JOSEIT JOURNAL OF SYSTEMS ENGINEERING AND INFORMATION TECHNOLOGY
http://jurnal.iaii.or.id/index.php/JOSEIT

Analysis Of the Behavior of Cyberattacks on Online Services Using the Cyber Threat Classification

Isaev Sergey Vladislavovich[1*], Kononov Dmitry Dmitrievich[2]
[12]Institute of Computational Modeling SB RAS, Russia

[*]*si@icm.krasn.ru*

Abstract

The paper contains a study of the dynamics of attacks on online services using the categorization of cyber threats by type in the corporate network of the Krasnoyarsk Scientific Center of the Siberian Branch of the Russian Academy of Sciences. The study was conducted using online service logs and allows solving pressing issues related to ensuring the built-in security of web services, such as: identifying both current and future cybersecurity risks. A summary of the most important logging and analysis techniques is provided. The authors describe the nature and content of the data sources and the software used. The extensive observation period of the study is one of its outstanding features. The structure of the processing system is provided and software tools for attack analysis and categorization are created. The paper shows that using categorized sampling allows for the detection of periodicity and the identification of patterns in specific types of attacks. A correlation matrix was created based on the type of attack. Except for Command Injection, Directory Browsing, and Java Code Injection attacks, which can be aggregated, the research found that most attack types had poor correlation. Based on the classification of cyber threats, the authors proposed a heuristic technique of risk comparison.

Keywords: Analysis, Security, Web, Internet, Attack, Corporate Network

1. Introduction

Currently, many companies use web technologies to organize business services at different levels (mail, cloud technology, hosting, video conferencing). It should be noted that web services are subject to information security risks as they operate on the open Internet. An important part of the functioning of modern information systems is the task of information security, which is complex and includes a number of measures at different levels, the implementation of which will help reduce the risks of cyber threats. An important part of security is the analysis of various activity logs generated by the system [1].

Of particular interest are the logs of the web servers Nginx and Apache, the analysis of which enables the detection of cyber-attacks on the system. In web systems, the volume of logs can be significant, which makes manual analysis difficult, in which case it is necessary to use automated tools for data processing and analysis [2]. Data analysis usually involves processing various software tools and is a multi-step process [3; 4]. The data obtained from the analysis can be used to model the information security system [5] or to compare the behavior patterns of the devices with real cyber-attacks [6].

Different approaches are followed when analyzing journals. One of the most popular methods is signature analysis. Log handlers use predefined signatures to identify and classify malicious events [7; 8th.]. Additional parameters and features can be extracted from the log elements, which can be used for subsequent analyzes such as clustering and anomaly detection [9]. Cyber-attack failures usually produce log entries that differ from those that reflect normal system behavior. It is therefore advisable to pay attention to individual log entries that do not fit into the overall picture.

With clustering, such data sets are identified by a high level of dissimilarity to all existing clusters or do not match any signature [10; 11]. However, not all unwanted system events manifest themselves as individual anomalous log entries, but rather as dynamic or sequential anomalies. Therefore, approaches are needed that allow sequences of recordings to be grouped or temporal patterns and correlations to be identified. Dynamic clustering allows for the identification of events that have multiple heterogeneous and distinct log entries over time [12; 13], allowing implicit abnormal behavior to be detected.

Existing works use different methods to analyze service logs. Often the authors describe the analysis methodology and use test data as an example, which does not allow evaluating the performance of the approach on real data. Or they use real data with short time intervals, which makes it difficult to analyze the dynamics of processes over different time periods. In this work, we study the security of the corporate network of the Krasnoyarsk Scientific Center (FITs KSC SB RAS) based on the analysis of web service logs.

The aim of the work is to analyze the security of web services in the dynamics of the last 2 years, to classify cyber-attacks by type and to identify dependencies between different attack parameters. In contrast to the

Accepted by editor: 25-08-2022 | Final revision: 15-09-2022 | Online publication: 25-09-2022

existing work, the analysis is carried out over long time intervals, thanks to which the dynamics of the behavior of the web services can be shown by hours, days, months and years. The work is a continuation of the study on the security of web services in corporate networks [14], in comparison with previous works, a classification of cyber threats by type was carried out and a risk assessment method was proposed.

