Analysis and forecast of undesirable cloud services traffic

Marina V. Tumbinskaya
E-mail: tumbinskaya@inbox.ru

Bulat I. Bayanov
E-mail: bayanov_bulat@mail.ru

Ruslan Zh. Rakhimov
E-mail: rahimov96@mail.ru

Nikita V. Kormiltcev
E-mail: kormiltcev@hotmail.com

Alexander D. Uvarov
E-mail: obg-96@mail.ru

Kazan National Research Technical University named after A.N. Tupolev
Address: 10, Karl Marx Street, Kazan 420111 Russia

Abstract

These days one of the main problems that must be solved to ensure information security in cloud services for corporations as well as for individual clients is to correctly identify and predict hacking in the network traffic. This paper presents statistics on information security threats, provides classification of information security threats for cloud services, identifies hackers' goals, and proposes countermeasures.

A vital task is to develop an effective method that could be used to protect cloud services from various network threats, as well as to analyze the network traffic. For these purposes, we chose a method based on an additive time series model, which allows us to predict the undesirable network traffic. To test this method, we obtained quantitative parameters for the undesirable traffic by simulating a network attack and collecting empirical data that describe this process. We used special software that simulates a network attack, and software that records and processes all the empirical data needed for the research.

Using the data obtained, we analyzed the efficiency of the method based on the additive time series model. We demonstrated that this method is also applicable for research into the general dynamics of the number of network attacks in cyberspace. This method also allows us to reveal how the dynamics of the number of hacker network attacks depends on season, date, or time. The results show that, based on data describing the network traffic, one can identify and predict the undesirable hacker threats.

Key words: forecasting; DDOS attack; cloud services; network traffic; modeling; additive time series model; autocorrelation function; error estimation.

Citation: Tumbinskaya M.V., Bayanov B.I., Rakhimov R.Zh., Kormiltcev N.V., Uvarov A.D. (2019) Analysis and forecast of undesirable cloud services traffic. *Business Informatics*, vol. 13, no 1, pp. 71–81

DOI: 10.17323/1998-0663.2019.1.71.81

Introduction

Development of the infrastructure of modern enterprises causes an increased demand for cloud technologies, because they are convenient, reasonably priced, mobile, quick, and reliable. Cloud technologies allow us to use cloud services. A cloud service [1] is an Internet service that makes it possible for its clients to outsource the maintenance of some elements of IT infrastructure [2].

According to the RightScale statistics, 95% of organizations used one or another cloud service model in 2017 [3]. According to Orange Business Services experts, the market for cloud services in Russia comes to about 24.6 Bn. roubles [4]. It was shown [2, 5] that modern IT companies are uncompetitive if they fail to use cloud technologies, thus foregoing profits. Cloud services have long been used in large corporations (Google Disk, iCloud from Apple, Cloud mail.ru).

Cloud services make it necessary to solve information security problems, since new technologies lead to the emergence of a large number of threats and vulnerabilities in information security systems. According to a Kaspersky Laboratory poll [6], 13% of Russian organizations face issues related to cloud infrastructure security at least once a year. Out of those companies, 32% lost their data due to such incidents. Therefore, it is crucial to ensure information security in cloud services.

The proposed novel method for analyzing and predicting the network traffic based on the additive time series model and integrated into security tools can ensure the necessary security level for regular storage, thus protecting it from various network attacks. This constitutes the scientific novelty of the paper. Unfortunately, many existing data security methods cannot reliably predict undesirable network traffic.

1. Possibility of interpreting the proposed method in WAF

As shown in [7], the majority of hacker attacks are based on typical hacker methods, which are brought to perfection. Therefore, we need to develop methods that employ continuous learning, and such methods should gradually replace the signature analysis. It was also noted [7] that some developers of web application firewalls (WAF) focus on renewing the signatures rather than on the signature analysis. To create a security model that ensures the necessary security level, WAF needs an extensive database of the undesirable traffic signatures and actions that can be applied to all types of web applications. The proposed method for analyzing and predicting the network traffic based on the additive time series model can be integrated into complicated WAFs in the future. Here the main goal will not be to predict the hacker’s and legitimate user’s actions, but to create a security model based on the URL as well as on the parameters and cookies. After the security model is developed, it needs to be tested, i.e., the traffic should be analyzed to prevent a hacker’s exploiting both known and unknown vulnerabilities.

