Towards Validation of Insider Threat Identification Algorithm with Synthetic Data

Oksana Nikiforova[1,*,†] and Vitaly Zabiniako[2,†]

[1] *Riga Technical University, Riga, Latvia*
[2] *"ABC software" Ltd., Riga, Latvia*

Abstract

This paper addresses the challenge of detecting insider threats in cybersecurity by proposing behavior model-driven approaches. It argues that existing datasets are incapable to capture nuanced user activities accurately and proposes an enhanced dataset generated by more elegant structure. The paper discusses the evolving threat situations and the need for proactive cybersecurity measures, presents a taxonomy of insiders, and emphasizes the importance of behavior-driven approaches. It mentions existing datasets limitations and introduces the proposed data generator structure, explaining its components and implementation logic. The paper illustrates a use case showcasing the application of generated data for insider threat identification. It concludes by stressing the significance of behavior-driven approaches and high-quality datasets in enhancing detection capabilities against insider threats.

Keywords

Insider threat identification, machine learning, cyber security, synthetic dataset generation

1. Introduction

Insider threat identification stands as a pressing concern within cybersecurity landscapes. Unlike external threats, insiders possess legitimate access to organizational systems, making their detection inherently challenging. The potential consequences of insider threats, including data breaches, intellectual property theft, and sabotage, underscore the criticality of effective identification mechanisms [1]. Addressing this issue requires a refined understanding of insider behaviors necessitating advanced detection methodologies. Thus, the recognition of insider threats as a relevant issue reflects the evolving threat landscape and underscores the imperative for proactive cybersecurity measures [2].

The state of the art of insider threat detection is characterized by a rapidly growing array of methodologies and techniques under continual development. Researchers and cybersecurity professionals are actively exploring diverse approaches to enhance the efficacy and robustness of detection systems [3], [4], [5]. These approaches encompass a

Baltic DB&IS Conference Forum and Doctoral Consortium 2024
* Corresponding author.
† These authors contributed equally.
oksana.nikiforova@rtu.lv (O. Nikiforova); vitalijs.zabinako@abcsoftware.lv (V. Zabiniako)
0000-0001-7983-3088 (O. Nikiforova); 0000-0002-1307-1815 (V. Zabiniako)

spectrum of methodologies, ranging from rule-based systems to advanced machine learning algorithms. Each approach brings its own strengths and limitations, reflecting the complex nature of insider threat detection. This diversity in approaches reflects a collective effort to address the evolving challenges posed by insider threats, with ongoing researches aimed at pushing the boundaries of detection capabilities.

Central to the advancement of insider threat detection algorithms is the availability of high-quality datasets for accurate testing and evaluation. However, acquiring and managing datasets that accurately represent the complexities of insider threat behaviors poses a significant challenge. Traditional datasets often lack the diversity and granularity necessary to effectively train and validate detection systems. Moreover, the sensitive nature of insider threat data complicates the sharing and accessibility of suitable datasets [6]. Addressing this challenge requires innovative approaches to dataset development, including synthetic data generation techniques. Overcoming this obstacle is crucial for advancing the state-of-the-art in insider threat detection algorithms.

To address the problem of obtaining suitable datasets for testing insider threat detection algorithms, our paper presents according solution: an algorithmic data genera-tor tailored specifically for this purpose. This data generator allows to precisely define and simulate realistic insider threat scenarios, producing datasets that encapsulate a wide range of behavioral patterns. The generated datasets serve as valuable resources for training, testing, and benchmarking insider threat detection algorithms.

The paper is organized as follows. The next section gives an overview of insider threats identification researches and data sets used for its validation, as well as an insight into some examples of datasets suitable for validation of insider threat identification algorithms. The third section describes the authors offered algorithm for such data generation, where the typical scenarios are defined for usage of common information systems and the predefined probabilities are set to simulate regular work of information systems users. The fourth section demonstrates a use case of the generated data application for validation of the insiders' threats identification approach offered by authors and published in [7], [8], [9]. The fifth section gives main conclusions of the research and states real-world implications and future directions of the research.