2. Method

The data sources for the analysis in this paper are the web services data for 2020-2021. and incomplete 2022 (volume 45 GB, 176 million items). The analysis was performed using the following software tools: UNIX tools, GAccess, libmaxmind, JSON tools, Python, Microsoft Excel. Figure 1 shows the phases of data processing. The primary processing involves aggregating logs from all web services and unifying the format for further processing.

For all data, the source is georeferenced, i.e. country code based on the IP address (GeoIP). Then error processing (both client and server) is performed with aggregation over different time intervals (year, month, day, hour). Attacks are also parsed, which involves classification by type and then aggregation by geospatial data. Attack classification by type is performed by OWASP [15] using a set of ModSecurity Core Rule Set [16] developed for web applications to identify cyber threats. The GSec GSec program suite in the Go and C languages serves to process the attacks, which carries out an automated attack type classification and aggregation of data in different time intervals.

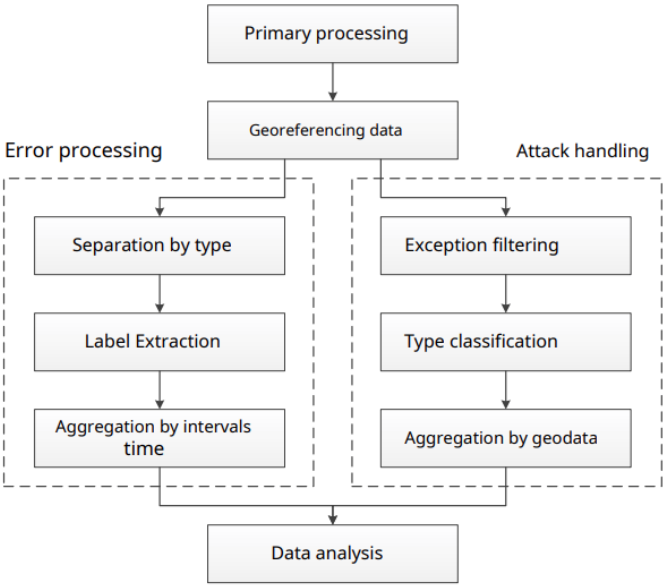


Figure 1. Stages of Data Processing

3. Result and Discussion

Analysis of the total number of attacks for 2020-2022 shows that the daily number of attacks varies slightly on average: 3664 in 2020, 3481 in 2021 and 3698 in 2022 (a 3% deviation from the annual average). At the same time, the maximum number of attacks varies greatly. At the same time, the maximum number of attacks varies widely, from 8,500 to 21,000 per day, which indicates the simultaneous operation of several uncoordinated sources. Figure 2 shows the overall dynamics of detected attacks by month for 2020-2022. We note a lack of pronounced periodicity in both daily and monthly totals.

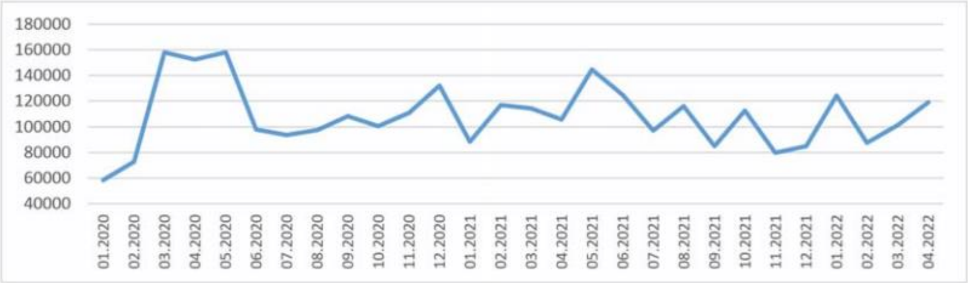


Figure. 2 General Dynamics of Attacks by Month

DOI 10.29207/joseit.v1i2.4944

Analyzing attacks by species, clear trends can be seen in the increase in the number of attacks by individual species. Fig. 3 shows the monthly number of POLICY/EXT_RESTR (forbidden extension) and WEB/FILE_INJ (file injection) attacks, which clearly show a two-fold increase in intensity. The number of these attacks clearly shows a 2-fold or more increase in intensity, which is not visible in Fig. 2.

