Analysis of Internet Service Log Data to Assess the Level of Cyber-threats in the Corporate Network*

Sergey Isaev[0000-0002-6678-0084], Dmitry Kononov[0000-0002-8757-5274]
and Andrey Malyshev[0000-0001-5669-1574]

Institute of Computational Modelling of the Siberian Branch
of the Russian Academy of Sciences, 50/44 Akademgorodok, Krasnoyarsk, 660036, Russia
ddk@icm.krasn.ru

Abstract. The article describes log analysis of Internet services of the Krasnoyarsk Science Center (Russia). The importance of log analysis as a method to improve the effectiveness of network security is shown. Data sources are described. The study examines the following systems: Netflow IP traffic, intrusion prevention system, corporate mail server, web server. The log data was used to distinguish the frequency of events and to identify malicious behavior. The article describes security threats identified during the analysis of logs. The analysis results allow optimizing protection systems against network attacks. Measures taken to improve network security are presented.

Keywords: Cyber-Threats, Security, Data Analysis, Log, Internet.

1 Introduction

Development of modern information technologies leads to increasing digitalization level and active use of various Internet services for scientific and business processes. The corporate network and services it provides become daily working tools, without which the full functioning of the organization is impossible. In this regard, the tasks of assessing the risks of cybersecurity and the level of cyber-threats for providing adequate protection are becoming more and more relevant. An important aspect of cybersecurity is the study of security logs [1]. Modern researchers use dynamic methods of analysis since traditional approaches with static metrics may skip intellectual low-frequency attacks [2]. An important parameter of a secure system is the response time to information security incidents. Minimization of this parameter to the extent of full attack prevention is described in [3]. Methods and algorithms of mail spam resistance are actively developing [4]. To provide information security various software is used: security scanners [5], complex security analysis systems [6], etc. Thus, revealing new cyber-threat signs and analysis methods is an urgent problem.

For many years, Krasnoyarsk Science Center has been studying the problems of cyber security analysis and ensuring the network protection [7]. The purpose of this work is to analyze log data on Internet services, identify potential risks, and optimize security protection mechanisms.

2 Data Sources

The corporate network of the Krasnoyarsk Science Center has a four-level architecture: 1) network core which provides routing and connectivity to external networks; 2) server network which hosts Internet services; 3) aggregation level which connects multiple organizations together; 4) local network for end users. All the information about network traffic is collected at the Internet connection points and server network. In addition, there are logs from the main Internet services: corporate mail server, web server, and proxy server for access to external resources. The sources of data analysis are the following:

1. Netflow IP traffic: more than 400 GB, more than 2 billion records.
2. Mail server log: 1 year, more than 1.5 million records.
3. Intrusion Prevention System (IPS) log: 1 year, about 200 thousand records.
4. WWW log: 1 year, about 12 GB, more than 44 million records.

3 IP Traffic Data Analysis

To assess the permanent threat level, IP traffic to unused network addresses was analyzed (Fig. 1).

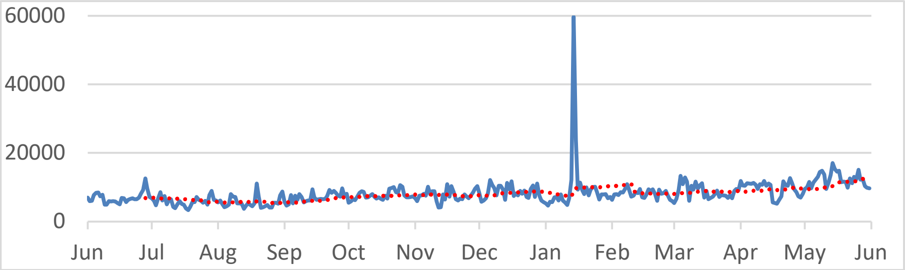

Fig. 1. Incoming connections to unused network addresses (daily, June 2019 – June 2020).

The analysis shows the permanent number of access attempts. The daily aggregation shows a trend in the number of access attempts from 5000 to 12000 per day. The detected peak of 60000 on 14.01.2020 is explained by the attack duration rather than intensity, as seen in the hourly aggregation (Fig. 2).