2. Background and Related Work

The recent years have shown an increase in inside threats, like those involving Edward Snowden, Chelsea Manning and Kim [10]. A cybersecurity report from 2018 told us that more than half of the threats they studied came from within organizations; for 27% it was a common occurrence to deal with insider risks [11]. Another report later on showed that there had been a noticeable rise in what people believed were attacks from insiders: 63% felt there was more activity going on. There is worry about the increasing difficulty of recognizing insider attacks compared to outside dangers, because those who are inside have been given permission and their actions can be complex [12]. Attacks from insiders, which frequently happen within normal working hours, create difficulties in examining large activity logs [13], [14].

An insider typically refers to an individual possessing authorized access to an organization's computer systems and networks [15]. Worries increase about the challenge of finding insider attacks compared to external dangers. Insiders, who might not have high-level technical abilities, mix the difference between harmful and authorized actions. We can separate insiders into three main categories: traitors, masqueraders, and unintentional offenders. Those who are traitorous pursue their own gain or money, masqueraders pretend to be lawful but do illegal work and unintentional perpetrators breach security by mistake. Usually, malicious insiders commit IT sabotage, steal intellectual property or conduct fraud for financial reasons [16].

Validation of insider threat's identification algorithms requires datasets that capture user activities within organizational networks, where users perform different activities in particular information systems under specific scenarios. Datasets play a crucial role in the study of insider threat detection. However, there is still an absence of a wide-ranging dataset from real-life situations that is available to the general public. Many existing datasets depend on attacks that are made up artificially. In a recent survey by [17], the datasets are classified as masquerader-based, traitor-based, substituted masqueraders and identification/authentication-based types along with other malicious ones. The datasets investigated by authors are as follows: Amazon Employee Access Challenge (AEAC), CERT, DARPA Insider Threat Evaluation (DITREC), Enron, Greenberg's dataset, Los Alamos National Laboratory (LANL), RUU, Schonlau, TWOS, University of New Brunswick datasets.

When utilizing such datasets, researchers must prioritize conformance to any data usage agreements and ethical guidelines. Furthermore, it is crucial to assess the diversity and realism of scenarios depicted within the dataset to accurately measure both the effectiveness and efficiency of insider threat identification algorithms.

Based on the analysis of insiders' threats identification algorithms, authors propose to categorize the algorithms into the taxonomy shown in Figure 1.

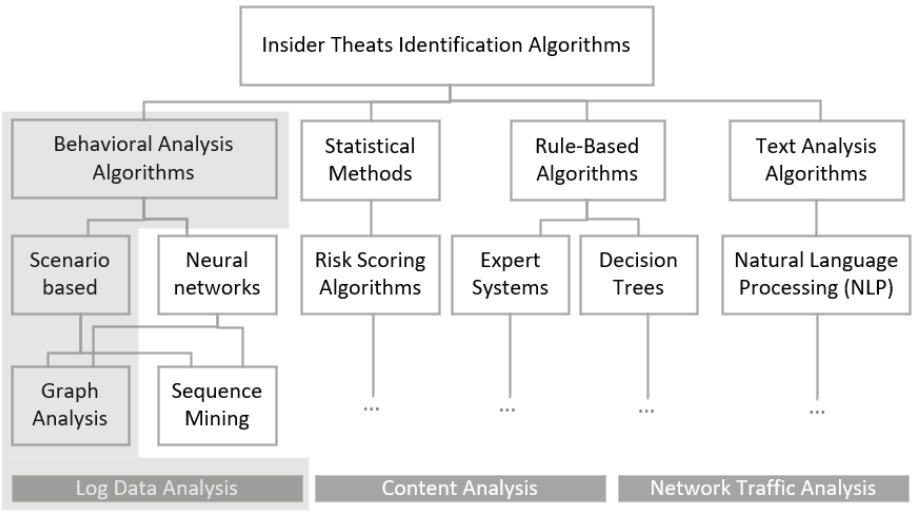

Figure 1: Insiders threats identification algorithms taxonomy.