2. Classification of security threats to cloud services

Let us consider the classification of security threats to cloud services. *Table 1* presents the most common threats according to [6]. Possible hackers' goals and security measures are presented for each threat. No single method alone can prevent all types of threats; therefore, it is impossible to block the threats completely. Statistical data for each threat that was successful can be stored in the system and used for future analysis and development of new security systems.

3. Simulating a network attack

To analyze the network traffic coming into the network nodes, information security specialists install ad hoc software at the network nodes. In this research, Wireshark software (v.2.6.1) was used. This software allows us to capture and analyze the network traffic for the most common network protocols (TCP, UDP, HTTP, etc.).

Papers [8, 9] contain the network traffic data that describe DDOS attacks. However, there is not enough data there for the purposes of this research; therefore, we collected the necessary data by simulating a network attack based on the algorithms presented in [10]. We used two nodes of a configured network. One of them was used as the victim's device, and the other one as the hacker's device. Virtual machines installed on the same computer served as those devices. Wireshark was installed on the victim's virtual machine, and LOIC (an open-source code for DDOS attacks)[1], which creates undesirable traffic, was installed on the hacker's virtual machine.

In our research, we assumed that hackers attack the network multiple times with various initial configurations of the malware, and we did not rule out the possibility that the victim could access the network as well. The network stream (the number of network packets per second) through the victim's network node is presented in *Figure 1.*

Table 1.

Types of security threats to cloud services, hackers' goals, and security measures

#	Security threat to cloud services	Hackers' goal	Security measures
1.	Data theft	Accessing a database (e.g., e–mail addresses of users)	Database decentralization and data encryption with an SSL certificate
2.	Data loss	Database modification or erasing information	Data backup, restricted access
3.	Account theft / hacked services	Database modification or erasing information	Two–factor authentication
4.	Unprotected nterfaces and API	Complete access to the database	Authentication, access control, encryption
5.	DDOS attacks	Preventing authorized users from accessing the cloud service	Access control
6.	Undesirable insider	Database access	Access control
7.	Cloud services used by hackers	Access to the cloud computing resources	Restriction of the system's computing power

[1] https://www.darknet.org.uk/2017/10/loic-download-low-orbit-ion-cannon-ddos-booter/