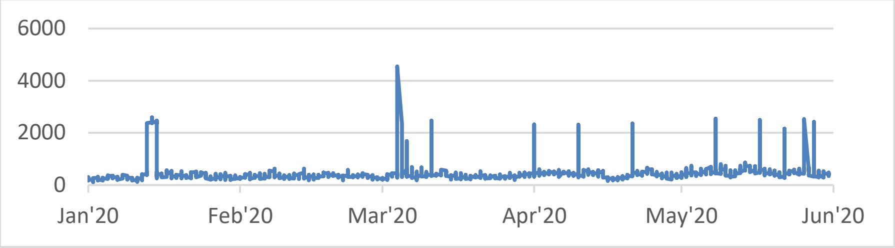

Fig. 2. Incoming connections to unused network addresses (hourly).

The hourly rate of access attempts to a single address is about 500 per hour, regardless of the time of day and days of week, with the peaks of up to 4500 per hour. Thus, the time interval for confident detection of network attacks should be no more than one hour. The number of unique threat sources per day (Fig. 3) changes from 1000 to 2000, which indicates the presence of a large and constantly operating network used for scanning Internet services.

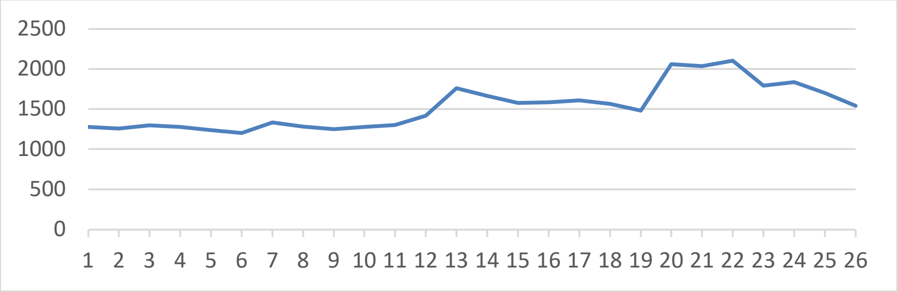

Fig. 3. Number of unique scan sources per day for April 2020.

Building the distribution by hours (Fig. 4) allows one to calculate the average deviation of about 1.5% with the maximum deviation around 07:00 KRAT of about 5%. The detected maximum corresponds to 00:00 GMT time, which indicates the prevalence of the systems with scheduled scans and attacks launched at midnight GMT.

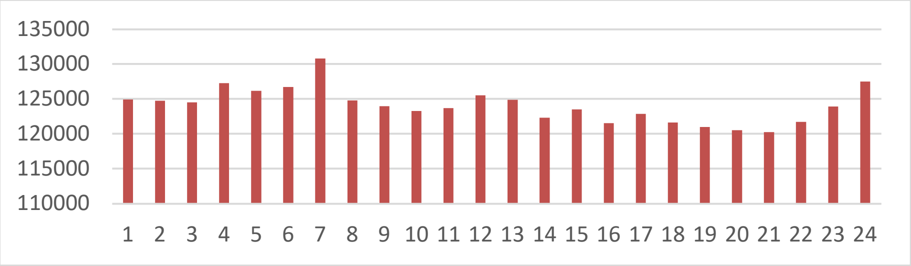

Fig. 4. Distribution of scans by hours.

The analysis of scanning frequency of individual services (Fig. 5) allows identifying popular services, which are the most attacked and under which the threats are masked: Telnet/23, MS SQL/1433, HTTP/80, Personal Agent/5555, SSH/22, HTTP Alternate/8080, RDP/3389. Thus, Telnet and MS SQL can be added to the existing blocking network ports (SSH, RDP, SMTP) to increase the protection efficiency.