The taxonomy the algorithms into four parts behavior-oriented, statistical model-driven, rule-based and algorithms working with text.

The set of user behavioral analysis based insider detection algorithms, among other, includes the following techniques:

- Markov Chains / Hidden Markov Model (HMM) [18];
- Clustering Algorithms [19];
- Graph-Based Approaches [20];
- Self-Organizing Maps (SOM) [21];
- Gaussian Mixture Models (GMM) [22];
- Nearest Neighbor Methods [23];
- One-Class Support Vector Machines (SVM) [24].

Hybrid methods combine multiple base models to improve the overall detection performance. By leveraging the diversity of multiple models, hybrid methods enhance the robustness and reliability of insider threat detection systems.

The focus of the paper is on the behavior model-driven approaches, where user behavior scenarios in the form of user activities graphs are applied to identify insiders' threats. All the algorithms for insiders' threats identification, which use behavior model as a core of the approach, requires data about users' activities in the information systems with the greater detail than the data offered in the data set described above. Audit logs that capture business-level actions (e.g., logins, document accesses, data edits, information searches, etc.) are of especial importance. Such logs provide the granular insights necessary for accurate training and testing of algorithms de-signed to detect insider threats. By examining such nuanced data and behavioral patterns, algorithms can better identify anomalies indicative of malicious intent or unauthorized activities.

For identification of insiders' threats at the business level a detailed analysis of users' activities is required. The data sources mentioned above has lesser efficiency regarding this aspect, so the improved data set is offered further, based on the proposed Data Generator implementation logic.

3. Data Generator General Structure

The data generator allows to generate simulated user activity data for a specified number of users within a specified time period with certain probability settings. The input data include the following information:

UserName - User identifier. In the generated data, this is a randomly chosen female name (from an international list of female names) and a surname (randomly chosen animal name).

SessionId - Session identifier. In the generated data, a session is a set of ac-tions performed, starting with the first action performed by a user in the in-formation system (usually the "login" action) and ending with the exiting from the information system (usually the "logout" action).

ActionTime - Time of action execution.
ActionName – Name (identifier) of the action.

The generated data simulates actions from e-mail system (like, open Outlook, view e-mail list, open e-mail, send e-mail, etc.), Web browser (like, open, view, open gmail, open google drive, download file, etc.), and some information system (like, open menu, search data, edit data, etc). All scenarios start with log-in action, some actions with low probability and some malicious actions, which are called as maliciousAction1, 2, 3 and are performed once for a specific user.

The foundation of the data generator lies in a developed script, enabling the artificial generation of a dataset mimicking user actions in information systems and recording it in a form of an audit log. The generator allows for the emulation of user actions within a specified time period with defined probabilities and according to predefined behavioral models (including options for simulating malicious activities). The data generation cycle itself begins, applying random value generation within the specified intervals for performed actions. Within this cycle, the fulfillment of other necessary conditions is also observed, such as: if the execution time of user actions approaches the end of the workday, the user sequentially ends work in all active sessions by performing the "logoff" action; and if work is completed on a Friday, the next set of actions will only commence execution on Monday morning.

After data generation, the algorithm performs additional session filtering to remove extremely short and extremely long sessions with action counts less and more than the specified number of actions should be performed during one session. This ensures additional quality of the generated data from the perspective of data volume. Then the algorithm arranges the generated data records in ascending order of action execution time and saves the final results.

In addition to the generated data from the historical period, real-time data streaming functionality has been implemented. The utilization of real-time data generation enables the immediate detection of security incidents by insider threat identification algorithms and real-time reporting of security incidents to the security administrator.

4. Use Case of Generated Data Application

The developed data generator has been applied to validate a custom algorithm created by the authors, which enables the identification of malicious user actions based on the audit logs of user-executed actions within the IS, thereby constructing the user behavior model in the form of a graph for specific IS usage. The user behavior model allows to make conclusions regarding anomalies in a specific user's behavior (or deviations in it), identified as malicious activities, and immediately alerts the data security administrator about malicious incidents when potential breaches occur [7], [8], [9].

