TRAFFIC DETECTION OF TRANSMISSION OF BOTNET THREAT USING BP NEURAL NETWORK

X.G. Li[†][*], *J.F. Wang*[†]

Abstract: With the gradual improvement of the telecommunication infrastructure in China, the Internet and other new technologies have been frequently used. The Internet technology also brings many network security threats, for example, botnet, while bringing convenience. Botnet is a network formed between hosts controlled by malicious code. One of the most serious threat to network security faced by the Internet is a variety of malicious network attacks on the carrier of botnet. Back propagation (BP) neural network is proposed to detect botnet threat transmission. In this study, a botnet detection model was established using BP neural network system. BP neural network classifier could identify the botnet traffic and normal traffic. Moreover a test was carried out to detect botnet traffic using BP neural network; the performance of the BP neural network classifier was evaluated by the detection rate and false positive rate. The results showed that it had high detection rate and low false positive rate, which indicated that the BP neural network had a good performance in detecting the traffic of botnet threat transmission.

Key words: *botnet, BP neural network, traffic, detection*

Received: August 22, 2018 **DOI:** 10.14311/NNW.2018.28.028
Revised and accepted: October 31, 2018

1. Introduction

Botnet is an novel attack technology based on the traditional malicious code, which is difficult to be detected and defended. At present, China has become a heavily affected area due to extremely serious transmission of botnet threat. In addition, government official websites and many Internet companies have also been attacked by botnets for many times. Survey shows that about 14.04 million hosts in China and 38.07 million overseas were installed with botnet in 2010, the number of hosts which were installed with botnet in China accounted for 30.3 % of the number around the world, and the number in Guangdong, Beijing and Zhejiang was the largest. Botnet has seriously affected the security operation of Internet, and has already jeopardized China's political economy and military. In order to ensure the

[*] Xingguo Li; National Key Laboratory of Fundamental Science on Synthetic Vision, Sichuan University, Chengdu, 610065, China; E-mail: xinggccse@yeah.net

[†] Junfeng Wang – Corresponding author; School of Aeronautics and Astronautics and College of Computer Science, Sichuan University, Chengdu, 610065, China, E-mail: lxg@scu.edu.cn

safe operation of the Internet, domestic and foreign researchers began to study how to effectively fight against botnets and reduce the harm caused by botnet through the traffic characteristics of botnets. The technology of back propagation (BP) neural network was proposed for traffic characteristic detection. Yun et al. [21] found the defect of repetitive scanning in the traditional botnet propagation strategy, analyzed the spread of botnet through back propagation (BP) neural network, proposed an algorithm to calculate the approximate optimal step size, and verified the effectiveness the algorithm for the spread of botnet through simulation experiments. Jiang et al. [10] put forward a BP neural network based botnet detection algorithm and took hosts as the analysis objects. The ordinary host and the robot were distinguished through the extraction of the traffic characteristics of hosts, and the detection rate was 99%, which suggested the good performance of the algorithm. Obeidat et al. [15] classified botnets using k-means, k-Medoids and neural network clusters and found that this method had an accuracy of 95.7 %. In this study, BP neural network was used to detect the traffic characteristics of transmission of botnet threat and test the effectiveness of detection technology.

2. Traffic detection of transmission of botnet threat using BP neural network

2.1 The concept of bonnet

Botnet is a platform used to launch malicious attacks controllably, which combines a variety of traditional malware technologies, including computer viruses, worms and Trojan horses. A large number of host have botnets. When the host receives the attack command from attackers, organized malicious attacks will be carried out by botnets, resulting in serious and large-scale hazards, including privacy security issues, such as stealing privacy and distributed denial of service attack [11], spam [2], phishing [17] and so on.

2.2 Botnet traffic

Both normal traffic and botnet traffic exist in botnet host, but there is a large difference in communication mode between botnet traffic and normal traffic. Botnet traffic should be distinguished from normal traffic in aspects of number of transmission control protocol (TCP) stream, the average value of time interval, the average value of time interval change, the average value of byte number, the average value of data packet number and the average value of the duration. Moreover, the average values of data packet number, TCP stream number and byte number are significantly lower than the normal traffic, but the average values of time interval, time interval change and duration are significantly higher than the normal traffic, which are shown in Tab. I. Detection of botnet can be realized through those differences.