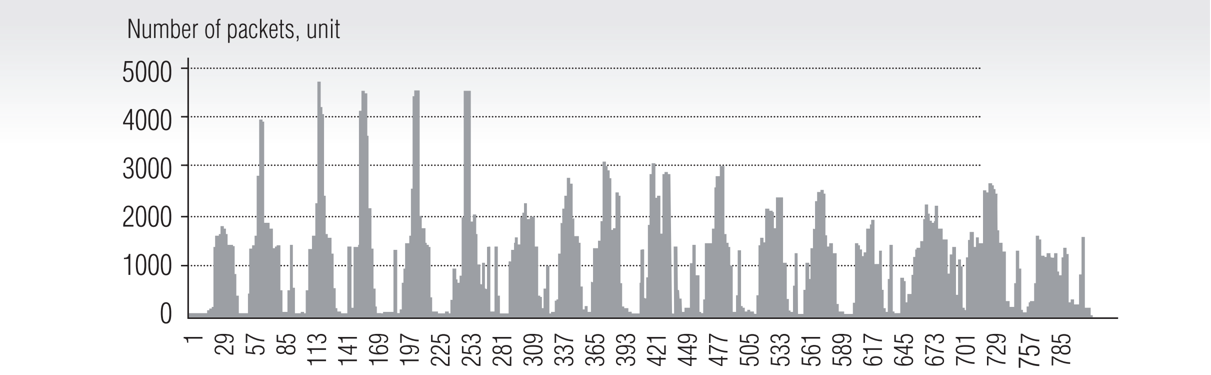

Fig. 1. Number of network packets passing through the victim's node, per second

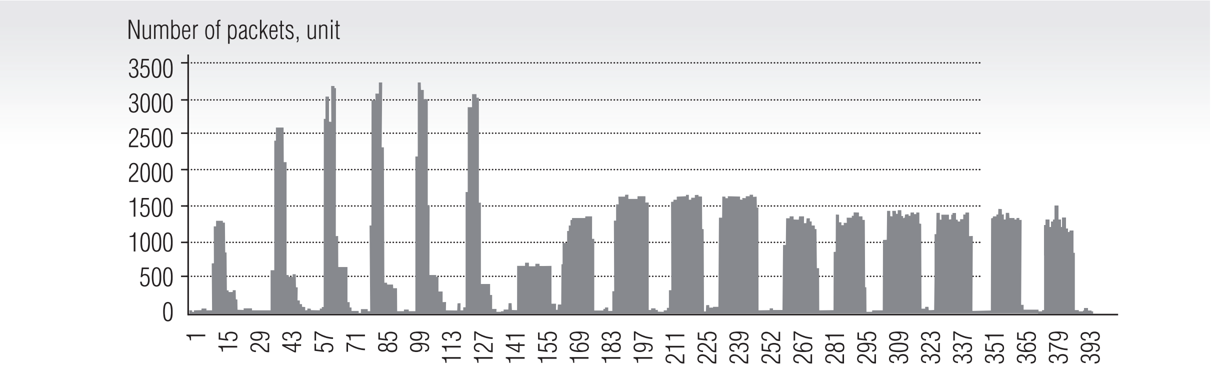

Fig. 2. Number of network packets from the hacker's node, per second

As we can see from *Figure 1*, we cannot distinguish the stream from a particular user from the overall network stream; therefore it is recommended to consider network streams from particular users.

For convenience of analysis, it is possible to filter the network stream through the victim's node and separate the packets coming from the hacker's node. The network stream from the hacker's node is presented in *Figure 2*. It represents a physical process, the intensity of which periodically increases by several orders of magnitude and describes the actions of a particular user.

We can find out the address of the node of the hacker attacking the network by analyzing the parameters describing the incoming network traffic, for example, the density of the distribution of the number of packets by their size in bits. *Table 2* presents the data obtained by simulating network attacks and desirable network traffic using Wireshark software.

The simulation results show that, when the network traffic is undesirable, more than 92% of the packets are 40–79 bits in size. At the same time, when the traffic is desirable, the percentage of the packets of this size is about 39%, while more than 42% of the packets have the size between 1280–2559 bits, and about 11% are sized between 640–1279 bits. It is also suspicious when the traffic is extremely intensive (in terms of the number of packets per unit of time) or it has other untypical parameters. As a sample for analysis, we chose the number of network packets coming from the hacker's node per second (*Figure 2*).

Table 2.

The percentage of received packets by packet size for desirable traffic and during a network attack

#	Packet size	Percentage of packets received by packet size for desirable traffic	Percentage of packets received by packet size during a network attack
1.	0–19	0.00%	0.00%
2.	20–39	0.00%	0.00%
3.	40–79	39.06%	92.79%
4.	80–159	3.81%	0.48%
5.	160–319	0.93%	3.30%
6.	320–639	1.35%	3.25%
7.	640–1279	11.05%	0.16%
8.	1280–2559	42.90%	0.02%
9.	2560–5119	0.82%	0.00%
10.	5120 and more	0.08%	0.00%

4. Predicting the network attacks with time series analysis

For statistical analysis, we chose the method based on time series analysis. According to the Cisco annual report on cybersecurity [11], in year 2018, 39% of organizations used automated tools to prevent hacker attacks, and the rest of them used machine learning (artificial intelligence) [12–15].

We solved the problem of predicting the network attacks using a time series additive model. This model assumes that each level of the time series (F) can be presented as a sum of three components: a trend (T), a seasonal component (S), and a random component (E):

$$F = T + S + E. \tag{1}$$

To determine the trend component, linear regression was used:

$$y = a \cdot x + b, \tag{2}$$

where y – the trend value;

x – the lag;

a and b – the regression coefficients.