In this paper, the authors demonstrate the validation of a list of malicious actions identified by their algorithm using a dataset of user activities generated by the data generator described above. This establishes mutual validation possibility, since we a priori know which data are malicious in the generated dataset and which sessions should be identified as malicious. Consequently, the algorithm should produce these sessions as a

result. The ability to utilize the generated data in the insider threat identification algorithm confirms that the data generator has produced data suitable for validating such an algorithm.

The validation was conducted using generated activity data for 20 users, simulated over a period of 30 days. In these three scenarios, the actions saveMail (email), downloadFile (web browser), and printDocument (IS) are executed with significantly lower probabilities compared to other actions in these scenarios. Real world data contains actions that are rarely performed, potentially representing malicious activities.

Furthermore, to simulate the execution of malicious actions, sessions for the user "Miranda Hyena" are generated, incorporating entirely malicious actions with specific names such as "maliciousAction1", "maliciousAction2", and "maliciousAction3". As a result, the trust coefficient for the list of users should be less than 100 only for the user Miranda Hyena, as shown in Figure 2, where the top five users are shown. The figure illustrates that the user *Miranda Hyena* occupies the highest position, with the only user with the numbers of malicious sessions identified according to various criteria. Columns 1 through 6 indicate the number of sessions recognized as malicious according to different applied criteria (the algorithm that calculates malicious sessions based on extreme difference in a subgrouping provided a more precise result with a count of 4 malicious sessions, indicating that the application of generated data for algorithm validation demonstrated instances of more accurate computation [25]). The presence of malicious sessions for a particular user corresponds to the trust percentage calculated by considering the number of unique actions in a session and the number of unique malicious actions in a session. Thus, the trust percentage can replace all six coefficients, allowing the security administrator in the real system to operate based solely on the total number of malicious sessions for each user.

UserName	Trust%	UniqueSteps	UniqueSuspiciousSteps	SuspiciousSessions	Sessions	1	2	3	4	5	6
Miranda Hyena	76.6	47	11	6	92	2	0	3	0	4	0
Scarlett Monkey	100.0	38	0	0	119	0	0	0	0	0	0
Zoey Vulture	100.0	39	0	0	158	0	0	0	0	0	0
Amelia Koala	100.0	37	0	0	156	0	0	0	0	0	0
Ava Panda	100.0	37	0	0	145	0	0	0	0	0	0

Figure 2: Malicious actions and sessions identified for the user *Miranda Hyena*.

The malicious sessions, for which trust percentage is less than 100% (indicating that the session contains at least one malicious action), are selected into a separate list, as shown in Figure 3. In the synthetic data, there are 4 sessions containing malicious actions (sessions with identifiers 11568, 11573, 11403 and 11579). It can be observed that the intentionally generated sessions with malicious actions are identified in the resulting list and even appear at the top of the list (see Figure 3). Figure 4 presents a fragment of detailed actions data for a particular session 11568 with malicious actions. The full session in the form of graph is shown in Figure 5.

SessionId	Trust%	Steps	UniqueSteps	UniqueSuspiciousSteps	SessionStartTime	SessionEndTime	Duration(sec)	UserName
11568	70.0	19	10	3	2024-02-26 09:24:00	2024-02-26 18:02:27	31107	Miranda Hyena
11573	75.0	13	8	2	2024-02-27 11:22:29	2024-02-27 18:00:35	23886	Miranda Hyena
11403	80.0	19	10	2	2024-02-01 13:15:31	2024-02-01 14:45:26	5395	Miranda Hyena
11579	80.0	18	10	2	2024-02-27 17:02:14	2024-02-27 18:00:31	3497	Miranda Hyena
13557	100.0	6	4	0	2024-01-31 09:04:21	2024-01-31 10:34:50	5429	Amelia Koala
13559	100.0	14	9	0	2024-01-31 09:17:08	2024-01-31 11:33:54	8206	Amelia Koala

Figure 3: A list of malicious sessions of the user *Miranda Hyena*.