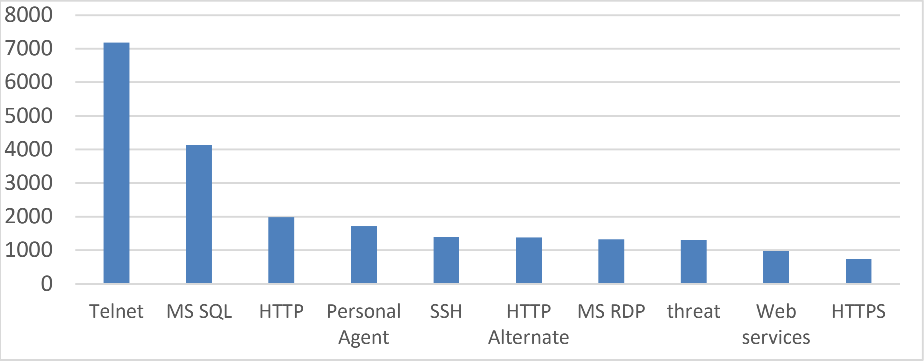

Fig. 5. Scan rates by service.

4 Mail Server Data Analysis

The analysis of the mail traffic reveals its periodicity during both the day and days of week (Fig. 6).

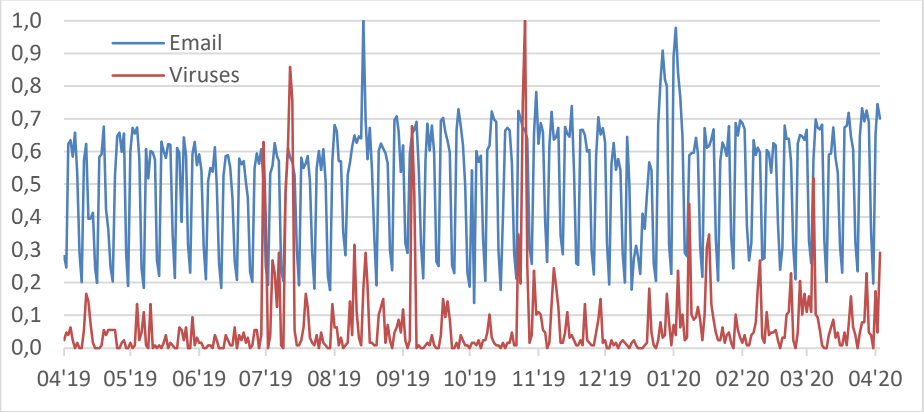

Fig. 6. Normalized daily number of emails and viruses detected during the year.

The number of viruses in the mail by hours in the recipient's time zone (Fig. 7) shows a good correlation with the number of mail spam (0.63), but time distributions by the sender have zero correlation, which may indicate different sources of mail spam and mail viruses.

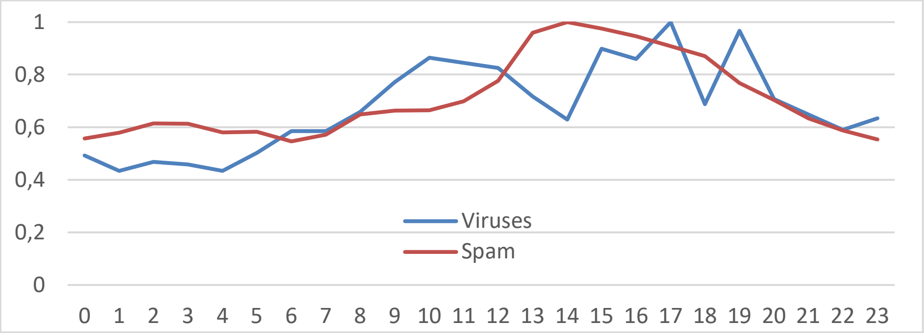

Fig. 7. Distribution of spam and viruses by the recipient's time zone.

Using geographical databases and data aggregation by the territory allows building a distribution by the country of spam sources (Fig. 8). The most active mail spam countries: Russia, USA, Germany, France, China.

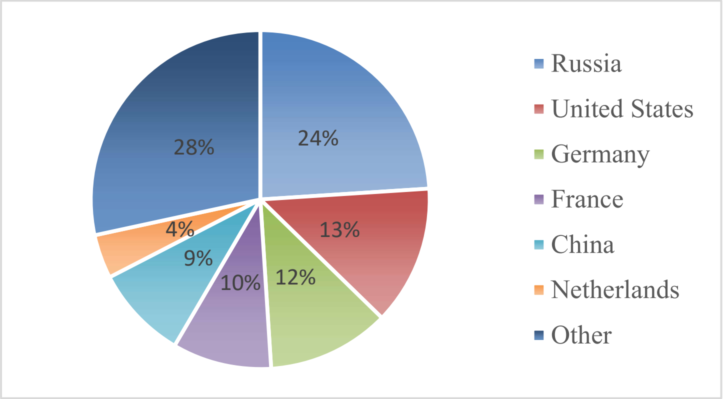

Fig. 8. Leading countries of mail spam.

