# Semi-structured information in the field of artificial intelligence and information security: processing results

# Semi-structured information in the field of artificial intelligence and information security: processing results

**N.I. Yusupova[1], O.N. Smetanina[1], M.M. Gayanova[1], N.P. Komendantova[2]**

[1]Ufa State Aviation Technical University, Ufa, K. Marks str., 12, 450000, Russia

[2]International Institute for Applied Systems Analysis, Laxenburg, Schlossplatz, 1, A-2361, Austria

E-mail maya.gayanova@gmail.com

**Abstract**. This article is devoted to the semantic analysis of weakly structured information in the field of "Artificial intelligence and information security". The methodology of this research included two stages and is based on the meta-analysis of existing studies. The received results allow development of further methodological recommendations on semi-structured information and artificial intelligence.

**Keywords:** Semi-structured information, Artificial Intelligence, Information Security, Semantic analysis.

## 1. Introduction

The processing of semi-structured information is an urgent task, many scientists are engaged in it at present day. There are no general solutions in this area, the authors make specific decisions depending on the subject area. This article discusses some of the results of semantic analysis of semi-structured information in the field of knowledge "Artificial Intelligence and Information Security".

## 2. Research methodology

The research includes two stages. The first stage is a preliminary analysis of the information of the selected subject area based on data from a known platform, the purpose of which is to highlight the most significant publications. The second stage is a semantic analysis of the most significant publications to highlight the most interesting scientific results.

The scientific electronic library eLIBRARY.ru was the source for empirical data collection [1]. eLIBRARY.ru is a Russian scientific electronic library integrated with the Russian Science Citation Index (RSCI), which currently contains over 34 million articles. The API integrated into eLIBRARY.ru allows us to get a list of publications and metadata about publications.

On the main page of eLIBRARY.ru there is a built-in search function and an advanced search in the library, where a specific query is entered. Further, the system asks where to look for information (in the title of the publication, in the annotation, in keywords, in the names of the authors' organizations, in the lists of cited literature and in the full texts of publications) and in what types of publications (articles in journals, books, conference materials, deposited manuscripts, dissertations, reports and patents).

We can also specify search topics, specific authors or journals, which will refine the search result. Then we can set the search parameters: search based on morphology, search for similar text, search in

publications that have full text on eLIBRARY.ru, etc. Further, we can set a specific year or an interval by years for the search and sorting type: by years or by relevance.

The system generates a list of publications for the request, which can be saved to a collection with the appropriate name. In personal selections, we can analyze publications according to various parameters, which are quantitative characteristics of publications that are issued by the system: the total number of publications, the number of articles in journals included in the Web of Science or Scopus citation bases, included in the RSCI or RSCI core. Further, it gives a weighted impact factor of the journals in which articles were published, the total number of authors, the average number of publications per author, the total number of citations, the average number of citations per article, the number of articles cited at least once, the number of self-citations and the Hirsch index.

Here we can also get data for statistical reports on the types of distribution of publications by topic, by keywords, by journals, by organizations, by authors, by years, by the number of co-authors, by the number of citations, citing publications on the topic, citing publications by keywords, citing publications by journals, citing publications by organizations, citing publications by authors, citing publications by years, citations by topic, citations by journals, citations by organizations, citations by authors, citations by years of citing publications, citations by years of cited publications.

Quantitative indicators are taken from the above-mentioned procedure. On their basis, indicators specific to this area are built and directions for further immersion are determined in order to identify hidden patterns.

The second phase, associated with immersion in the semantics of the body of scientific texts, suggests the involvement of knowledge experts in the given field of science. It should be noted that if the analysis is carried out automatically, only with the built-in tools of eLIBRARY.ru, without the participation of an expert, articles that are not related to the given field of science can be included in the selection, and the terms are only mentioned. When a knowledge expert is connected, non-material articles are rejected, and the number of articles in the collection decreases, and only articles that present significant results remain.

## 3. Results of the first stage

For a query in eLIBRARY.ru [1], four combinations of keywords from the field of knowledge were selected: "Artificial Intelligence and Information Security" (AI&IS), "Machine Learning and Cybersecurity" (ML&C), "Data Mining and Cybersecurity" (DM&C), "Machine Learning and Information Security "(ML&IS), "Data Mining and Information Security" (DM&IS). The search was carried out only in the title of the publication, in the annotation and in keywords, which significantly reduced the number of publications in the collection.