	Number of TCP stream	The average value of time interval	The average value of time interval change	The average value of byte number	The average value of data packet number	The average value of the duration
Normal traffic	10.9427	−19.3522	−4.6336	17632.4154	33.1331	13.5455
Botnet traffic	12.3164	0.6245	−0.0001	4202.1326	13.4784	57.3256

Tab. I *The difference between normal traffic and botnet traffic.*

3. BP neural network based botnet detection model

3.1 Preprocessing of traffic features

When BP neural network was used in botnet detection, host was taken as the analysis object and the features of host pair communication flow were extracted. Each host pair has six features. The six input values represented different aspects of traffic features, and the units and ranges of the contents expressed were quite different. If these data were directly input into the neural network, the data would become huge after weighted accumulation, making it difficult for the network to converge in a limited time. In order to improve the detection rate of botnet, it is necessary to pre-process the eigenvalues [4] and normalize the parameters before training the neural network classifier.

Traffic characteristics were normalized using z-score method:

$$x' = \frac{x - \bar{x}}{\alpha_x}, \tag{1}$$

where x is the value of a traffic feature and $\bar{x}$ and α_x stand for the average value and standard deviation of the value.

The formula of the average value $\bar{x}$ was $\bar{x} = \sum_{i=1}^{n} x_i$, and the formula of the standard deviation α_x was $\alpha_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$.

3.2 Structure construction of BP neural network classifier

Like human neural network, BP neural network is a multi-layer feedforward neural network. Generally, the three-layer network includes input layer, hidden layer [13] and output layer, which can complete the mapping of any dimensional space. According to 6 dimensions of the characteristic vector of host pair, the input layer of BP neural network is determined to be 6 neurons, which are respectively TCP stream number [9], the average value of time interval, the average value of time interval change, the average value of byte number, the average value of data packet number and the average value of the duration. There are two corresponding output results, which are non-botnet and botnet.

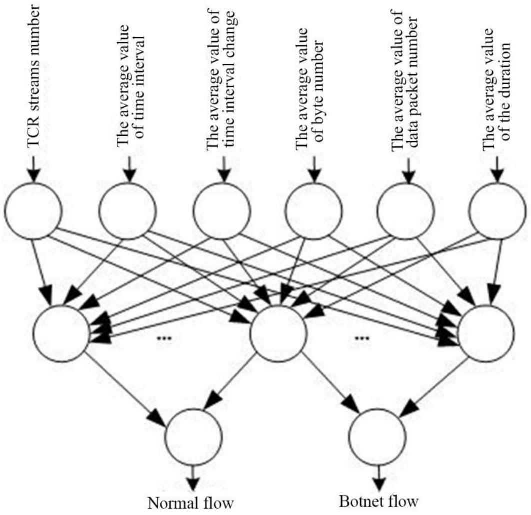
TCR streams number
The average value of time interval
The average value of time interval change
The average value of byte number
The average value of data packet number
The average value of the duration
...
...
Normal flow
Botnet flow

Fig. 1 *Model diagram of BP neural network classifier.*

The determination of the number of hidden layer units often requires experience and multiple tests. The empirical formula is as follows:

$$\mathrm{abc} = \sqrt{a+b} + c, \tag{2}$$

where *abc* stands for the number of hidden neurons, a stands for the number of input neurons, b stands for the number of output neurons, c stands for empirical constant, and the data range is [1, 10].

Due to relatively large network traffic [3], and the number of hidden layer is proportional to the neural network training speed, therefore the BP neural network classifier is not applicable to multiple hidden layer. Single hidden layer structure should be adopt, especially when the data traffic is relatively large. The single hidden layer has a faster training speed, and through the BP neural network algorithm, a large amount of learning on the training samples can be conducted, the threshold and the weight can be adjusted, and the error of the output value is relatively small, which has no large influence on the network accuracy.

3.3 The training of BP neural network

The designed neural network classifier was trained with a large amount of normal traffic and botnet traffic. Moreover, the training process of BP neural network is a process of constant modification. The flow chart of BP neural network training algorithm [5] is shown in Fig. 2.