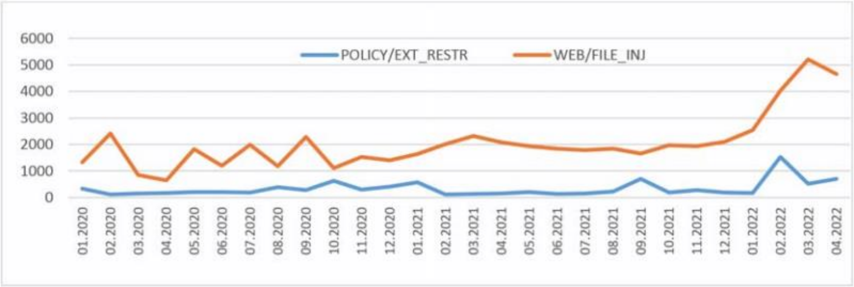


Figure 3. Dynamics of Classified Attacks by Month

The aggregation chart by day for 2022 (Fig. 4) does not show the upward trend noted in Fig. 3. The upward trend noted in Fig. 3 is not visible. We can see a separate peak around February 26, 2022 that can be associated with massive cyber-attacks on Russian internet resources. Therefore, the most effective way to detect an increase in the risk of cyber threats is to analyze classified threats with an aggregation of up to one month. Fig. 5 shows distribution charts for the distributions for 2022, 2021 and 2020: unclassified attacks (a), WEB/CMD_INJ (command injection) attacks (b) and WEB/FILE_INJ (file injection) attacks (c).

While the unclassified distribution of attacks has similar parameters for different years, when applying the classification, the distribution parameters change quite significantly, especially for the year 2022, which is characterized by an increase in the risk of cyber threats. In order to select a set of indicators, a correlation matrix of their distributions by days for the entire observation period 2020-2022 was constructed (Figure 6). Most indicators show weak correlation, except for WEB/CMD_INJ (command injection), WEB/DIR_TRAVERSAL (directory search) and WEB/JAVA_INJ (java code injection), which can be aggregated.

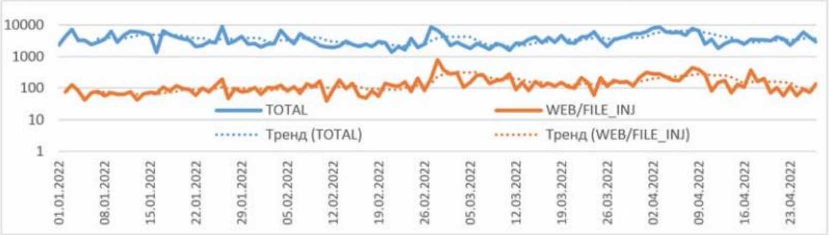
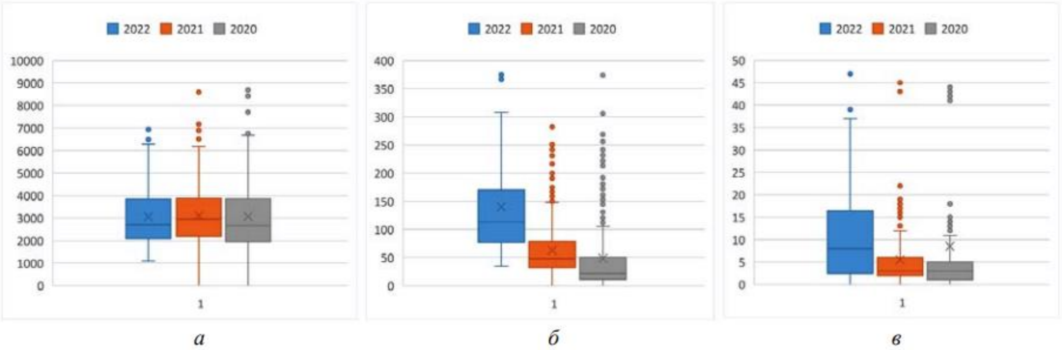