The obtained generalized quantitative data obtained from eLIBRARY.ru, which is the most important characteristics of publications on these topics, is summarized in Table 1.

Table 1. Key characteristics of publications on the analyzed topics

| | Research Topics | | | |
|---|---|---|---|---|
| Quantitative characteristics | ML&C | DM&C | ML&IS | DM&IS |
| Total publications | 27 | 28 | 107 | 419 |
| Articles in journals | 16 | 19 | 61 | 281 |
| Articles in journals included in Web of Science or Scopus | 1 | 1 | 6 | 21 |
| Articles in journals included in the RSCI core | 3 | 6 | 21 | 44 |
| Articles in RSCI journals | 2 | 6 | 19 | 34 |
| The weighted impact factor of the journals in which the articles were published | 0,410 | 0,685 | 0,412 | 0,428 |
| Total number of authors | 69 | 73 | 263 | 987 |
| Average number of publications per author | 0,39 | 0,38 | 0,41 | 0,42 |

| | | | | |
|---|---|---|---|---|
| Total number of citations | 30 | 117 | 192 | 1452 |
| Average number of citations per article | 1,11 | 4,18 | 1,79 | 3,47 |
| Number of articles cited at least once | 9 | 12 | 33 | 189 |
| Total number of self-citations | 0 | 0 | 8 | 25 |
| Hirsch index | 3 | 5 | 6 | 18 |

The conclusion from the Table 1 is that the largest number of publications, the largest number of articles in journals, articles in journals included in the Web of Science or Scopus citation bases, the total number of authors, the total number of citations, the average number of citations per article, the number of articles cited at least once, the number of self-citations and the Hirsch index is in the field of "Data Mining and Information Security".

Further, according to data from eLIBRARY.ru, a graph was built reflecting the articles appearance dynamics from 2011 to 2020 (figure 1).

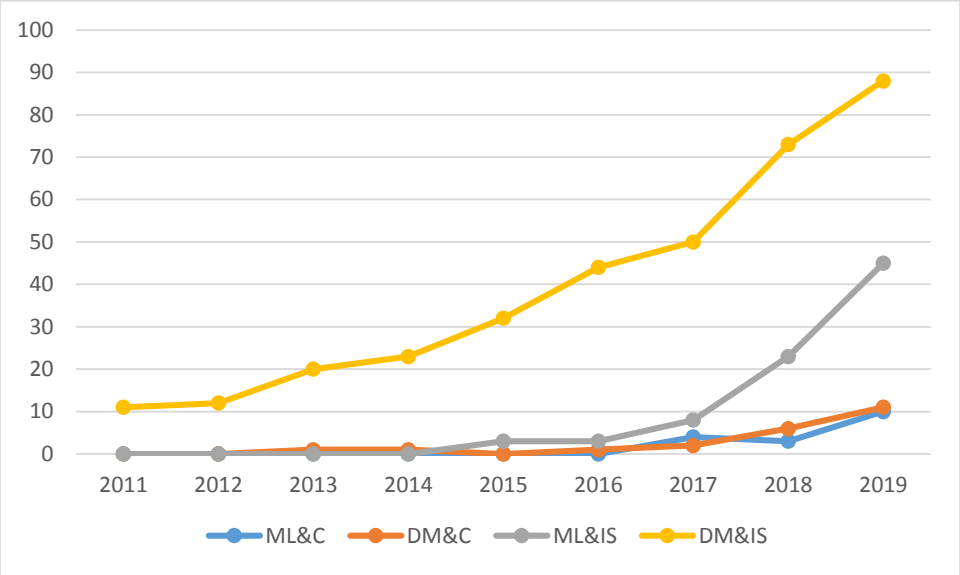


Figure 1. Articles appearance dynamics

From Fig. 2 it is easy to see that since 2011 there has been a steady increase in the number of articles over the years, which shows that this area of knowledge is relevant and the interest of researchers in this topic is growing. 2020 is not included as it is not yet complete.

