinguish “smart” airports equipped with a large number of intelligent devices. Among them is the international airport in Dubai, equipped by HDL, a global manufacturer of Smart Home automation systems, with a large number of smart devices. HDL has completed the integration of airport building automation systems, in particular, lighting automation equipment, video surveillance control systems, baggage handling and conference room management in several terminals [3].

Among the ICS of the critical transport industry, it is also worth noting intelligent traffic control systems that receive information from cars, smart traffic lights and various sensors that roads are equipped with.

The specificity of modern ICS, from the point of view of information security, lies in the following factors:

1. A large number of smart devices that are part of modern ICS have vulnerabilities that allow attackers

to use such devices as points of illegitimate penetration into ICS.

2. The heterogeneity and large volume of data generated by smart devices, along with network devices, servers and personal computers, greatly complicate and slow down the process of security analysis. On large amounts of data, anomalies caused by low-intensity and short-term computer attacks also become invisible.

3. ICS have a communication environment, which, together with the fact that the mutual control of the components with each other is implemented by exchanging data, allows us to assume that data flows from the ICS components will not always be controlled by a person. This entails security problems associated with hidden illegitimate influence on the operation of the system.

4. ICS implement irreversible physical processes along with reversible informational ones. The information processes of such systems are aimed at ensuring the correct functioning of the ICS network infrastructure, which forms the conditions for the flow of physical processes. Even a slight exit of the physical process out of control, achieved through a destructive information impact on the parameters of the network infrastructure, can lead to catastrophic consequences. For ICS, of particular importance is the early detection of computer attacks, which makes it possible to obtain a certain margin of time for a reaction that counteracts the violation of the correctness of the course of physical processes.

The problem of detecting computer attacks is caused by the active digitalization of critical infrastructure components, their integration with smart devices (Internet of things devices). As a rule, manufacturers of such devices are focused on large-scale sales of their products, which requires a fast pace of production. Under such conditions, the level of security of the created devices cannot be high, since thorough testing and security checks require time and the involvement of specialists. A similar situation is observed with the release of updates to device firmware - updates are released quite rarely, and a significant part of security problems remain unresolved.

The relevance of the task of preventing computer attacks, especially in terms of recognizing attacks, is emphasized by the trend observed over the past few years to increase the number of computer attacks on industrial and critical infrastructure. Along with the digitalization of technological infrastructure, there is an explosive growth in the number of cyber-attacks on critical infrastructure objects represented by automated and cyber-physical systems all over the world.

2 METHODS AND MATERIALS

Currently, the most common way to implement attacks on end devices, whether it is a user's personal computer or sensors and controllers in the Internet of Things system, are network attacks [4]. With the help of network attacks, it is possible to: exploit vulnerabilities in the device software, remote control through embedded malicious software, create botnet networks, as well as disable nodes associated with critical areas of human activity, which can lead to serious consequences.

Distributed denial-of-service (DDOS) attacks have led to significant financial losses worldwide [5]. The results presented are consistent with the growing number of devices connected to the Internet. This growth is driven by the rise of the Internet of Things (IoT) and is characterized by the concept of connecting any device, anywhere, anytime. One of the most dangerous malicious traffic on the Internet is high-load DDOS attacks.

Existing approaches to detecting DDOS attacks are primarily aimed at detecting malicious network traffic in lightly loaded networks, and to a small extent cover vulnerabilities caused by network loads [6]. A limited number of studies offer a solution to the problem of detecting and preventing DDOS attacks in large-scale networks based on strict filtering of network packets, the use of various security policies and methods of mathematical statistics [7-9].

To prevent computer attacks on ICS, it is necessary to detect anomalies in the operation of system components as early as possible with high accuracy.

The solution of the monitoring problem is complicated by a number of aspects typical for modern ICS:

1. A large amount of data circulating in the system. A large number of intelligent components that perform measurements and control leads to intensive data generation in the system. Not all of this data is necessary for security analysis, however, in order to leave only the necessary data, it is necessary to process the entire amount of information.

2. High data heterogeneity. Many components of the system can be produced by different vendors and, therefore, have a different format, which also complicates the information processing process. In this case, it is necessary to analyze both network traffic and data from end devices.