Figure 4. Dynamics of Unclassified Attacks and File Injection Attacks



a *б* *в*

Figure 5. Range diagrams of attacks by year: a – unclassified; b – WEB/CMD_INJ; c – WEB/FILE_INJ

Based on our analysis, we can see that individual classified attack types contain more information about risk dynamics than unclassified ones. By selecting independent classified attack types and calculating their statistical indices for time samples, the following heuristic method for assessing changes in cyber threat risks, based on comparing parameters of sample distributions. For the samples V1 and V2, which contain N independent indicators, we introduce the following risk variation estimation function R:

DOI 10.29207/joseit.v1i2.4944

$$R(V_1,V_2) = \frac{1}{N}\cdot\sum_{i=1}^{N} K_i,$$

$$\text{где} \quad K_i = \begin{cases} 1, \text{ если } \mu_i > 0{,}6745\cdot\sigma_i, \\ 0, \text{ если } -0{,}6745\cdot\sigma_i \le \mu_i \le 0{,}6745\cdot\sigma_i, \quad \mu_i \\ -1, \text{ если } \mu_i < -0{,}6745\cdot\sigma_i; \end{cases}$$

Sample mean value of the i-th attribute of sample V_2; σi - standard deviation of sample i-го feature of sample V_1. According to the proposed procedure, if the mean of all N characteristics of the sample V2 is greater than the third quartile of the sample V1, then the value of the change in risk is 1, which can be interpreted as a significant increase in risk for all indicators. If the mean value of all N characteristics of sample V2 is less than the first quartile of sample V1, then the value of the change in risk is -1 (a reduction in risk for all indicators). The value of R(V1,V2) [-1,1], which allows you to use this indicator for analyzes using artificial intelligence methods, in particular the Shortliffe method.

IP georeferencing information was extracted from logs and attack sources were analyzed by type. Correlations of the 2020 and 2021 samples by attack type were calculated for countries in the top 15 by attack intensity. Assuming that the correlation of attack indicators of different types (attack profile) is determined by the software used to conduct an attack, then the high correlation of such samples for a country at different time periods can be interpreted as a fixed set of software, used for the attack software (the attacked vulnerabilities). The resulting chart in Figure 7 shows that the countries with the most stable attack patterns are China, Russia, Germany, the UK, the United States and Poland.

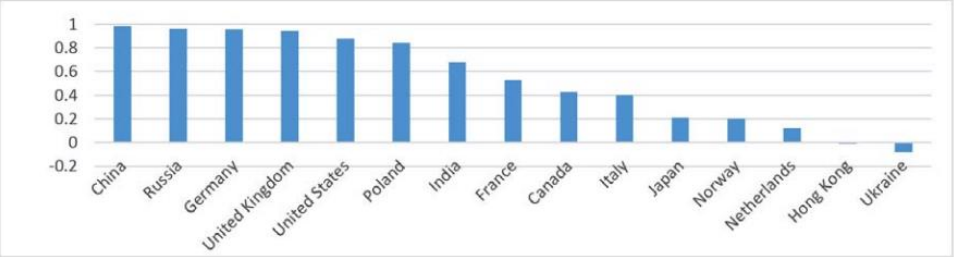


Figure 7. Correlation of Attack Patterns in 2020 and 2021 by Country

Low correlation countries (Holland, Hong Kong, Ukraine) do not have a consistent set of attack software. Countries with low correlation (Holland, Hong Kong, Ukraine) do not have a consistent set of attack software and are likely to be used by different groups of attackers Controlling botnet networks.

4. Conclusion

The paper examines the dynamics of attacks on web services by country and highlights the main groups of countries with a constant attack profile and high intensity. Pairwise correlations of different attack types were compared and the attacks with high correlation were identified, which can be aggregated when assessing the risks. The method of comparing cybersecurity risks for different time periods using attack type classification. The method does not depend on the time intervals to be compared and the sample size, since it is based on statistical indicators. The cybersecurity risk assessment method can be used in other areas where there is a classification of indicators.