In Equation (2), coefficients a and b are determined from the previous values in the original sample using the following equation:

$$b = \frac{\sum(x - \overline{x})(y - \overline{y})}{\sum(x - \overline{x})^2}, \tag{3}$$

$$a = y - b \cdot x, \tag{4}$$

where $\overline{x}$ – the mean lag value;

$\overline{y}$ – the mean value in the original sample.

Figure 3 presents the original data on the number of network packets coming from the hacker's node alongside the trend line obtained by Equation (2), where $a = 0.973$, $b = 615.87$. The trend line goes up because of the increase in the intensity of the network traffic.

Now we have to determine the seasonal component, which is periodical and can be obtained from the autocorrelation function (ACF). *Figure 4* presents a plot of the auto-

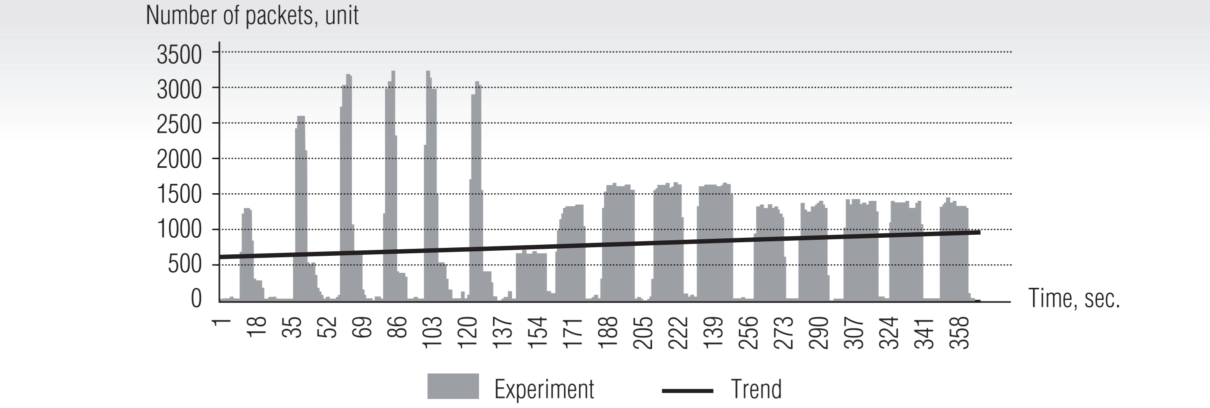

Fig. 3. The number of network packets received from the hacker's node per second, including the trend component

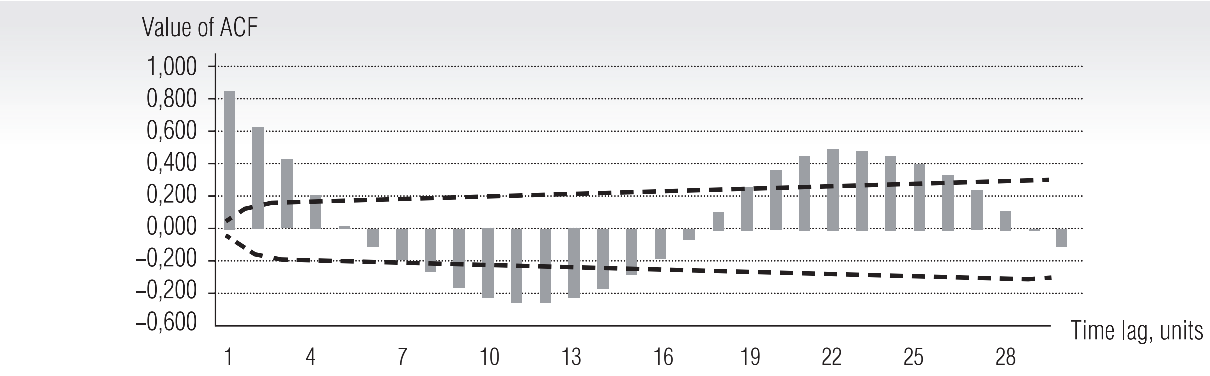

Fig. 4. Autocorrelation function with the account of the white noise

correlation function of. the lag number. The dashed line corresponds to the white noise (the boundary of the statistical significance of correlation coefficients is the error of the autocorrelation function). The autocorrelation function was calculated for a time interval of up to 30 lags.