Further, an analysis of the performance of the most active researchers by topic was carried out (Figures 2-5).

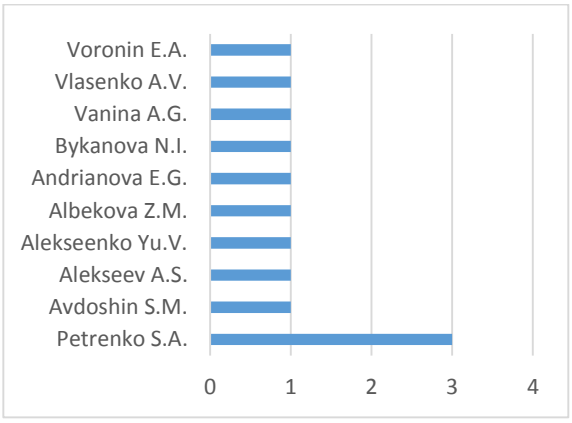


Figure 2. Comparative analysis of performance on the topic "Machine Learning and Cybersecurity"

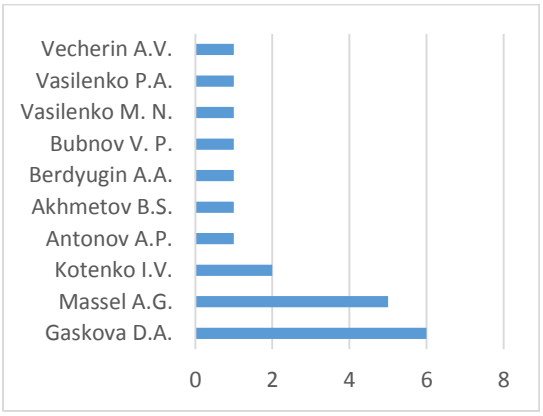


Figure 3. Comparative analysis of performance on the topic "Data Mining and Cybersecurity"

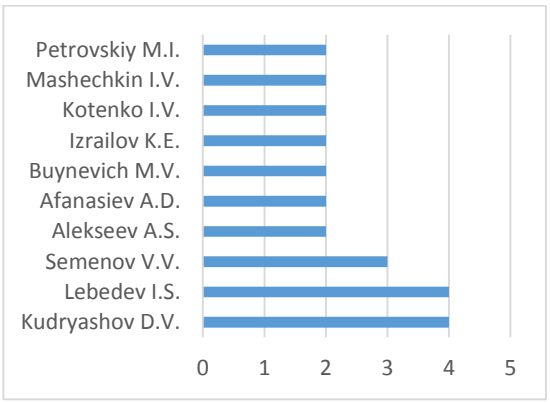


Figure 4. Comparative analysis of performance on the topic « Machine Learning and Information Security»

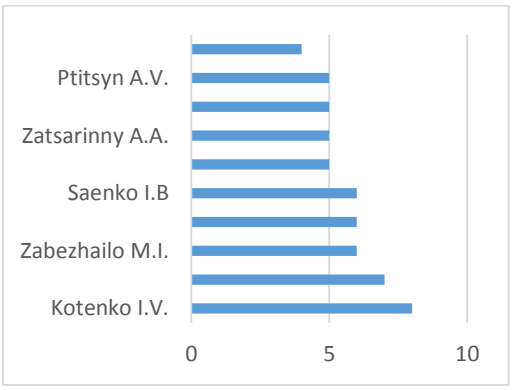


Figure 5. Comparative analysis of performance on the topic «Data Mining and Information Security»

Although the query was based on keyword combinations from the Artificial Intelligence and Information Security domain, there is not much intersection between authors by subject. Only Professor Kotenko I.V. from SPIIRAS, St. Petersburg, is one of the ten most active researchers on three topics and takes the first position in the topic "Data Mining and Information Security". Alekseev A.S. also meets twice. from Don State Technical University, in the topics "Machine Learning and Cybersecurity" and "Machine Learning and Information Security".

Further analysis of the publication activity of research organizations and scientific schools was carried out according to data from eLIBRARY.ru. The results of such analysis is shown in Figures 6-9.

