Detecting Abnormal Interactions among Intranet Groups Based on Netflow Data

Tao Yu¹ and Ruiqi Yue²
¹Beijing Institute of Tracking and Telecommunications Technology, Beijing, China; ²Science and Technology on Information Systems Engineering Laboratory, National University of DT, Changsha, China; Email:magnia@sina.com

Abstract. This paper proposes a method for detecting abnormal interactions among intranet groups based on netflow data. Firstly, the netflows of each group are aggregated, and two anomaly detection indicators are constructed, i.e., the group network traffic and the uncertainty of group network traffic distribution. Secondly, the time series of two anomaly detection indicators of each group are analyzed, and four prediction models are used for prediction. Finally, the best-performing model is selected as the prediction benchmark, and the difference between the predicted result and the real data is used to detect whether there is an interaction anomaly among groups. The experimental results show that the proposed method can effectively detect the abnormal interaction among groups in intranet.

1. Introduction

With the increase of network penetration, APT attacks and other new network threats, the organization or enterprise intranet is no longer as secure as previously known. Traditional security technologies based on boundary protection have been unable to perceive and analyze the security situation in the intranet. Therefore, data-driven intranet security tools and analysis methods have attracted great attention in recent years. In many data, network flow is very important for intranet security monitoring. It is a group of IP (Internet Protocol) packets with the same attributes and passing through a certain observation point in the network in a specific period of time. The attributes include the header information (such as IP address, port number), the attributes of the packet itself (such as the length of the packet), the transmission mode of the packet (such as the IP address of the next hop) and so on.

Abnormality refers to the rare behavior deviating from the normal pattern [2]. Generally, network flow anomaly detection is based on the prediction results of time series characteristics. Time series analysis method includes feature extraction method based on basic statistical method [3], feature extraction method based on transformation [4], feature extraction method based on fractal theory [5], and feature extraction method based on model. Among them, the commonly used model-based feature extraction methods include AR model, MA model, ARMA model, ARIMA model and other traditional time series models, as well as RNN, LSTM, SVR and other new time series models based on neural network or support vector machine.

In the field of network anomaly detection, among domestic researchers, Crown Male et al. [6] extracted 9 traffic characteristics used to identify the intrusion behavior of unauthorized users in WLAN, and designed an intrusion detection scheme using support vector machine. Jingtao et al. [7] proposed a method of anomalous scanning behavior monitoring based on numerical analysis, which realized real-time monitoring of the main network worm viruses and malicious scanning detection of hackers. Wang Yongchao et al. [8] proposed a method of using AR model to get the growth rate of
abnormal data packets, and based on this, analyzed the possibility of worm infection in the current network. Cai Jiamei et al. [9] proposed an intranet distributed security audit system based on user behavior, and applied the neural network based on genetic algorithm to audit log analysis and anomaly detection. Jiang Wei et al. [10] proposed a network behavior analysis system model based on IPFIX protocol, and classified normal traffic behavior, sudden traffic behavior and illegal traffic behavior. Among foreign researchers, Krishnamurthy et al. [11] proposed a method that can efficiently detect heavy-hitters flow in multiple dimensions. Stainford et al. [12] studied the detection method of port scanning behavior. Sherif et al. [13] compared the effects of five common machine learning methods in analyzing network behavior and detecting botnets, and successfully detected and prevented botnets from intruding in the command and control phase (C&C). Estan et al. [14] studied the resource consumption in the process of network circulation, and proposed a clustering algorithm which can automatically discover significant circulation patterns. Bakoben et al. [15] used an approximation method to cluster the entity relationships in the enterprise network, which served as the basis for the analysis of Intranet behavior. Xu et al. [16] used data mining and information theory to extract the characteristics of Internet traffic and applied them to behavior modeling.

On the basis of network flow data, this paper proposes an anomalous interaction detection method based on time series prediction results, aiming at the increasing complexity of Intranet space security situation, the increasing security uncertainties in Intranet and the insufficiency of corresponding countermeasures, and carries out experimental analysis in real intranet environment.

2. Construction of Anomaly Detection Index

The collected network flow data are aggregated according to the functional departments of intranet users to generate network traffic transmission data among different groups. In a fixed time period T, the interaction matrix N of m groups is expressed as:

$$N = \begin{bmatrix}
n_{r_1,r_1} & n_{r_1,r_2} & \cdots & n_{r_1,r_m} \\
n_{r_2,r_1} & n_{r_2,r_2} & \cdots & n_{r_2,r_m} \\
\vdots & \vdots & \ddots & \vdots \\
n_{r_m,r_1} & n_{r_m,r_2} & \cdots & n_{r_m,r_m}
\end{bmatrix}$$

(1)

It is used to express the total flow from the group to the group in T, that is, the sum of the network flow from all the source addresses belonging to the group and the target addresses belonging to the group in T. The total amount of traffic flowing out of the group in T, and the traffic flowing through the whole network in T. The total network flow of the group in time T is depicted.