Analysis of the autocorrelation function shows that the original data is periodical. There is a high correlation for lags 22 and 23. Therefore, for the seasonal component in the additive model the period will be about 23 lags. Thus, the length of one season is $N = 23$ (the lag's number can take values $n = 1, 2, \ldots, N$), where one lag corresponds to one second.

The values of seasonal component S_n are determined as mean values of the differences between current the value F_n and the trend component T_n calculated for each lag number n:

$$S_n = \sum_{k=1}^{K} \frac{(F_n)_k - (T_n)_k}{K}, \qquad (5)$$

where k – the season number;

K – the total number of seasons.

Then the total number of lags for the entire time series is $M = N \cdot K$.

Using the values obtained for the trend component (2) and the seasonal component (5), we can calculate the predicted values for F using Equation (1) (in this model, the random component is not considered) [16]. *Figure 5* presents the plots for sample values F_n and predicted values F. The discrepancies between the plots for F and F_n can be evaluated by calculating the mean absolute percentage error (MAPE).

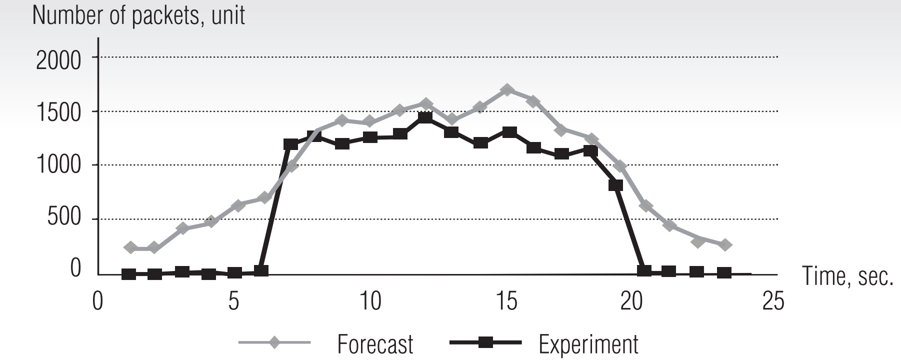

Fig. 5. Current values for the test sample and predicted values for the number of network packets per second

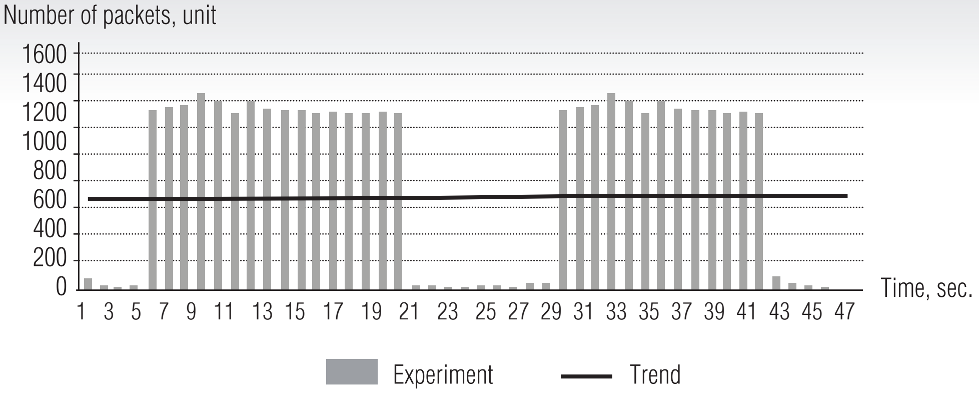

Fig. 6. The number of network packets coming from the hacker's node per second, with the account of the trend component for two closest seasons from the original sample

This estimate cannot be used to calculate the error of the prediction model used in this research because the current test sample includes values close to 1. Therefore, we used the root-mean-square error (RMSE) instead; it is equal to 353. This follows from the following equation:

$$\mathrm{RMSE}=\sqrt{\frac{1}{N}\sum(y-\hat{y})^2}, \qquad (6)$$

where N – the original sample size;

y – the predicted value,

$\hat{y}$ – the current value.