In terms of the number of virus sources, the United States, France, Russia, and Vietnam are the leaders. The weekend activity and threat level is 2-3 times lower than on workdays, while Tuesday and Wednesday are the highest threat level days. In terms of the time of day, the threat level at night is 2 times lower than during working hours.

5 Intrusion Prevention System Data Analysis

Analysis of the Intrusion Prevention System log shows no periodicity both with regard to the days of week and time of day (Fig. 9). The number of blocked network addresses is approximately 3 times lower than the number of unique scanning sources per day (on average, 500 and 1500, respectively), which indicates that only every third source takes actions leading to its blocking.

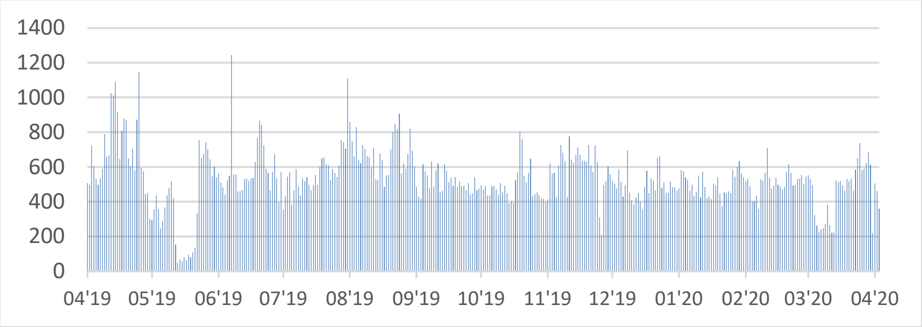

Fig. 9. Histogram of the number of blocks during the year.

In the hourly frequency distribution of the blocking system over the SSH and RDP protocols (Fig. 10), there is a peak (about 150% of the average) at around midnight GMT, which is an additional indicator of a large threat scanning network launched at 00:00 GMT.

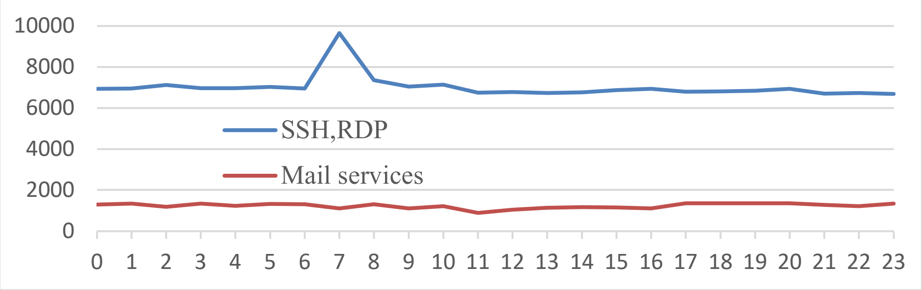

Fig. 10. Hourly distribution of IPS responses.

The largest number of responses (165 thousand) falls on the SSH service, followed by SMTP (about 30 thousand) and RDP (about 3 thousand). The significant predominance of SSH may be due to the specific of the blocking system: all connections from the threat sources are blocked, and SSH has the minimum port number (22) among popular services and, thus, is checked first. The analysis of the geographical location of threat sources shows the leadership of China (39%) and the United States (12%). Thus, it is possible to make a conclusion about purposeful invasion attempts from China, since other data indicates that it does not have a leading position.

6 WWW Data Analysis

The analysis of web services logs shows periodicity both in terms of the days of week and time of day. In addition, there is a tendency for the number of requests associated with the WWW services expansion to increase and an increasing number of visitors is also observed. The request analysis in terms of the country shows the following results: Russia – about 80% of all visits, USA – 7%, Germany – 2%. The analysis of error logs shows the following: Russia – 54%, USA – 27%, China – 3%. The most popular browsers are Chrome – 47%, Firefox – 16%, Internet Explorer – 6%. Web spiders and bots amount to about 9% of the total number of requests and 32% of the total number of errors.