References

[1] Landauer M., Skopik F., Wurzenberger M., Rauber A. System log clustering approaches for cyber security applications: A survey. Computers & Security. 2020, Vol. 92, P. 101739.

[2] He P., Zhu J., He S., Li J. et al. Towards Automated Log Parsing for Large-Scale Log Data Analysis. IEEE Transactions on Dependable and Secure Computing. 2017, Vol. 15, No. 6, P. 931–944.

[3] Moh M., Pininti S., Doddapaneni S., Moh T. Detecting Web Attacks Using Multi-stage Log Analysis. IEEE 6th International Conference on Advanced Computing (IACC). 2016, P. 733–738.

[4] Zhu J. et al. Tools and Benchmarks for Automated Log Parsing. IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP). 2019, P. 121–130.

[5] Efimova Yu. V., Gavrilov A. G. [Modeling an information security system based on the analysis of system logs]. Inzhenernyi vestnik Dona. 2019, No. 6 (57), P. 40 (In Russ.).

[6] Bolodurina I. P., Parfenov D. I., Zabrodina L. S. et al. [Modeling the identification of a cyber attack profile based on the analysis of the behavior of devices in the network of a telecommunications service provider]. Vestnik Yuzhno-Ural'skogo gosudarstvennogo universiteta. 2019, No. 4, P. 48–59 (In Russ.).

[7] He P., Zhu J., Zheng Z., Lyu M. R. Drain: an online log parsing approach with fixed depth tree. Proc. of the International Conference on Web Services (ICWS). IEEE, 2017, P. 33-40.

[8] Reidemeister T., Jiang M., Ward P. A. Mining unstructured log files for recurrent fault diagnosis. Proc. of the Int. Symp. on Integrated Netw. Mgmt. IEEE, 2011, P. 377–384.

[9] Sidorova D. N., Pivkin E. N. [Algorithms and methods of data clustering in the analysis of information security event logs]. Bezopasnost' tsifrovykh tekhnologii. 2022, No. 1 (104), P. 41–60 (In Russ.).

[10] Juvonen A., Sipola T., Hamalainen T. Online anomaly detection using dimensionality reduction techniques for http log analysis. Computer Networks. 2015, No. 91, P. 46–56.

[11] Wurzenberger M., Skopik F., Landauer M., Greitbauer P., Fiedler R., Kastner W. Incremental clustering for semi-supervised anomaly detection applied on log data. Proc. of the 12th International Conference on Availability, Reliability and Security, ACM (2017), P. 31:1–31:6.

[12] Aharon M., Barash G., Cohen I., Mordechai E. One graph is worth a thousand logs: uncovering hidden structures in massive system event logs. Proc. of the Joint Eur. Conf. on Machine Learning and Knowledge Discovery in Databases. Springer, 2009, P. 227–243.

[13] Jia T., Yang L., Chen P., Li Y., Meng F., Xu J. Logsed: anomaly diagnosis through mining time-weighted control flow graph in logs. Proc. of the 10th Int. Conf. on Cloud Comp. (CLOUD). IEEE, 2017, P. 447–455.

[14] Kononov D., Isaev S. Analysis of the dynamics of Internet threats for corporate network web services. CEUR Workshop Proceedings. The 2nd Siberian Scientific Workshop on Data Analysis Technologies with Applications 2021. 2021, Vol. 3047, P. 71–78.

[15] Helmiawan M. A., Firmansyah E., Fadil I., Sofivan Y., Mahardika F. and Guntara A. Analysis of Web Security Using Open Web Application Security Project 10. 8th International Conference on Cyber and IT Service Management (CITSM). 2020, P. 1–5.

[16] OWASP ModSecurity Core Rule Set. Available at: https://owasp.org/www-projectmodsecurity-core-rule-set/ (accessed: 13.05.2022)

DOI 10.29207/joseit.v1i2.4944