The value obtained indicates that the prediction model is less than optimal. To make our additive predicting model more accurate, we used the trend and seasonal components from earlier seasons (the most recent ones), as shown in *Table 3* and *Figure 6.*

Both the amplitude and the duration of these seasons are close to the corresponding values for the following season, and this fact can improve the quality of the prediction.

Figure 7 presents the plots for future actual values for the test sample and the predicted values, taking into account the corrections to the components of the time series additive model.

In this case, the estimated RMSE is 201, which is a considerable improvement compared to the earlier RSME of 353. This leads

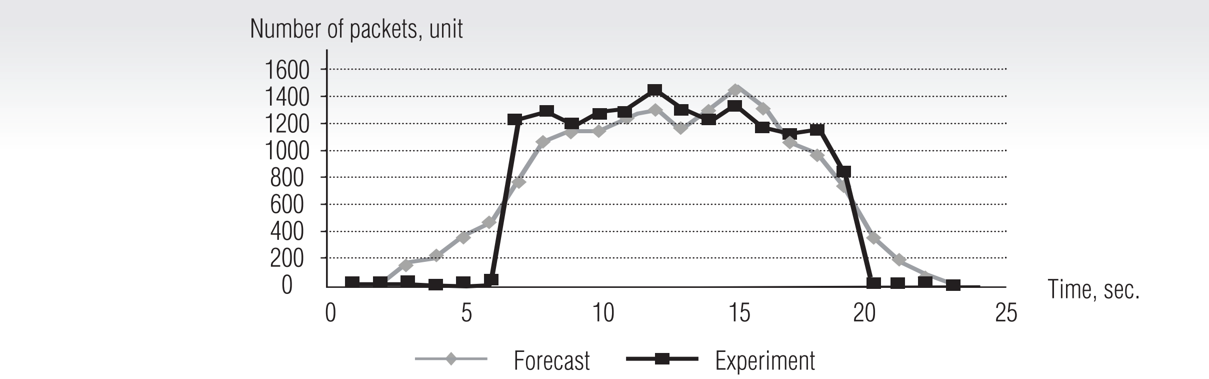

Fig. 7. Actual values for the test sample and predicted values for the number of network packets per second, with corrections

us to the conclusion that a predictive model of network attacks is much more accurate when it is based on recent experimental data rather than on the entire sample.

To calculate the relative error for the prediction model, we can calculate the ratio of the RMSE estimate to the maximum value in the test sample. Here we chose the maximum instead of the mean value because the sample considered contains many values close to 1. This leads to a relatively small mean value, which makes it impossible to estimate the relative error (MAPE) reliably. We found that the ratio of the RMSE to the maximum value in the test sample is 13%.

Therefore, the proposed prediction model of undesirable network traffic has a reasonably small relative error, and it can serve as an efficient tool for the detection of network attacks.

If necessary, the proposed model for the prediction of DDOS attacks could be used to study the general dynamics of the number of DDOS attacks in cyberspace [17]. If we use the number of attempted DDOS attacks in each quarter of years 2017 and 2018 as empirical training data, we can predict the number of DDOS attacks in the first half of year 2019.

The analysis of the data presented in Figure 8 shows that there are two periods in the dynamics of the number of DDOS attacks, namely 60 and 7 days. Apparently, the activity peaks (Feb 15, 2019; April 10, 2019, and June 5, 2019) of the envelope curve fall between relatively long holidays (March, May, and June). Short-scale periodic peaks are probably caused by the activity during particular days of the week. Therefore, a relatively simple prediction model allows us to find a connection between the periods in DDOS attacks and the calendar features for 2019.