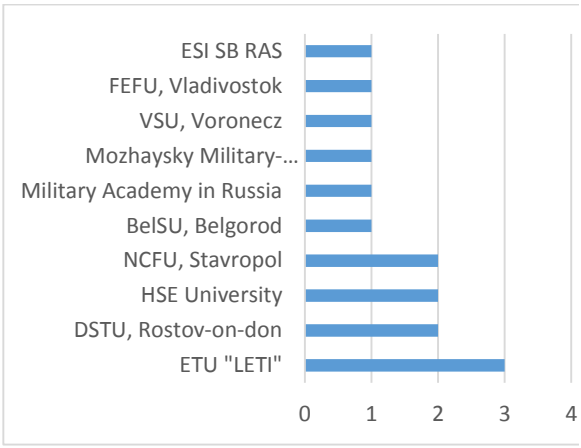


Figure 6. Performance analysis of the most active organizations by topic «Machine Learning and Cybersecurity»


Figure 7. Performance analysis of the most active organizations by topic «Data Mining and Cybersecurity»

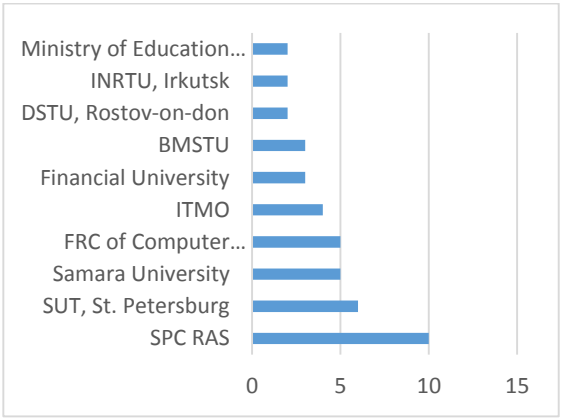


Figure 8. Performance analysis of the most active organizations by topic « Machine Learning and Information Security»

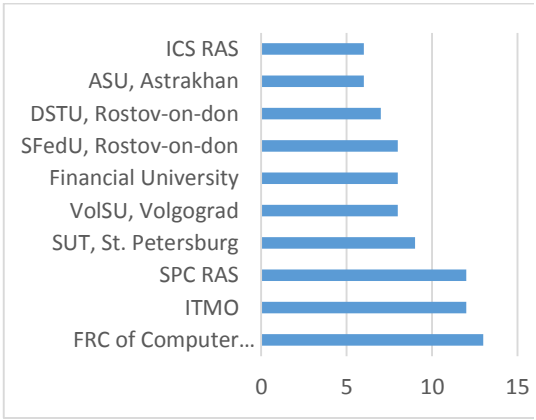


Figure 9. Performance analysis of the most active organizations by topic «Data Mining and Information Security»

There is a stronger intersection of organizations here. The undisputed leader is the St. Petersburg Federal Research Center of the Russian Academy of Sciences, which is present in three out of four topics and takes the first position in the topic "Machine Learning and Information Security". We can also note the Bonch-Bruevich Saint-Petersburg State University of Telecommunications, the Federal Research Center Computer Science and Control of the Russian Academy of Sciences and the National Research University ITMO, noted in two topics and occupying leading positions in them.

Also of interest are the journals in which the authors publish the results of their research. eLIBRARY.ru allows you to extract data on publication activity by journals; the diagrams based on these data are presented in Figures 10-13.

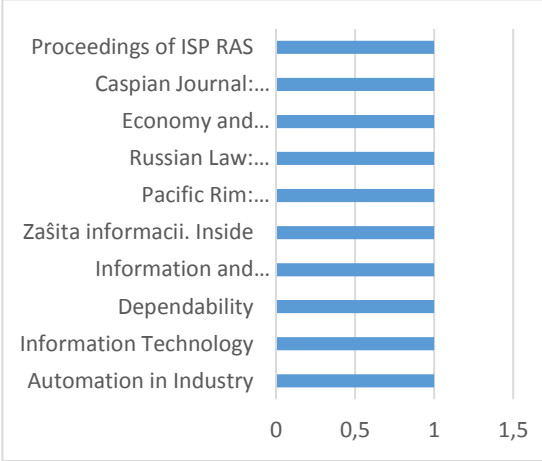


Figure 10. Analysis of journal activity by topic «Machine Learning and Cybersecurity»