When processing web service logs, requests were divided into two non-intersecting groups: legitimate and erroneous requests. Legitimate requests are those that are processed by a web application or web service in normal mode and whose HTTP response code is one of 1XX, 2XX, 3XX. Erroneous requests (or errors) are those that are processed incorrectly either on the client side (HTTP response code 4XX) or on the server side (5XX). The error analysis is important because it allows revealing malicious activity. Figures 11 and 12 show the number of requests and errors per day. As one can see from the graphs, the number of requests and errors depends on holidays since most of the services are used during business hours.

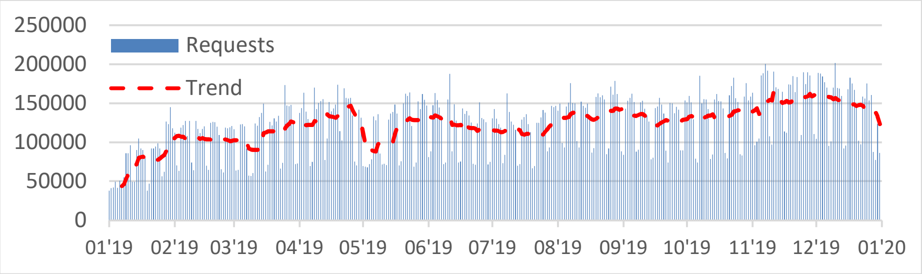

Fig. 11. Number of requests per day during the year.

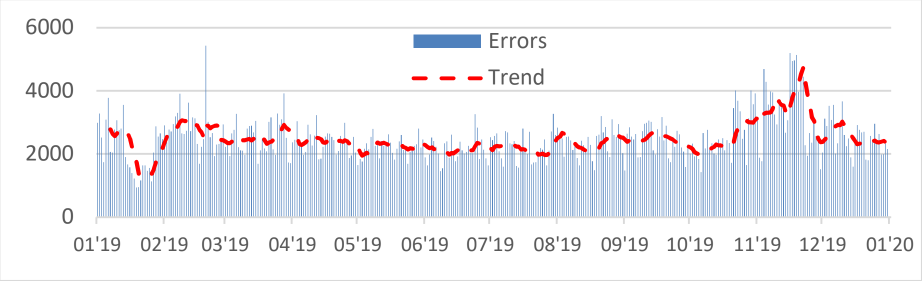

Fig. 12. Number of errors per day during the year.

The correlation coefficient for the server and client requests is 0.984 (Fig. 13). This indicates that most of the requests are carried out from Krasnoyarsk (KRAT) and nearby time zones.

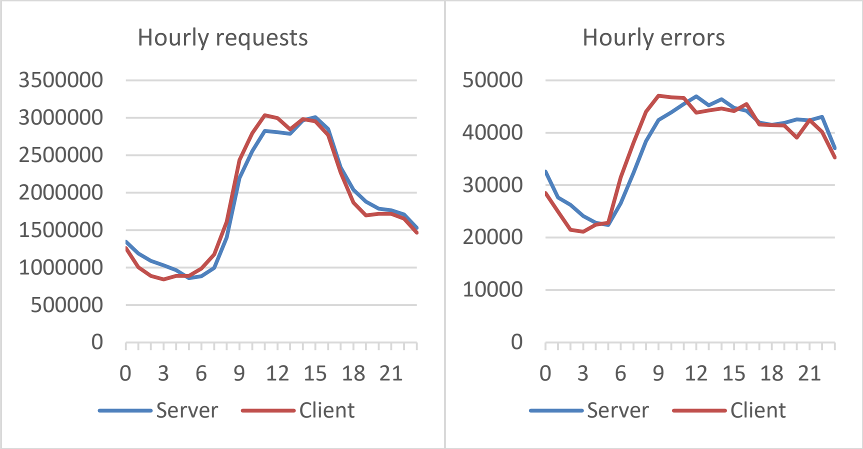

Fig. 13. Number of requests and errors per hour.