From the traffic sequence of each group, we can only get the total flow from the group at each time point, but we cannot get further information about the direction of these flows. In order to characterize the distribution characteristics of outflow from each group, the uncertainty of flow distribution within time T is defined as:

$$RU_{FlowSize_{r_i}} = \frac{H(P_{flowsize})}{\log m} = \frac{-\sum_{j=1,2,…,m}^{n_{r_i,j}} \frac{n_{r_i,j}}{n_{r_i}} \log \frac{n_{r_i,j}}{n_{r_i}}}{\log m}$$

(2)

The statistical probability of the flow from the group to each group in time T is the lowest uncertainty when it is equal to 0. When it is equal to 1, it means that the flow from the group flows to each group equally and has the greatest uncertainty. The characteristics of network traffic distribution of population in time T are depicted.

The uncertainties of group network traffic and group network traffic distribution characterize the group characteristics of group-related network traffic from two perspectives respectively. Therefore, the anomaly detection index set used in this paper is \(TEST^=\).

For a certain period of time T and a group that need anomaly detection, the predicted value of the group’s detection index in that period is assumed by the prediction model, while the actual value of the group’s detection index in that period is assumed. The Theil Inequality Coefficient of the two indicators was calculated respectively.
\[
\mu(n_{r_i}) = \frac{|n_{r_i,\text{predict}} - n_{r_i,\text{true}}|}{n_{r_i,\text{predict}} + n_{r_i,\text{true}}} \tag{3}
\]

\[
\mu(RU_{\text{FlowSize}}_{r_i}) = \frac{|RU_{\text{FlowSize}}_{r_i,\text{predict}} - RU_{\text{FlowSize}}_{r_i,\text{true}}|}{RU_{\text{FlowSize}}_{r_i,\text{predict}} + RU_{\text{FlowSize}}_{r_i,\text{true}}} \tag{4}
\]

Setting threshold sum, if > and >, defines the occurrence of abnormal interactions with respect to groups at time \(T\).

3. Data Analysis and Model Training

3.1. Time Series Generation

This paper studies the internal network environment of a unit, and the network structure is connected with the external Internet. The whole network is divided into seven sub-networks. All network data packets will pass through the central switch, and aggregate the data packets into network flows in the core switch. Seven subnets correspond to six functional departments and one server. If only personal computers are considered in each department, the group set \(R=\{\text{Department 1, Department 2, Department 3, Department 4, Department 5, Department 6, server}\}\). By collecting the data from the core switch, the network flow data generated by 661 hosts in the intranet during the 17 days from November 11 to November 27, 2016 are obtained. There are 195,037,491 data records in total, and the data size is 29.63GB.

If the sampling time interval \(T = 10\) minutes, the 24-hour network flow data can generate a 144-length traffic time series with each sequence value as the traffic interaction matrix in that time period. From November 11 to November 27, 2016, there are 11 working days and 6 rest days. Taking 10 working days and 5 rest days, the time series with length 1440 and 720 are generated respectively, and recorded as \(\{1\} \) and \(\{1\} \), as shown in Table 1. Situational validation data are retained on November 25, working day and November 27, rest day.

| Weekdays | Day 11 | 14 | 15 | 16 | 17 |
|----------|--------|----|----|----|----|
| Sequence | \(x_1\)-\(x_{144}\) | \(x_{145}\)-\(x_{288}\) | \(x_{289}\)-\(x_{432}\) | \(x_{433}\)-\(x_{576}\) | \(x_{577}\)-\(x_{720}\) |

| Weekdays | Day 18 | 21 | 22 | 23 | 24 |
|----------|--------|----|----|----|----|
| Sequence | \(x_{721}\)-\(x_{864}\) | \(x_{865}\)-\(x_{1008}\) | \(x_{1009}\)-\(x_{1152}\) | \(x_{1153}\)-\(x_{1296}\) | \(x_{1297}\)-\(x_{1440}\) |

| Paliday  | Day 12 | 13 | 19 | 20 | 26 |
|----------|--------|----|----|----|----|
| Sequence | \(y_1\)-\(y_{144}\) | \(y_{145}\)-\(y_{288}\) | \(y_{289}\)-\(y_{432}\) | \(y_{433}\)-\(y_{576}\) | \(y_{577}\)-\(y_{720}\) |

3.2. Analysis and Prediction of Traffic Sequences of Different Groups

Statistical tests on the normality, stationarity, randomness and autocorrelation of time series in Table 1 show that the traffic series of all groups deviate from the normal distribution, so the effect of traditional prediction methods depending on the error assumption of normal distribution may be affected; moreover, the traffic series of each group are non-random. The results of ADF test and Ljung-Box test show that the flow sequence of most groups is unstable and self-correlated, and only sector 3 is stable. The sequence has the characteristics of strong periodicity of working day and stable sequence of rest day.