TimeFrom	ActionFrom	ActionTo	TimeTo	Duration(sec)	SuspiciousStep
2024-02-26 09:24:00	login	openOutlook	2024-02-26 09:35:34	694	0
2024-02-26 09:35:34	openOutlook	viewEmailsList	2024-02-26 09:42:27	413	0
2024-02-26 09:42:27	viewEmailsList	openEmail	2024-02-26 09:43:58	91	0
2024-02-26 09:43:58	openEmail	viewEmailsList	2024-02-26 10:00:46	1008	0
2024-02-26 10:00:46	viewEmailsList	openEmail	2024-02-26 10:55:20	3274	0
2024-02-26 10:55:20	openEmail	maliciousAction1	2024-02-26 11:31:53	2193	1
2024-02-26 11:31:53	maliciousAction1	createNewMail	2024-02-26 12:08:26	2193	1
2024-02-26 12:08:26	createNewMail	sendMail	2024-02-26 12:44:59	2193	0

Figure 4: Detailed data for actions of the session with identifier *11568*.

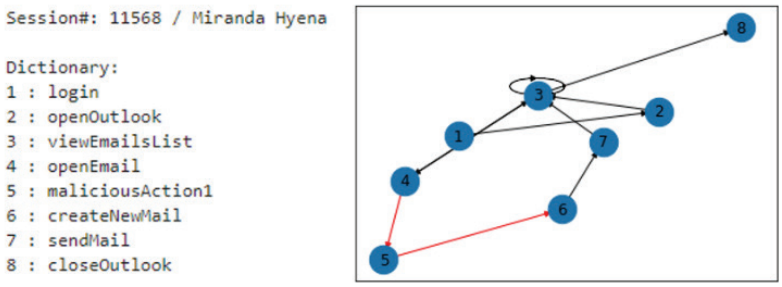

Figure 5: Actions graphs of the session with identifier *11568*.

To summarize, the results of identifying malicious users align with expectations - the user who had sessions with malicious actions embedded in the synthetic data has the highest maliciousness coefficient. The results of identifying malicious sessions align with expectations as well - sessions with the highest maliciousness were identified, containing artificially inserted malicious actions. The visualization results of user session graphs also align with expectations – transition to / from artificially introduced atypical malicious actions are visually highlighted with color.

More accurate results in identifying malicious actions are ensured by behavior analysis in subgroups compared to individual and group behavior analysis. It is recommended not to directly apply malicious actions identification to individual behavior models in groups with 1-2 users, but rather to analyze subgroup behavior models in groups with three or more users. More accurate results in identifying malicious sessions are provided by parameters based on searching for actions with extremely low probabilities compared to calculating the average probability of sessions. Overall, the trust percentage of sessions,

calculated by considering the number of unique actions in a session and the number of unique malicious actions in a session, can replace all six coefficients.

As for the evaluation of the generated data, it can be seen that sessions with data allowing for the identification of malicious actions were created. The synthetic dataset can serve as a basis for validating insider threat identification algorithms.

5. Conclusion

To sum up, this research addresses a major gap in the realm of insider threat identification algorithms by introducing an appropriate solution tailored explicitly for behavior-based approaches. While various existing datasets exist for testing such algorithms, none provide the level of granularity necessary to thoroughly evaluate business logic, particularly concerning behavior-based algorithms. Our proposed solution bridges this gap, offering researchers and practitioners an easy-configurable platform to test and refine detection strategies effectively.

One key feature of our solution is the flexibility it provides in experimenting with action probabilities within the data generator configuration. This enables the creation of diverse user behavior scenarios, allowing for thorough testing of algorithm robust-ness under various circumstances. By simulating different types of user behavior, researchers can gain valuable insights into the effectiveness of their algorithms across a spectrum of potential threat scenarios.