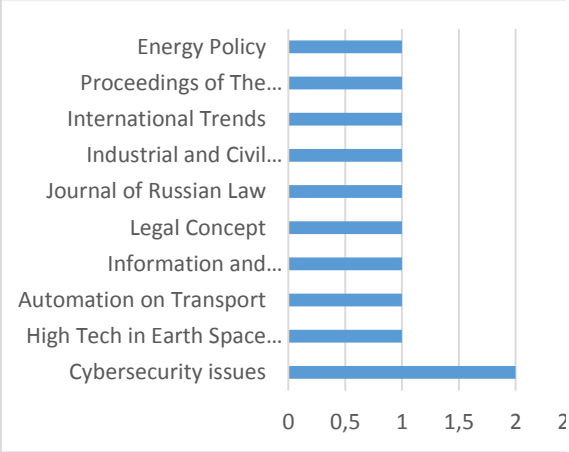


Figure 11. Analysis of journal activity by topic «Data Mining and Cybersecurity»

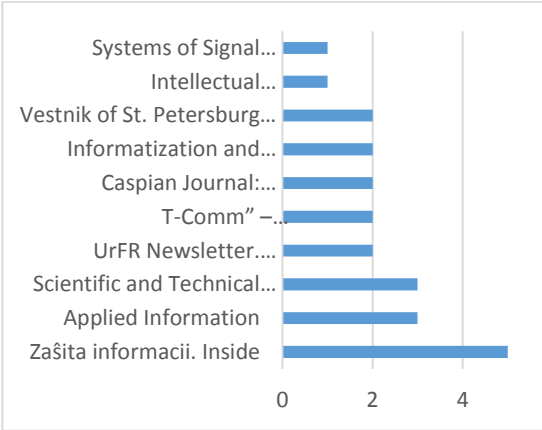


Figure 12. Analysis of journal activity by topic «Data Mining and Information Security»

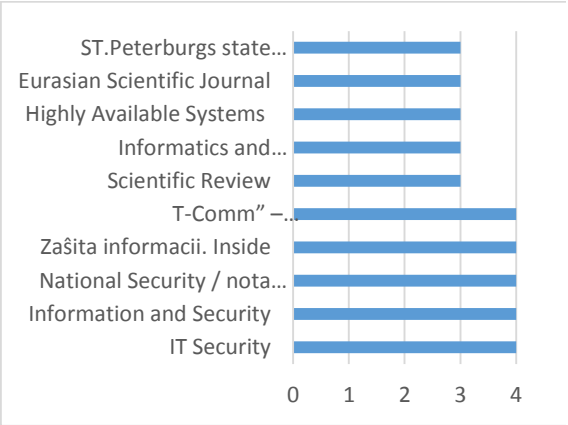


Figure 13. Analysis of journal activity by topic «Data Mining and Information Security»

It can be seen that 4 combinations of keywords do not overlap in active journals. The only exception is the journal "Zaŝita informacii. Inside ", it published the results of research on three topics: "Machine Learning and Cybersecurity", "Machine Learning and Information Security" and "Intelligent Analysis and Information Security". Also worth mentioning is the «Caspian Journal: Management and High Technologies» magazine, which is mentioned twice.

## 4. Results of the second stage

Based on the preliminary processing of information at the first stage, the most popular publications of active authors were found. Analysis of the content of these articles allows us to highlight the most interesting results in the given field of science.

In the papers of Petrenko S.A. from ETU LETI [2, 3] approaches to ensuring the security of information systems of banks and e-commerce using multifunctional BI-platforms are considered. The analysis of key problems in the organization of big data is carried out and problems of implementation, operation and maintenance of BI systems from the point of view of ensuring information security are identified. The cybersecurity ontology (meta-ontology) was proposed as a way to represent knowledge about the qualitative characteristics and quantitative laws of information warfare.

In the papers of S.M. Avdoshin from the Higher School of Economics [4], an analysis of successful cyber-attacks on the blockchain systems industry has been carried out. A list of the main types of successful thefts by hackers is presented.