The first step of BP neural network training was to set the expectation error or training times, input the training sample, and let the BP neural network correct

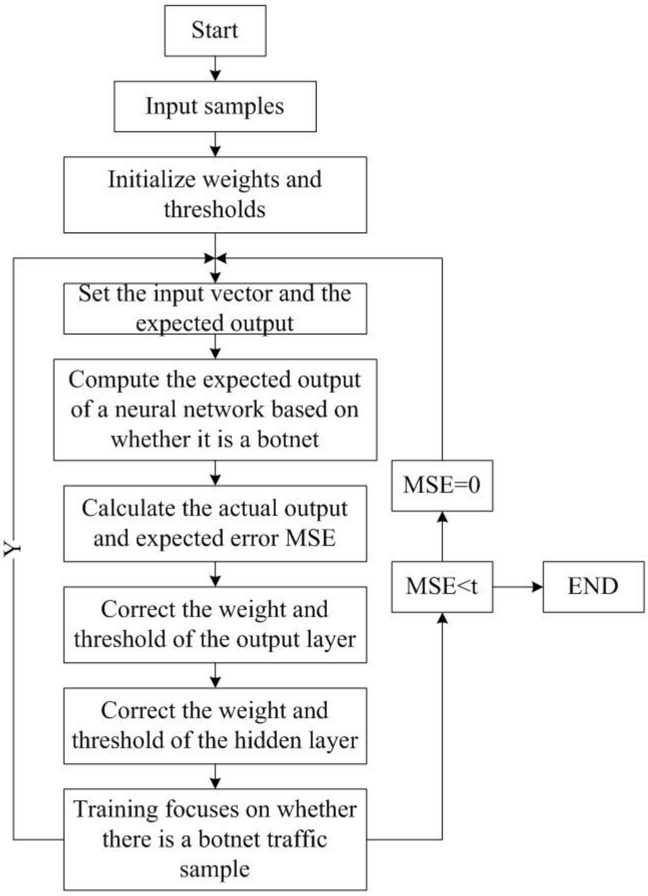
Start
Input samples
Initialize weights and thresholds
Set the input vector and the expected output
Compute the expected output of a neural network based on whether it is a botnet
Calculate the actual output and expected error MSE
Correct the weight and threshold of the output layer
Correct the weight and threshold of the hidden layer
Training focuses on whether there is a botnet traffic sample
Y
MSE=0
MSE<t
END

Fig. 2 *The flow of BP neural network training algorithm.*

the weight and threshold, so that the error number decreased along the negative gradient direction until the goal was completes. The training steps of BP neural network [19] is shown as follows:

- The parameters were initialized, and the expected error and the largest training times were set.
- Training samples were input, and six input neurons were assigned with six traffic eigenvalues.
- Expected error was set; the output neuron was set as 1 if it was botnet traffic and as 0 if it was normal traffic.
- y_j was calculated and output, the neurons number of input layer, hidden layer and output layer was set as i, k and j, and then x_j was output; the weight that connected the input layer and hidden layer was v_{ij}, the threshold of hidden layer node was θ_j, the learning rate was β, and the expected error was t.

- The output node n_k and the error MSE were calculated and output, and the connection weight between the hidden layer and the output layer was set as w_{jk}; the output node threshold was θ_k, and the expected output a_k.
- The weight and threshold of the output layer and the hidden layer was corrected, if MSE $< e$, the training ends, otherwise it will continue.

  The error formula is:

  $$\mathrm{MSE} = (t_k - n_k) f'(\mathrm{net}_k) = (t_k - n_k) n_k (1 - n_k).$$

  Weight correction formula:

  $$w_{jk}(m+1) = w_{jk}(m) + \Delta w_{jk} = w_{jk}(m) + \beta \mathrm{MSE}_k y_j.$$

  Threshold correction formula:

  $$\theta_k(m+1) = \theta_k(m) + \beta \mathrm{MSE}_k.$$

- Step 2 $\sim$ 6 were repeated until the expected error or the largest training times was achieved.

After the training, BP neural network can be used to detect botnet traffic and distinguish normal traffic from botnet traffic.

4. Testing and the results analysis

4.1 Data preparation

The data collected in this test was the network traffic of the campus network of Sichuan University, China, including the traffic generated by normal network applications such as hypertext transfer protocol, simple mail transfer protocol (SMTP), file transfer protocol (FTP), etc., and the traffic generated by the simulated botnet C&C channel [18] was injected into the database, which regards the traffic in the real network environment as the background traffic. The collected number of samples that was detected as the normal traffic was 420, the number of samples that was detected as the botnet traffic was 80, and the total number of samples was 500.

4.2 Testing

In order to verify the effectiveness of BP neural network classifier in traffic detection of transmission of botnet threat, this experiment evaluated the performance of neural network classifier using two indicators: detection rate (DR) and false positive rate (PFR).

$$\mathrm{DR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}},$$

where TP stands for the number of samples that was correctly detected as the botnet traffic, and FN stands for the number of samples that cannot be detected as the botnet traffic.