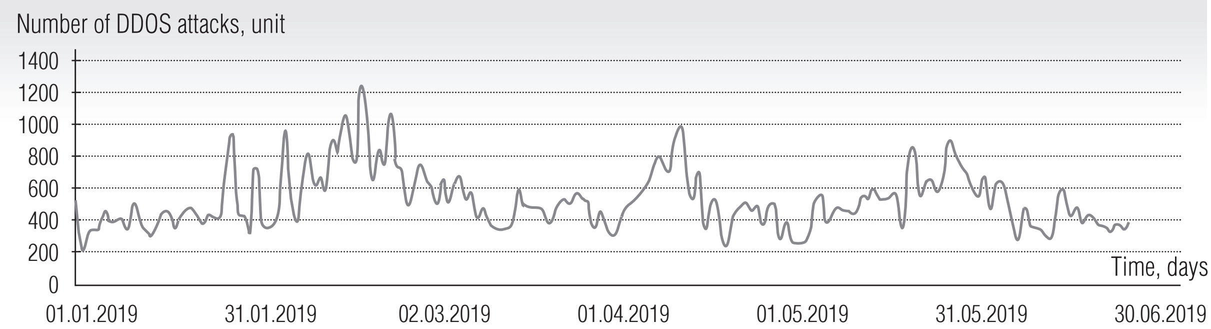

Fig. 8. Predicted numbers of DDOS attacks

Table 3.

Estimation of the trend and seasonal components

#	Current values, number of packets per second	Trend component estimate, number of packets per second	Seasonal component estimate, number of packets per second
1.	65	661	–596
2.	21	662	–641
3.	9	663	–654
4.	18	663	–645
5.	1088	664	424
6.	1398	665	733
7.	1301	666	635
8.	1363	667	696
9.	1343	668	675
10.	1375	669	706
11.	1283	670	613
12.	1378	671	707
13.	1387	672	715
14.	1304	673	631
15.	1276	674	602
16.	1302	675	627
17.	1295	676	619
18.	1380	677	703
19.	1391	678	713
20.	1062	679	383
21.	15	679	–664
22.	23	680	–657
23.	11	681	–670
24.	10	682	–672
25.	19	683	–664
26.	24	684	–660
27.	13	685	–672
28.	36	686	–650
29.	36	687	–651
30.	1313	688	625
31.	1342	689	653
32.	1360	690	670
33.	1439	691	748
34.	1380	692	688
35.	1290	693	597
36.	1384	694	690
37.	1329	695	634
38.	1306	695	611
39.	1315	696	619
40.	1296	697	599
41.	1309	698	611
42.	1298	699	599
43.	93	700	–607
44.	37	701	–664
45.	21	702	–681
46.	9	703	–694

Also, we should note that efficiency of the proposed model is higher when DDOS attacks have almost identical statistical parameters. If each implementation of a DDOS attack differs statistically, it is harder to detect and predict the hacker's actions.

Conclusion

This paper reports the results of network traffic analysis aimed at predicting the threats in cloud services. The statistics on information security threats to data storage and transmission that we present here validate the need for the development of new methods of data protection. Such methods typically use ad hoc hardware and software to analyze the information security threats. We implemented the malware that simulated network attacks, as well as the software that captured and processed the empirical data we needed for this study. We simulated a network attack (a DDOS attack) and saved the necessary parameters to files convenient for analysis and further processing. Out of many prediction models, we chose the additive time series model. The results obtained with the help of this model show that if we know the behavior of the statistical parameters of different implementations of a DDOS attack, we can detect and predict the hacker's actions for this type of attacks. The high efficiency of the proposed model is proven by comparison of the predicted values with the future actual values. The model's accuracy is characterized by the RMS error, which is equal to 201. The results of our research demonstrate that statistical methods of network traffic analysis can be employed in the tools used to protect the cloud services from various network attacks. ■