3. The need for operational data analysis. To solve the problem of preventing computer attacks, it is necessary to carry out the processes of detecting anomalies, searching for a self-regulation algorithm and applying it to the current configuration of the ICS in a time less than the time for the attack to spread and achieve its goal.

4. The need for periodic use of data on previously implemented computer attacks on the system. A system that does not update its protection scenarios and does not take into account the features of previously successfully implemented computer attacks does not have the property of cyber stability. To ensure this property, it is advisable to use data on previous attacks.

2.1 Setup Two-stage data aggregation

To solve the above problems, it is proposed to implement an approach consisting in a two-stage data aggregation separated by a normalization stage [10].

Aggregation is the process of combining components into a single whole, therefore, it will reduce the amount of data being analyzed. Aggregation is proposed to be performed in two stages: aggregation by time; aggregation by type of ICS component.

Time aggregation is performed for each component of the ICS separately, it consists in taking for further data processing not several parameter values that arrive over a short time interval, but one aggregated value for a larger time interval, consisting of several small time intervals [11].

An important aspect in the implementation of stage 1 of aggregation is the choice of the time interval - if the period is chosen too large, there is a risk of missing an anomaly in the data values from the device. If the period is chosen too small, the amount of data will not be significantly reduced, and the data processing will take a long time.

It is proposed to introduce the parameter φ, which is a significant period of the functioning of the ICS component associated with the physical process in which this component participates. Let's define the set of physical processes running in the system: $FP_r = \{fpr_1, fpr_2, \dots, fpr_k\}$.. Let us assign to each physical process fpr_i its period ψ_i:

$$\forall\, fpr_i \in FP \exists Period: fpr_i \to \psi_i, i = \overline{1,K}. \quad (1)$$

Then for each component of the ICS, modeled on the graph by the vertex v_j and involved in the physical process fpr_i, the aggregation period φ_j must be strictly less than the period of the physical process ψ_i:

$$\forall\, v_j, j = \overline{1,N}, v_j \in fpr_i\ \varphi_j < \psi_i \quad (2)$$

The potential loss of data is important because before the aggregated value is written to the database, the processing is done in memory, so this information can be lost in the event of a sudden failure.

If necessary, depending on the scale of the ICS, the amount of data and the possibilities for storing information, several more levels of aggregation can be performed on the aggregated values of the parameters, when one is formed from several aggregated values. The number of such levels may vary, but the total aggregation period must be less than the period of the physical process.

For each level of aggregation and set of aggregated values, a certain time of their life in the system is assigned. This time is already determined by the structure and speed of the system for preventing computer attacks, since it is directly related to the speed of performing analytical operations [12]. After the expiration of the lifetime, information from each level of aggregation must be deleted in order to make room for new information coming from the ICS components.

Then we denote the aggregation period as $Pagg$ and compare it with the timestamps Δt, the time interval for accumulating data that will be aggregated in the future, and $tlife-$ the lifetime of the aggregated data in the system: $Pagg(\Delta t, tlife)$. The value of Δt can be described as the minimum of the values of the allowable failure time and the aggregation time without losing the quality of the ICS and the accuracy of the analysis.

The stages of aggregation, as already noted, are separated by the stage of normalization - the reduction of various values obtained from the components of the ICS to a single format. This stage is necessary for further work with the data set, in addition, it is preparatory for stage 2 of aggregation - according to the type of ICS component.

Normalization is implemented in several steps:

1) processing of messages received from the end devices of the ICS;

2) normalization of measurement values from end devices;

3) assignment of metadata.

The first step is to process messages from end devices.

ICS in order to extract the measured values - it is from them that time series will be formed in the future, which are the object of methods for detecting computer attacks.

The following steps are taken to process messages:

1) determination of the ICS component from which messages were received;

2) formation of a list of possible data presentation formats for this component;

3) definition of the message format.

After recognizing message formats from all end devices, it is necessary to choose which of the formats

is prevailing and normalize the parameters extracted from messages in accordance with this format [13]. This will reduce the time and computational costs of data preprocessing. The scheme of the normalization stage is shown in Figure 1.