The false positive rate is the probability of being wrongly detected as the botnet traffic:

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}$$

where TN stands for the number of samples that was correctly detected as the normal traffic, and FP stands for the number of samples that cannot be detected as the normal traffic.

First, whether there is error between the experimental results and the real results should be judged, then the samples of TP, FN, FP and TN should be counted. Finally, the required data should be calculated according to the above formulas, which are shown in Tab. II.

Expected error	TP	FN	FP	TN	Detection rate	False positive rate
0.001	76	4	0	420	0.9500	0
0.002	79	1	0	420	0.9875	0
0.003	80	0	0	420	1	0
0.004	80	0	0	420	1	0
0.005	80	0	0	420	1	0
0.006	80	0	0	420	1	0
0.007	80	0	0	420	1	0
0.008	80	0	1	419	1	0.00238
0.009	80	0	3	417	1	0.00714
0.010	80	0	4	416	1	0.00952

Tab. II *Test results under different errors.*

It was found form Tab. II that when the expected error was set within the range of 0.001 ∼ 60.010, the expected error was proportional to the detected botnet traffic sample, i.e., the detected botnet traffic sample increased with the increase of the expected error, and the detection rate increased. However, the number of samples whose normal traffic was misreported as botnet traffic also increased, and the false positive rate increased. When the expected error was 0.001 and 0.002, the number of samples that cannot detect botnet traffic were 4 and 1 respectively, i.e., the detection rate was 95 % and 98.75 % respectively, but when the expected error was set to 0.003 and above, the detection rate remained at 100 %. When the expected error was 0.008, false reports started to appear, and the number of samples that misreported the normal traffic as the botnet traffic was 1, so the false positive rate cannot be maintained at 1 %, and the false positive rate increased with the increase of the expected error. As can be seen from Tab. II, the detection rate and the false positive rate maintained at a relatively good level when the expected error was between 0.001 and 0.01. Through the observation and analysis of the test results, it was found that it had a higher detection rate and lower false positive rate, indicating that the BP neural network classifier had a better performance in the traffic detection of botnet and had effectiveness and feasibility in the traffic detection of transmission of botnet threat.

According to Tab. II, the variation diagrams of detection rate and false positive rate with the change of expected error are shown in Fig. 3 and Fig. 4.

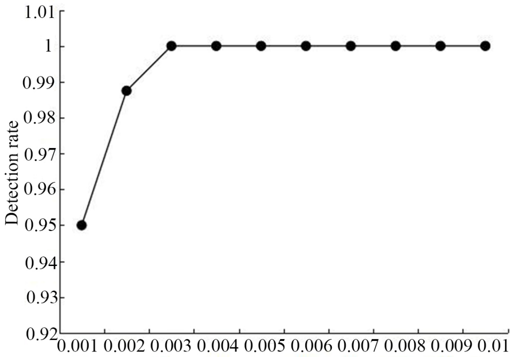

Fig. 3 *The variation diagram of detection rate.*

It can be seen from Fig. 3 that the detection rate increased with the increase of expected error. In the change process of expected error from 0.001 to 0.01, it can be seen that the detection rate changed from 95 % to 100 % and reached 100 % when the expected error was 0.003, and the value will not decrease. The detection rate changed from 95% to 98.75 % when the expected error increased from 0.001 to 0.002 and changed from 98.75 % to 100 % when the expected error increased from 0.002 to 0.003, indicating that the detection rate increased with the expected error, and the change rate of detection rate decreased. Therefore, if the expected error was lower than 0.001, the BP neural network system does not have a good detection rate. However, when the expected error was 0.001 ∼ 0.01, the BP neural network system had a better detection rate.

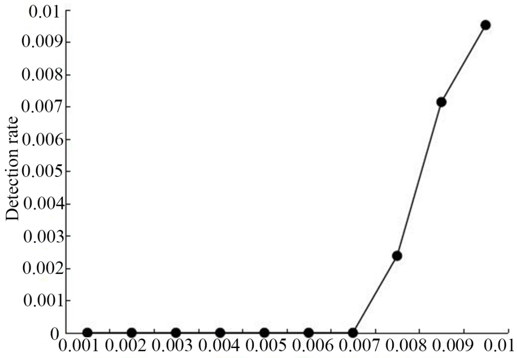

Fig. 4 *The change of false positive rate.*

As can be seen from Fig. 4, the false positive rate increased with the increase of the expected error within a certain range. When the expected error changed from 0.001 to 0.007, the false positive rate was always 0, starting from 0.008, the false positive rate increased with the increase of the expected error. But the false positive rate did not exceed 1 % in the range of 0.001 ∼ 0.01. However, when the expected error exceeded 0.01, the false positive rate of BP neural network system will exceed 0.01, and there will be no relatively low false positive rate of BP neural network system. While in the range of 0.001 ∼ 0.01, BP neural network system has low false positive rate.