The correlation was calculated and queries and errors were normalized by server and client time (Fig. 14). The high value of the correlation coefficient for errors was found to be caused by incorrect operation of websites and web services. However, the rest of the errors show the presence of scans and attacks performed by web spiders.

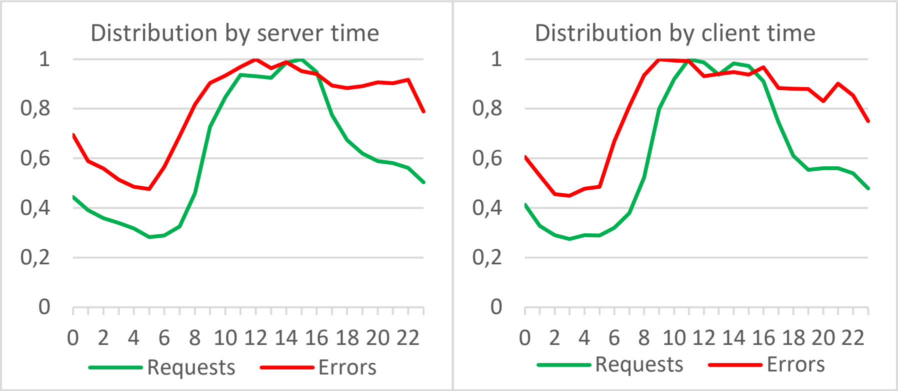

Fig. 14. Distribution of requests and errors by server and client time.

Since errors indicate the attempts to access non-existent or non-public resources, there is a high probability of an increasing number of threat sources from the United States and China since their portion of errors is several times greater than the portion of requests. The majority of errors are caused by detecting vulnerabilities in popular Content Management Systems (CMS). In addition, the browsers analysis in terms of requests and errors shows an increased percentage of errors from web spiders, which may indicate a high risk of threat.

7 Measures Taken

As a result of the research, some measures were taken to increase the security of Internet services of the Krasnoyarsk Science Center. In particular, the following was performed: 1) the threshold time interval for more confident detection of network scanning was increased; 2) new TCP ports to the monitoring system to track malicious activity were added; 3) firewall settings to more effectively blocking unwanted hosts were optimized; 4) web server settings to prevent attacks on the known CMS vulnerabilities was updated; 5) network settings of internal switches were optimized to block unwanted traffic between different divisions.

8 Conclusion

In this work, we analyzed data logs from the corporate Internet services of the Krasnoyarsk Science Center. The main sources of cybersecurity threats were identified. New signs of threat sources were determined which can be used to improve corporate network security systems. In general, the applied security tools allow detecting and blocking threats at early stages. The results of the study allow optimizing protection systems against network attacks, taking into account the identified sources of threats which were previously not taken into account in standard security tools. The measures taken increase the responsiveness to emerging threats and cybersecurity of the organization as a whole.

References

1. Khan, S., Parkinson, S.: Discovering and utilising expert knowledge from security event logs. Journal of Information Security and Applications **48**, 102375 (2019)
2. Landauer, M., Wurzenberger, M., Skopik, F.: Dynamic log file analysis: An unsupervised cluster evolution approach for anomaly detection. Computers & Security **79**, 94–116 (2018)
3. Kim, D., Kim, Y.-H., Shin, D., Shin, D.: Fast attack detection system using log analysis and attack tree generation. Cluster Computing **22(2)**, 1827–1835 (2019)
4. Mujtaba, G., Shuib, L., Raj, R. G., Majeed, N., Al-Garadi, M. A.: Email Classification Research Trends: Review and Open Issues. IEEE Access **5**, 9044–9064 (2017)
5. Zhang, K., Zhao, F., Luo, S., Xin, Y., Zhu, H.: An Intrusion Action-Based IDS Alert Correlation Analysis and Prediction Framework. IEEE Access **7**, 150540-150551 (2019)
6. Sapegin, A., Jaeger, D., Cheng, F., Meinel, C.: Towards a system for complex analysis of security events in large-scale networks. Computers & Security **67**, 16–34 (2017)
7. Kulyasov, N., Isaev, S.: Research of network anomalies in the corporate network of the Krasnoyarsk Scientific Center. Siberian Journal of Science and Technology **19(3)**, 412–422 (2018)