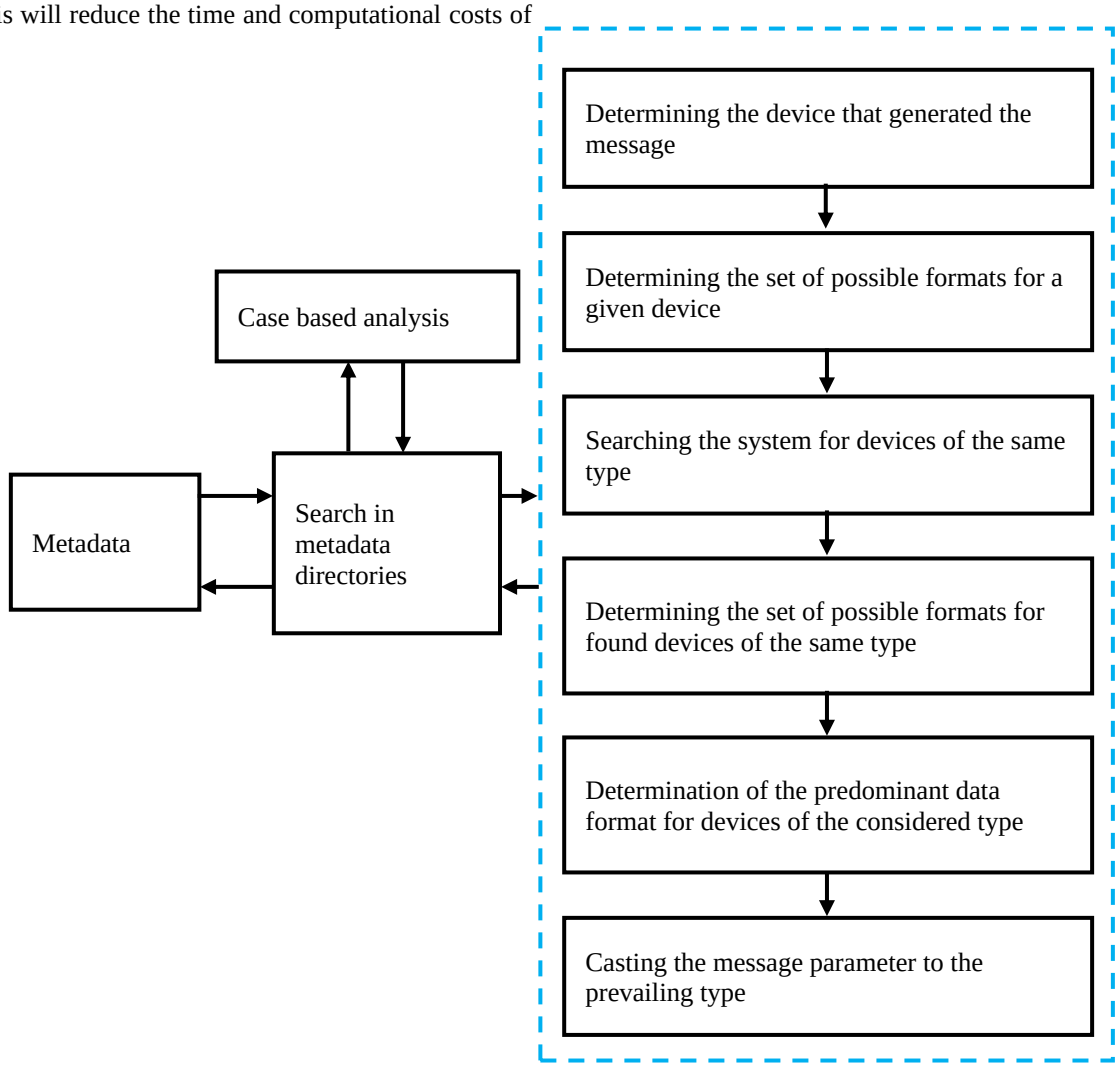

Figure 1: Scheme of the normalization stage.

Stage 2 of aggregation is performed according to the type of ICS component, this aggregation stage can not always be performed and not in all systems, since the condition for its implementation is the presence of devices of the same type that measure the same parameters of the environment or physical process and are located in close proximity to each other. friend. For different ICS, proximity can be expressed by different distances.