The results of research by D.A. Gaskova and A.G. Massel from the Melentiev Energy Systems Institute of Siberian Branch of the Russian Academy of Sciences, Irkutsk, are given in [5, 6, 7]. In [5], a risk-oriented decision support approach was formulated in identifying critical objects in the energy sector from the point of view of cyber threats. The novelty of the research lies in the development and application of a fractal stratified model of the relationship between risks, energy facilities and technologies.

In [6], the concepts of Smart Grid and digital energy were discussed, and the main information technologies proposed for their implementation were analyzed. It was proposed to include intelligent technologies in this list to support strategic decisions on energy development.

In [7], a risk-oriented approach was described for analyzing threats and assessing the risk of cybersecurity breaches at energy facilities, taking into account damage caused by damage or destruction of the facility, using quantitative and qualitative parameters.

The papers of the most active author, Professor Igor Kotenko from SPIIRAS are given in [8-12]. In [8], the analysis of existing methods of attacker behavior, characteristics of the attacker's profile and their application for predicting future steps of an attack was given. The analysis made it possible to identify the main advantages and limitations of approaches to predicting attacks and building an attacker's profile, existing problems and prospects in this area.

The [9] examined an approach to online decision-making in telecommunication systems based on the use of hierarchical fuzzy situational networks, which allows making effective decisions in the context of dynamically changing external factors and presents the mathematical foundations for the creation and application of hierarchical fuzzy situational networks for the implementation of fuzzy logical inference in telecommunication networks.

In [10], an approach to the analysis of cybersecurity data based on a combination of a set of machine learning methods and big data technologies for network attacks and anomaly detection was presented. The approach was characterized by several levels of data processing, including extraction and decomposition of datasets, compression of feature vectors, training, and classification.

The paper [11] discussed the development and application of a semantic model for security assessment. The proposed model was presented as an ontology of metrics based on the relationship between security-related data sources, the primary characteristics of the underlying security data, and the objectives of the security assessment.

In [12], methods and techniques for managing the safety of complex heterogeneous systems with an emphasis on the correlation of events and safety assessment were considered. The proposed approach is based on a comprehensive analysis of large heterogeneous security data for event correlation, including syntactic and semantic analysis of security events and information.

The works of I.S. Lebedev from the St. Petersburg Federal Research Center of the Russian Academy of Sciences are presented in [13, 14]. In [13], the issues of ensuring the cybersecurity of autonomous unmanned objects were considered. The prerequisites were identified that determine the need for external monitoring systems. An approach to the analysis of the state of cybersecurity of an autonomous object based on classification methods was proposed and allows one to identify the current state using the processing of digitized acoustic information.

The article [14] presents methods for creating signatures of executable files based on the frequency distributions of their informative characteristics, which can be used to identify programs. A new method for identifying executable files was proposed and the results of experiments on their identification using a statistical criterion were presented.

Works by A.A. Grusho from Federal Research Center Computer Science and Control of the Russian Academy of Sciences are presented in [15-17]. In [15], the problem of revealing weakly expressed anomalies in the model of generalized systems was considered.

In [16], security policy rules for a multilevel access system of a distributed information system were constructed, and on the basis of the infrastructure associated with metadata, the possibility of implementing this security policy in a distributed information computing system was shown.

In [17], an approach was formulated to the study of some types of fraud in the digital economy using causal relationships.

## 5. Conclusion

Thus, the proposed method of processing semi-structured scientific information allows to get an insight of the most interesting scientific results in the subject area of interest in recent years. The analysis of publications based on data from the eLIBRARY.ru platform is a preliminary stage in the study of the corpus of scientific texts on the selected topic. It should be noted that when working with eLIBRARY.ru data, articles are automatically selected by keywords, without taking into account the specifics of the type and specifics of scientific research. To select the necessary material from the selection, you have to resort to a knowledge expert in the given field is science.

Carrying out the analysis "manually" is a rather cumbersome procedure, so it is advisable to develop an information system to support researchers with generalized data based on the analysis of the corpus of scientific texts. Such a system should provide information on the most interesting scientific results obtained recently.