Moreover, our synthetic data generation capability offers the opportunity to enhance real historical IS audit data. By seamlessly integrating our generated data with existing datasets, researchers can augment the pool of users' behavioral models and activities, thereby enriching the dataset for more comprehensive analysis. This integration enables the exploration of new user behaviors and the evaluation of algorithm performance in scenarios not usually encountered in real-world data.

Overall, our proposed solution empowers researchers and practitioners to advance their understanding of security risks and enhance their defense mechanisms. By providing a versatile platform for generating suitable datasets for algorithm testing and refinement, our solution contributes to the ongoing efforts to combat insider threats and safeguard sensitive information in today's dynamic organizational environments.

Acknowledgements

The research leading to these results has received funding from the research project "Competence Centre of Information and Communication Technologies" of EU Structural funds, contract No. 5.1.1.2.i.0/1/22/A/CFLA/008 signed between IT Competence Centre and Central Finance and Contracting Agency. The research title is "Development of a method for analysis and automatic grouping of information system users with similar behavior, using an AI/ML approach". The project is co-financed by the Recovery Fund of the Action Program "Latvian Recovery and Resilience Mechanism Plan 5.1.r. 5.1.1.r. of the reform and investment direction "Increasing productivity through increasing the amount of investment

in R&D" reforms "Management of innovations and motivation of private R&D investments" 5.1.1.2.i. investment "Support instrument for the development of innovation clusters" implementation rules within the competence centers" framework.

The intellectual property "System and Method for Detecting Atypical Behavior of Users in an Information System by Analyzing their Actions Using a Markov Chain and an Artificial Neural Network" is submitted to World Intellectual Property Organization on 2021/02/26.

References

[1] M.N. Al-Mhiqani, A. Rabiah, Z. Abidin, W. Mohamed, et al., A Review of Insider Threat Detection: Classification, Machine Learning Techniques, Datasets, Open Challenges, and Recommendations, Applied Sciences 10(15), 1-41 (2020). DOI: 10.3390/app10155208

[2] Cost of Insider Threats Global Report, 2022. URL: https://protectera.com.au/wp-content/uploads/2022/03/The-Cost-of-Insider-Threats-2022-Global-Report.pdf

[3] F.R. Alzaabi, A.Mehmood, A Review of Recent Advances, Challenges, and Opportunities in Malicious Insider Threat Detection Using Machine Learning Methods, EEE Access PP(99), 1-1 (2024). DOI: 10.1109/ACCESS.2024.3369906

[4] B.B.Sarhan, N.Altwaijry, Insider Threat Detection Using Machine Learning Approach. Applied Sciences 13(1), 259 (2022). DOI: 10.3390/app13010259

[5] L.L. Ko, D.M. Divakaran, Y.S. Liau, V.Thing, Insider Threat Detection and its Future Directions, International Journal of Security and Networks 12(3), (2016). DOI: 10.1504/IJSN.2017.10005217

[6] B. Lindauer, J. Glasser, M. Rosen, K. Wallnau, Generating test data for insider threat detectors, Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications 5(2), 80-94 (2014). DOI: 10.22667/JOWUA.2014.06.31.080

[7] O. Nikiforova, V. Zabiniako, J. Kornienko, E-Step Control: Solution for Processing and Analysis of IS Users Activities in the Context of Insider Threat Identification Based on Markov Chain, Intelligent Systems and Applications, pp. 345-359. Lecture Notes in Networks and Systems (2024). DOI: 10.1007/978-3-031-47721-8_23

[8] P. Garkalns, O. Nikiforova, V. Zabiniako, J. Kornienko, Analysis of the Behavior of Company Employees as Users of Information Systems or Tools, Based on Employees Clustering with K-means Algorithm,IEEE 64th International Scientific Conference on Information Technology and Management Science of Riga Technical University (ITMS), pp. 1-7. (2023). DOI: 10.1109/ITMS59786.2023.10317652

[9] O. Nikiforova, A. Romanovs, V. Zabiniako, J. Kornienko, Detecting and Identifying Insider Threats Based on Advanced Clustering Methods, IEEE Access (12), 30242-30253 (2024). DOI: 10.1109/ACCESS.2024.3365424

[10] M.N. Al-Mhiqani, R. Ahmad, Z. Abidin, et al.: A new intelligent multilayer framework for insider threat detection. Computers & Electrical Engineering (97) (2022). DOI: 10.1016/j.compeleceng.2021.107597.