This approach is relevant for industrial production systems, for energy distribution systems, since in industrial systems it is possible to single out production workshops, in which a certain number of sensors of the same type are located, measuring, for example, the temperature in the workshop, and for energy distribution systems - close energy consumers. Then the values obtained from several components of the same type can be aggregated into one value of a certain generalizing device [14]. Let's denote this value as $value_a$, then:

$$value_a = f(Value\{value_i | i = 1, \dots, n\}), \quad (3)$$

Where:

1) $Value$ is the set of values of the parameter of the aggregated components of the ICS;

2) $value_i$ is the parameter received from the i −th aggregated component of the ICS;

3) n is the number of aggregated ICS components.

A feature of this stage of aggregation is the ability to process asynchronous messages. The scheme of the stage is shown in Figure 2.

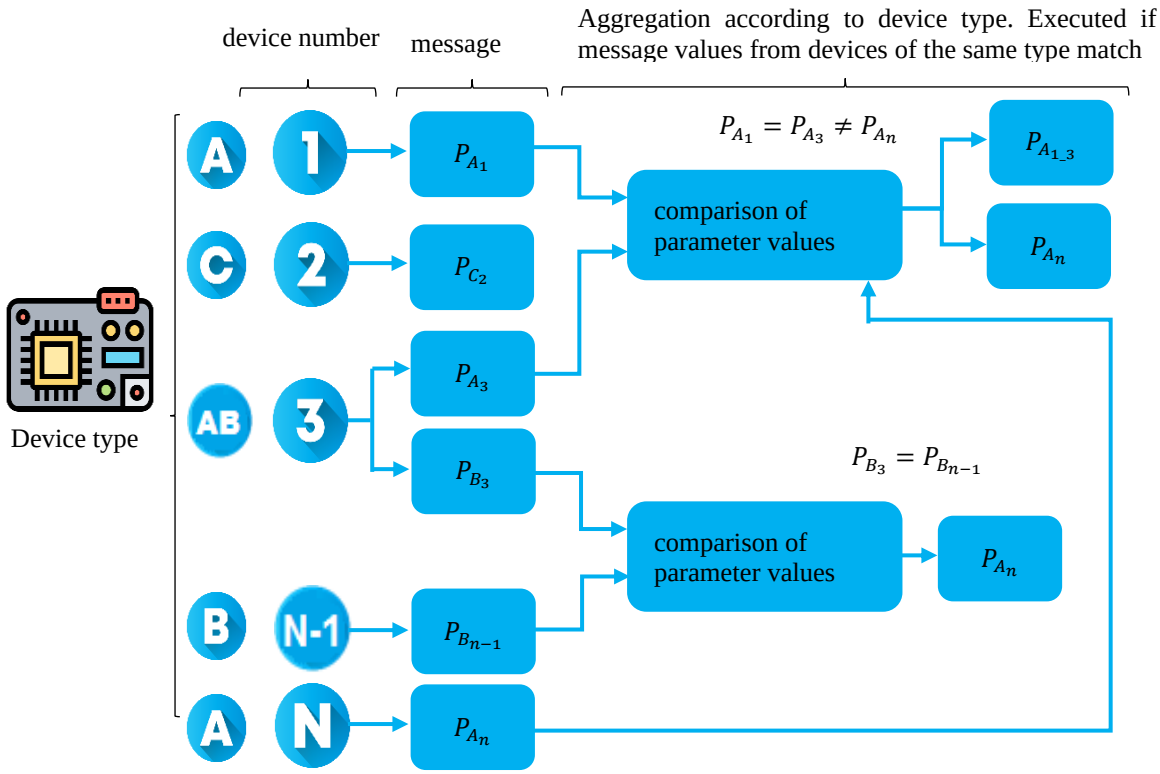

Figure 2: Scheme of aggregation by type of ICS component.

Restrictions on the use of this stage of aggregation:

1) the absence of closely spaced devices that measure the same parameters;

2) a significant difference in the measured values of the parameter.

The difference in the measured values, greater than some given $\Delta value$, determined separately for each ICS, may indicate a possible anomaly caused by a breakdown of one of the devices or a computer attack on it. Also, the difference can be associated with the feature of the subject area of the ICS, or rather, with the nature of the environment in which the measurement devices are located.

This type of aggregation makes it possible to further reduce the amount of data to be processed and, therefore, significantly speed up the process of further security analysis.