References

1. Maksimov K.V. (2018) Planning of activities in the IT-company in conditions of uncertainty taking into account the use of cloud services *Applied Informatics*, vol. 13, no 1, pp. 25–31 (in Russian).
2. Tumbinskaya M.V. (2017) Providing protection from targeted information in social networks. *Mordovia University Bulletin*, vol. 27, no 2, pp. 264–288 (in Russian).
3. Weins K. (2017) *Cloud computing trends: 2017 state of the cloud survey*. Available at: https://www.rightscale.com/blog/cloud-industry-insights/cloud-computing-trends-2017-state-cloud-survey (accessed 10 October 2018).
4. Krupin A. (2014) *We go into the clouds: Russian premiere of the IaaS-platform Flexible Computing Express*. Available at: https://servernews.ru/813983?k292300 (accessed 10 October 2018) (in Russian).
5. Tumbinskaya M.V. (2017) Process of distribution of undesirable information in social networks. *Business Informatics*, no 3, pp. 65–76.
6. Tadviser (2018) *Security threats in the cloud*. Available at: http://tadviser.ru/a/170054 (accessed 10 October 2018) (in Russian).
7. Baranov P.A., Beybutov E.R. (2015) Securing information resources using web application firewalls. *Business Informatics*, no 4, pp. 71–78.
8. Garcia S., Grill M., Stiborek J., Zunino A. (2014) An empirical comparison of botnet detectionmethods. *Computers and Security*, no 45, pp. 100–123.
9. Kosenko M.Yu., Melnikov A.V. (2016) Issues of protecting business information systems from botnets attacks. *Voprosy Kiberbezopasnosti*, no 4, pp. 20–28 (in Russian).
10. Gneushev V.A., Kravets A.G., Kozunova S.S., Babenko A.A. (2017) Modeling network attacks of attackers in a corporate information system. *Industrial Automatic Control Systems and Controllers*, no 6, pp. 51–60 (in Russian).
11. Cisco (2018) *Cisco annual report on cybersecurity for 2018*. Available at: https://www.cisco.com/c/ru_ru/about/press/press-releases/2018/03-12.html (accessed 10 October 2018) (in Russian).

12. Glushenko S.A. (2017) An adaptive neuro-fuzzy inference system for assessment of risks to an organization’s information security. *Business Informatics*, no 1, pp. 68–77.

13. Kartiev S.B., Kureichik V.M. (2016) Classification algorithm based on random forest principles for forecasting problem. *Software & Systems (Programmnye produkty i sistemy)*, no 2, pp. 11–15 (in Russian).

14. Afanas’ev A.P., Dzyuba S.M., Emelyanova I.I. (2017) Horner’s Scheme for investigation of solutions of differential equations with polynomial right-hand side. *Business Informatics*, no 2, pp. 33–39.

15. Tomilin A., Tumbinskaya M., Tregubov V., Smolevitskaya M. (2017) The BESM-6 virtualization project. Proceedings of the *2017 Fourth International Conference on Computer Technology in Russia and in the Former Soviet Union (SoRuCom 2017). Moscow, 3–5 October 2017*, pp. 241–245.

16. Pevtsova T.A., Ryabukhina E.A., Gushchina O.A. (2015) Calculation of seasonality index. *Mordovia University Bulletin*, vol. 25, no 4, pp. 18–36 (in Russian).

17. Kupreev O., Badovskaya E., Gutnikov A. (2018) *DDoS attacks in the third quarter 2018.* Available at: https://securelist.ru/ddos-report-in-q3-2018/92512/ (accessed 12 November 2018) (in Russian).

About the authors

Marina V. Tumbinskaya

Cand. Sci. (Tech.);

Associate Professor, Department of Information Protection Systems,
Kazan National Research Technical University named after A.N. Tupolev,
10, Karl Marx Street, Kazan 420111, Russia;

E-mail: tumbinskaya@inbox.ru

Bulat I. Bayanov

Student, Kazan National Research Technical University named after A.N. Tupolev,
10, Karl Marx Street, Kazan 420111, Russia;

E-mail: bayanov_bulat@mail.ru

Ruslan Zh. Rakhimov

Student, Kazan National Research Technical University named after A.N. Tupolev,
10, Karl Marx Street, Kazan 420111, Russia;

E-mail: rahimov96@mail.ru

Nikita V. Kormiltcev

Student, Kazan National Research Technical University named after A.N. Tupolev,
10, Karl Marx Street, Kazan 420111, Russia;

E-mail: kormiltcev@hotmail.com

Alexander D. Uvarov

Student, Kazan National Research Technical University named after A.N. Tupolev,
10, Karl Marx Street, Kazan 420111, Russia;

E-mail: obg-96@mail.ru