[11] Insider Threat Report, https://crowdresearchpartners.com/wp-content/uploads/2017/07/Insider-Threat-Report-2018.pdf,

[12] Insider Threat Report, https://www.cybersecurity-insiders.com/wp-content/uploads/2019/11/2020-Insider-Threat-Report-Gurucul.pdf

[13] N. Saxena, E. Hayes, E. Bertino, P. Ojo, K. Choo, P. Burnap, Impact and Key Challenges of Insider Threats on Organizations and Critical Businesses. Electronics 9(9), 1460 (2020). DOI: 10.3390/electronics9091460
[14] P. Legg, O. Buckley, M. Goldsmith, S. Creese, Automated Insider Threat Detection System Using User and Role-Based Profile Assessment. IEEE Systems Journal 99(2). 1-10 (2015). DOI: 10.1109/JSYST.2015.2438442
[15] C.P. Pfleeger, Reflections on the Insider Threat. In: Insider Attack and Cyber Security. Advances in Information Security (39) Springer, Boston, MA (2008). DOI: 10.1007/978-0-387-77322-3_2
[16] L. Liu, O. De Vel, Q. Han, J. Zhang, Y. Xiang, Detecting and Preventing Cyber Insider Threats: A Survey. IEEE Communications Surveys & Tutorials 20(2), 1397-1417 (2018). DOI: 10.1109/COMST.2018.2800740.
[17] I. Homoliak, F. Toffalini, J. Guarnizo, Y. Elovici, M. Ochoa, Insight Into Insiders and IT: A Survey of Insider Threat Taxonomies, Analysis, Modeling, and Countermeasures. ACM Comput. Surv. 52(2) (2020). DOI: 10.1145/3303771
[18] P. Dymarski, Hidden Markov Models, Theory and Applications DOI: 10.5772/601
[19] C.C. Aggarwal, C.K. Reddy, Data Clustering: Algorithms and Applications (2014) DOI: 10.1201/9781315373515
[20] P. Foggia, G. Percannella, C. Sansone, M. Vento, A Graph-Based Clustering Method and Its Applications. Advances in Brain, Vision, and Artificial Intelligence. 277-287 (2007). DOI: 10.1007/978-3-540-75555-5_26
[21] U. Asan, S. Ercan, An Introduction to Self-Organizing Maps, Computational Intelligence Systems in Industrial Engineering: with Recent Theory and Applications, pp. 299-319 (2012). DOI: 10.2991/978-94-91216-77-0_14
[22] V. Garcia, F. Nielsen, R. Nock, Hierarchical Gaussian Mixture Model, Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP, pp. 14-19. (2010). DOI: 10.1109/ICASSP.2010.5495750
[23] N. Bhatia, Survey of Nearest Neighbor Techniques, International Journal of Computer Science and Information Security. 8(2), 302-305 (2010). DOI: 10.48550/arXiv.1007.0085
[24] M. Amer, M. Goldstein, S. Abdennadher, Enhancing one-class Support Vector Machines for unsupervised anomaly detection, Proceedings of the ACM SIGKDD Workshop on Outlier Detection and Description, ODD 2013. pp. 8-15. (2013). DOI: 10.1145/2500853.2500857.
[25] O. Nikiforova, V. Zabiniako, Beyond Information System User Behavior Models: The Power of User Groups in Preventing Insider Attacks. Intelligent Systems and Applications. Lecture Notes in Networks and Systems (2025) *accepted for publication*