3 RESULTS AND DISCUSSION

The most promising in terms of accuracy in detecting DDOS attacks are hybrid methods based on the use of machine learning algorithms. The effectiveness of DDOS attack detection methods based on a hybrid approach is determined by a combination of heuristic parameterization of input data, the use of statistical analysis for their processing, and machine learning methods. An approach that can be used to detect anomalies is to split the source traffic into non-overlapping time windows.

Bursts become easier to track if time intervals are calculated to overlap each other. Thus, as a result of preprocessing, the researcher receives a data set composed of windows of the following type:

$$W_{i+x*tof} = \begin{cases} \text{Time series for parameter}_1 \\ \text{Time series for parameter}_2 \\ \text{Time series for parameter}_3 \end{cases} \quad (4)$$

where $w -$ window, $i -$ is the window number, $tof -$ is the time offset, $x -$ is the offset number. Figure 3 demonstrates more clearly the overlay of windows on top of each other.

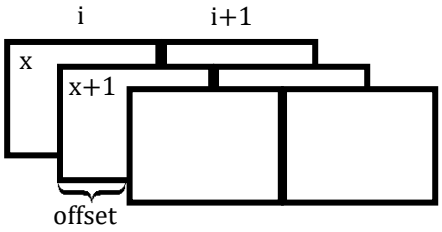

Figure 3 – Example of intersection of time windows

Thus, if the time interval is selected at 5 sec, and the offset is 2.5 sec, then the following windows will be compiled:

$$[[0sec-5sec],[2.5sec-7sec]],$$
$$[[5sec-10sec],[7sec-12sec]],$$
$$\ldots$$
$$[[n*5sec-(n+1)*5sec],[n*5+2.5sec-(n+1)*5+2.5sec]], \quad (5)$$
$$\ldots$$

In this case, the data grouping should be performed by the first interval, since it is in relation to this offset that the dynamics of changes in the composition of the sent traffic are considered.

In order to characterize the behavior of sent traffic in a given time window, it is proposed to monitor the degree of dependence of various packet parameters. As a metric to determine this degree of dependence, the multiple correlation coefficient [15] can be used, which for sequences x, y, z is calculated using the following formula:

$$R_{y(x,z)} = \sqrt{\frac{r_{xy}^2+r_{zy}^2-2*r_{xy}*r_{zy}*r_{xz}}{1-r_{xz}^2}} \quad (6)$$

where rxy, rzy, rxz are pair correlation coefficients, which are calculated as follows:

$$r_{xy} = \frac{\sum(x_i-<x>)*(y_i-<y>)}{\sqrt{\sum(x_i-<x>)^2*\sum(y_i-<y>)^2}} \quad (7)$$

The deviation of the multiple correlation coefficient $R_{y(x,z)}$ from the threshold value indicates the presence of an anomaly in the sent traffic.

Figure 4 shows a graph comparing multiple correlation coefficients at the junction of sent traffic without an attack with a TCP SYN attack [16].

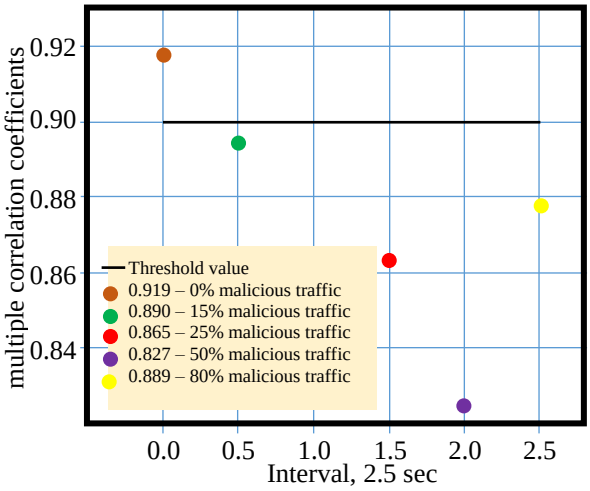

Figure 4 – Comparison of multiple correlation coefficients for different percentages of traffic with and without TCP SYN attack.

The figure shows 5 time intervals containing different percentages of malicious network traffic representing a TCP SYN attack and harmless sent traffic. Also highlighted in the figure is the threshold value that separates traffic without an attack above the line from traffic with an attack below the line. The threshold value of 0.9 was obtained empirically as a result of a series of experiments.