Discussion

With the development of Internet technology, its application in daily life is more and more extensive, which brings about increasingly serious network security problems [8]. Botnet, a controllable platform which contains many malware technologies, brings great threats to Internet security [14]. For example, distributed denial-of-service attacks based on botnet restrict resources and bring huge losses to enterprises and governments [1]. It is urgent to find effective botnet detection and defense methods to maintain network security and solve network threats. The traditional network detection methods mostly use offline detection methods, i.e., detecting botnets through analysis of historical network data, but with the update and variation of botnets, the accuracy of historical data information is not high, which can lead to a low detection rate.

The advent of traffic based detection methods brings new vitality into botnet detection [16, 20]. Network traffic has long-term and short-term characteristics. Network abnormality can be identified through traffic detection. Gianvecchio et al. [6] used traffic pattern recognition to distinguish normal traffic from botnet traffic and achieve good classification results by using entropy and Bayesian based classifiers. Haddadi et al. [7] explored five different traffic outputs using two different protocol filters and five different classifiers and found that the best performance was achieved using Tranalyzer traffic outputs and hypertext transfer protocol filter which carried decision tree C4.5 algorithm based classifier. Based on the traffic, this study extracted the features of normal traffic and zombie traffic. Six features, including the average value of time interval, the average value of time interval, the average value of byte number, the average value of data packet number and the average value of the duration, were used to distinguish the botnet traffic from the normal traffic. Then BP neural network detection model was established and BP neural network was trained. It was found that the proposed method had high detection rate and low false positive rate, indicating that it was effective. The botnet detection technology of BP neural network relies on the header information of the data packet and is in no correlation with the isnformation related to the inside of the data packet. It can detect the botnets encrypted in the communication. Neural network classifier [12] does not relay on malicious attack of botnet and can play its function even when botnet is slient. Neural network can memorize traffic features of botnet after training and then detect new traffic.

5. Conclusion

In this study, the BP neural network system was applied to the detection of botnet. Through input neural network and a large amount of training on the preprocessed botnet traffic characteristics, the traffic characteristics of the communication between botnets were learned, and then the detection of botnet traffic was applied. Through experiments, it was confirmed that within the expected error, BP neural network had a higher detection rate and a lower false positive rate for botnet traffic detection, which indicated the effectiveness and feasibility of BP neural network for traffic detection of transmission of botnet threat. However, this test also had its limitations. This test selected a certain type of botnet traffic for traffic detection, and the traffic characteristics of different types of botnets have certain differences. Therefore, when BP neural network was applied to different types of botnet detection, it was necessary to collect training samples for the new botnets and train BP neural network again to achieve good detection effect. To sum up, BP neural network had the effectiveness and feasibility for the traffic detection of transmission of botnet threats. The network is closely related to everyone. BP neural network plays an important role in fighting against botnet and reducing the harm caused by botnet. It is necessary for us to deeply study BP neural network and improve traffic detection of transmission of botnet threat.

References

[1] ALOMARI E., MANICKAM S., GUPTA B.B., KARUPPAYAH S., ALFARIS R. Botnet-based Distributed Denial of Service (DDoS) Attacks on Web Servers: Classification and Art. *International Journal of Computer Applications*, 2012, 49(7), pp. 24–32. doi: 10.5120/7640-0724.

[2] AL-ZOUBI A.M., ALQATAWNA J., FARIS H. Spam profile detection in social networks based on public features. *In: Proceedings of the 8th International Conference on Information and Communication Systems (ICICS 2017)*,Irbid, Jordan. IEEE, 2017, pp. 130–135.

[3] CEJKA T., BARTOS V., SVEPES M., ROSA Z., KUBATOVA H. NEMEA: A framework for network traffic analysis. In: *International Conference on Network and Service Management*, Montreal, QC, Canada, 2017, pp. 195–201.