## 6. Acknowledgment

The results of the research presented in this article were supported by Grants RFBR 19-07-00780.

The study is conducted with financial support from the Ministry of Education and Science of the Russian Federation as part of the basic part of the state assignment to higher education educational institutions # FEUE-2020-0007.The work is partially supported by the RFBR grant # 18-01-00796.

## References

[1] Moskaleva O, Pislyakov V, Sterligov I, Akoev M and Shabanova S 2018 Russian Index of Science Citation: Overview and review *Scientometrics* **116** 449–62

[2]. Petrenko S and Makoveichuk K 2020 Development of BI-Platforms for Cybersecurity Predictive Analytics pp 273–88

[3] Petrenko S, Makoveichuk K and Olifirov A 2020 New Methods of the Cybersecurity Knowledge Management Analytics pp 296–310

[4] Lazarenko A and Avdoshin S 2019 Financial Risks of the Blockchain Industry: A Survey of Cyberattacks *Proceedings of the Future Technologies Conference (FTC) 2018* ed K Arai, R Bhatia and S Kapoor (Cham: Springer International Publishing) pp 368–84

[5] Massel A and Gaskova D 2020 Identification of Critical Objects in Reliance on Cyber Threats in the Energy Sector *Acta Polytech. Hungarica* **17** 61–73

[6] Massel L and Massel A 2018 Intelligent support tools for strategic decision-making on Smart Grid development ed N Voropai, F-J Lin, G Chang, A Kler and R Lis *E3S Web Conf.* **69** 02009

[7] Daria G and Massel A 2018 Intelligent System for Risk Identification of Cybersecurity Violations in Energy Facility 2018 3rd Russian-Pacific Conference on Computer Technology and Applications (RPC) (IEEE) pp 1–5

[8] Doynikova E, Novikova E and Kotenko I 2020 Attacker Behaviour Forecasting Using Methods of Intelligent Data Analysis: A Comparative Review and Prospects Information 11 168

[9] Kotenko I, Saenko I and Ageev S 2019 Hierarchical fuzzy situational networks for online decision-making: Application to telecommunication systems *Knowledge-Based Syst.* **185** 104935

[10] Kotenko I, Saenko I and Branitskiy A 2020 Machine Learning and Big Data Processing for Cybersecurity Data Analysis *Data Science in Cybersecurity and Cyberthreat Intelligence* (Springer) pp 61–85

[11] Doynikova, E., Fedorchenko, A., Kotenko, I., 2020. A semantic model for security evaluation of

information systems. Journal of Cyber Security and Mobility 9, 301–330. doi:10.13052/JCSM2245-1439.925
[12] Kotenko I, Fedorchenko A and Doynikova E 2020 Data Analytics for Security Management of Complex Heterogeneous Systems: Event Correlation and Security Assessment Tasks *Advances in Cyber Security Analytics and Decision Systems* (Springer) pp 79–116
[13] Semenov V, Sukhoparov M and Lebedev I 2019 Approach to Side Channel-Based Cybersecurity Monitoring for Autonomous Unmanned Objects *International Conference on Interactive Collaborative Robotics* pp 278–86
[14] Salakhutdinova K, Lebedev I, Krivtsova I, Bazhayev N, Sukhoparov M, Smimov P, Markelov D, Davvdov A, Pecherkin S, Kolcherin D, Shaparenko Y and Iskanderov Y 2017 A Frequency Approach to Creation of Executable File Signatures for their Identification *2017 IEEE 11th International Conference on Application of Information and Communication Technologies (AICT)* (IEEE) pp 1–7
[15] Grusho A, Grusho N and Timonina E 2020 Method of Several Information Spaces for Identification of Anomalies *International Symposium on Intelligent and Distributed Computing* pp 515–20
[16] Grusho A A, Grusho N A and Timonina E E 2019 Using Metadata to Implement Multilevel Security Policy requirements *Informatics Appl.* **13** 85–9
[17] Grusho A A, ZABEZHAILO M I, Grusho N A and Timonina E E 2019 Architectural Decisions in the Problem of Identification of Fraud in the Analysis of Information Flows in Digital Economy *Informatics Appl.* **13** 22–8