4 CONCLUSION

To detect cyber-attacks, it is necessary to ensure efficient processing of large volumes of heterogeneous data of both the physical and information infrastructure of the system. An approach to the pre-processing of physical infrastructure data, in particular, end devices, is described, based on two stages of data aggregation separated by a normalization stage. The approach provides a reduction in the dimension of the processed data and the reduction of data from various components of the ICS to a single form.

The results were obtained based on the approach of dividing the original sent traffic into non-overlapping time windows. Thus, this study demonstrates the ability to detect DDOS attacks that make up at least 15% of the analysed time interval of sent traffic.

REFERENCES

[1] Tvaronavičienė M. et al. Cyber security management of critical energy infrastructure in national cybersecurity strategies: Cases of USA, UK, France, Estonia and Lithuania //Insights into regional development. – 2020, pp. 802–813.

[2] Pandey R. K., Misra M. Cyber security threats—Smart grid infrastructure //2016 National power systems conference (NPSC). – IEEE, 2016, pp. 1–6.

[3] Salam A. O. A. Automation and control of DIA transportation tunnel //2007 Mediterranean Conference on Control & Automation. – IEEE, 2007, pp. 1–7.

[4] Zhao K., Ge L. A survey on the internet of things security //2013 Ninth international conference on computational intelligence and security. – IEEE, 2013, pp. 663–667.

[5] Singh A., Gupta B. B. Distributed denial-of-service (DDoS) attacks and defense mechanisms in various web-enabled computing platforms: issues, challenges, and future research directions //International Journal on Semantic Web and Information Systems (IJSWIS). – 2022. – T. 18. – №. 1. pp. 1–43.

[6] Javaheri D. et al. Fuzzy logic-based DDoS attacks and network traffic anomaly detection methods: Classification, overview, and future perspectives //Information Sciences. – 2023.

[7] Malikovich K. M. et al. Differentiated Services Code Point (DSCP) Traffic Filtering Method to Prevent Attacks //2021 International Conference on Information Science and Communications Technologies (ICISCT). – IEEE, 2021. – pp. 1-4.

[8] G. S. Rajaboevich, K. M. -X. Mirpulatovich and A. J. Tileubaevna, "Method for implementing traffic filtering in SDN networks," 2022 International Conference on Information Science and Communications Technologies (ICISCT), Tashkent, Uzbekistan, 2022, pp. 1-3, doi: 10.1109/ICISCT55600.2022.10146873.

[9] Hwoij A., Khamaiseh A., Ababneh M. SIEM architecture for the Internet of Things and smart city //International Conference on Data Science, E-learning and Information Systems 2021. – 2021. pp. 147-152.

[10] Adrienko N., Adrienko G. Spatial generalization and aggregation of massive movement data //IEEE Transactions on visualization and computer graphics. – 2010. – T. 17. – №. 2. pp. 205–219.

[11] Sagatov M., Irgasheva D., Mirhusan K. Construction Hardware Protection Infocommunication Systems from Network Attacks //Proceedings of International Conference on Application of Information and Communication Technology and Statistics in Economy and Education (ICAICTSEE). – International Conference on Application of Information and Communication Technology and Statistics and Economy and Education (ICAICTSEE), 2015. – pp 271.

[12] Lawrence, R. The space efficiency of XML/R. Lawrence // Information and Software Technology. – 2004. – Vol. 46. № 11. pp. 753–759.

[13] Miloslavskaya N., Tolstoy A. New SIEM system for the internet of things //New Knowledge in Information Systems and Technologies: Volume 2. – Springer International Publishing, 2019. pp. 317–327.

[14] Kotenko I., Chechulin A. Attack modeling and security evaluation in SIEM systems //International Transactions on Systems Science and Applications. – 2012. – T. 8. pp. 129–147.

[15] Kubokawa, Tatsuya, Éric Marchand, and William E. Strawderman. "A unified approach to estimation of noncentrality parameters, the multiple correlation coefficient, and mixture models." Mathematical Methods of Statistics 26 (2017): 134-148.

[16] Berguiga, Abdelwahed, and Ahlem Harchay. "An IoT-Based Intrusion Detection System Approach for TCP SYN Attacks." Computers, Materials & Continua 71.2 (2022).