[4] CHEN H., LIU J., CHANG X., CHEN D., XUE Y., LIU P., LIN H., HAN S. A review on the pretreatment of lignocellulose for high-value chemicals. *Fuel Processing Technology*, 2017, 160, pp. 196–206. doi: 10.1016/j.fuproc.2016.12.007.

[5] DENG Z., CHOI K.S., CAO L., WANG S. T2FELA: type-2 fuzzy extreme learning algorithm for fast training of interval type-2 TSK fuzzy logic system. *IEEE Transactions on Neural Networks and Learning Systems*, 2017, 25(4), pp. 664–676. doi: 10.1109/TNNLS.2013.2280171.

[6] GIANVECCHIO S., XIE M., WU Z., WANG H. Humans and Bots in Internet Chat: Measurement, Analysis, and Automated Classification. IEEE/ACM Transactions on Networking, 2011, 19(5), pp. 1557–1571. doi: 10.1109/TNET.2011.2126591.

[7] HADDADI F., ZINCIR-HEYWOOD A.N. Benchmarking the Effect of Flow Exporters and Protocol Filters on Botnet Traffic Classification. *IEEE Systems Journal*, 2016, 10(4), pp. 1390–1401, doi: 10.1109/JSYST.2014.2364743.

[8] Handbook of research on modern cryptographic solutions for computer and cyber security. IGI Global, 2016.

[9] HOLT M.P., PUNKOSDY G.A., GLASS D.D., SHEVACH E.M. TCR Signaling and CD28/CTLA-4 Signaling Cooperatively Modulate T Regulatory Cell Homeostasis. *Journal of Immunology*, 2017, 198(4), pp. 1503–1511. doi: 10.4049/jimmunol.1601670.

[10] JIANG H., SHAO X. Botnet detection algorithm based on neural network. *Caai Transactions on Intelligent Systems*,2013, 8(2), pp. 113–118.

[11] KOLIAS C., KAMBOURAKIS G., STAVROU A., VOAS J. DDoS in the IoT: Mirai and Other Botnets. *Computer*, 2017, 50(7), pp. 80–84, doi: 10.1109/MC.2017.201.

[12] ILIN C.T., PRASAD M., SAXENA A. An Improved Polynomial Neural Network Classifier Using Real-Coded Genetic Algorithm. *IEEE Transactions on Systems Man & Cybernetics Systems*, 2017, 45(11), pp. 1389–1401. doi: 10.1109/TSMC.2015.2406855.

[13] MATTICK J.S. Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms. *Bioessays*, 2010, 25(10), pp. 930–939. doi: 10.1002/bies.10332.

[14] NIU W.N., ZHANG X.S., SUN E.B., YANG G.W., ZHAO L.Y. Two Stage P2P Botnet Detection Method Based on Flow Similarity. *Journal of University of Electronic Science & Technology of China*, 2017, 46(6), pp. 902–906 and 948. doi: 10.3969/j.issn.1001-0548.2017.06.019.

[15] OBEIDAT A. Hybrid Approach for Botnet Detection Using K-Means and K-Medoids with Hopfield Neural Network. *International Journal of Communication Networks & Information Security*, 2017, 9(3), pp. 305–313.

[16] QIU Z., MILLER D.J., KESIDIS G. Flow based botnet detection through semi-supervised active learning. In:*IEEE International Conference on Acoustics, Speech and Signal Processing*, New Orleans, LA, USA, 2017, pp. 2387–2391. doi: 10.1109/ICASSP.2017.7952584.

[17] SPEARS T. Phishing for Phools: The Economics of Manipulation & Deception. *Quantitative Finance*, 2017, 17(2), pp. 1–3, doi: 10.1080/14697688.2016.1252193.

[18] WANG X.F., ZHANG D.F., XIN S.U. Design of Mobile Botnets Integrated SMS and HTTP Protocol C&C Channel. *Journal of Chinese Computer Systems*, 2014, 35(7), pp. 1458–1463.

[19] WANG X.P. Research and Optimization of BP Neural Network Algorithm. In: *International Conference on Measuring Technology & Mechatronics Automation*, Nanchang, China, 2015, pp. 818–822.

[20] YEH Y.R., SUN M.K., HUANG C.Y. A Malware Beacon of Botnet by Local Periodic Communication Behavior. In: *IEEE Computer Software and Applications Conference*, Tokyo, Japan, 2018, pp. 653–657.

[21] YUN Y., HU G.Y., LI H.B., LUO J. Simulation of optimal step size propagation of BotNet based on OPNET. *Journal of Pla University of Science & Technology*, 2012, pp. 12–19.