For each group, ARIMA, SVR, LSTM and RNN are used to predict the traffic sequence. The first 80% of each sequence is used as training set and the last 20% as test set. For working days, the first 1152 time slices are taken as test sets, and the second 288 time slices are taken as test sets. For rest days, the first 576 time slices are taken as training sets, and the last 144 time slices are taken as test.
sets. For RNN and LSTM models, batch size is set to 10, epoch is set to 20 during training, learning rate is set to 0.01, the optimizer used is Adam, and the loss function is mean square error (MSE). For ARIMA model, the optimal autoregressive parameter $p$, difference parameter $D$ and moving average parameter $q$ are determined according to the best Akaike information criterion (AIC). For SVR model, the kernel function is Radial Basis Function (RBF), and the insensitive parameters, penalty coefficient $C$ and kernel function parameters are determined by the loss function of optimized training data.

In order to compare the training effects of different models, the root mean square error (RMSE) of the test set was calculated as a measure of prediction accuracy.

$$RMSE = \sqrt{\frac{1}{n_s} \sum_{s=1}^{n_s} (\hat{D}_s - D_s)^2}$$

Among them, the number of predictions, and the actual value and the real value of the second prediction, respectively. Since the last 20% of the total sequence is used as the test set, for the traffic sequence of working days, $n_s = 288$, for the traffic sequence of rest days, $n_s = 144$.

The prediction results of different models are compared as shown in Tables 2 and 3. It is noticed that RNN and LSTM are better than ARIMA model and SVR model in predicting traffic series of different groups in workday, because the traffic series of each group in workday show strong periodicity and weak burst, and the neural network can extract this feature better. On the rest day, on the contrary, the flow sequence of most populations is basically stable and there are only a few obvious mutations, so ARIMA model has a good prediction effect, while RNN's prediction RMSE performance is the worst.

### Table 2 Prediction of RMSE for Workday Traffic Sequences of Groups

| Model | Dept.1 | Dept.2 | Dept.3 | Dept.4 | Dept.5 | Dept.6 | Server |
|-------|--------|--------|--------|--------|--------|--------|--------|
| ARIMA | 1109.24| 1148.81| 970.38 | 1980.09| 831.67 | 1563.2 | 635.13 |
| RNN   | 895.37 | 968.34 | 950.98 | 1962.35| 854.34 | 985.23 | 532.54 |
| LSTM  | 863.72 | 899.5  | 943.26 | 1943.78| 857.23 | 993.56 | 521.78 |
| SVR   | 952.76 | 997.46 | 945.6  | 2045.67| 879.29 | 1389.23| 611.59 |

### Table 3 Prediction of RMSE by Daily Flow Series of Groups at Rest

| Model | Dept.1 | Dept.2 | Dept.3 | Dept.4 | Dept.5 | Dept.6 | Server |
|-------|--------|--------|--------|--------|--------|--------|--------|
| ARIMA | 55.71  | 59.62  | 26.11  | 53.5   | 86.5   | 102.57 | 84.76  |
| RNN   | 62.10  | 62.23  | 32.45  | 53.18  | 136.58 | 111.24 | 92.06  |
| LSTM  | 60.63  | 56.37  | 26.57  | 56.26  | 110.58 | 109.39 | 83.68  |
| SVR   | 63.25  | 57.56  | 26.82  | 54.40  | 84.60  | 108.29 | 82.06  |

#### 3.3. Analysis and Prediction of Uncertainty Sequence of Flow Distribution in Different Groups

The time series of uncertainties in the flow distribution of each group is constructed by the flow interaction matrix in each period according to formula (2). The uncertain sequence of flow distribution of each group has strong volatility, and also has the characteristics of strong periodicity of working day series and stable rest day series.

For each group, ARIMA, RNN, LSTM and SVR are used to predict the traffic uncertainty series. The first 80% of the workday series and the rest day series are taken as training sets, and the last 20% as test sets. For RNN and LSTM models, batch size is set to 10, epoch is set to 5 during training, learning rate is set to 0.05, the optimizer used is Adam, and the loss function is mean square error.
(MSE). The parameters of ARIMA model are determined according to the best Akaike information criterion (AIC). For SVR model, the kernel function is Radial Basis Function (RBF), and the insensitive parameters, penalty coefficient C and kernel function parameters are determined by the loss function of optimized training data. The predicted results are shown in tables 4 and 5.

| Model | Dept.1 | Dept.2 | Dept.3 | Dept.4 | Dept.5 | Dept.6 | Server |
|-------|--------|--------|--------|--------|--------|--------|--------|
| ARIMA | 0.048  | 0.048  | 0.048  | 0.076  | 0.049  | 0.068  | 0.062  |
| RNN   | 0.06   | 0.05   | 0.06   | 0.08   | 0.07   | 0.07   | 0.05   |
| LSTM  | 0.05   | 0.05   | 0.05   | 0.08   | 0.08   | 0.06   | 0.05   |
| SVR   | 0.05   | 0.05   | 0.05   | 0.08   | 0.09   | 0.06   | 0.05   |

| Model | Dept.1 | Dept.2 | Dept.3 | Dept.4 | Dept.5 | Dept.6 | Server |
|-------|--------|--------|--------|--------|--------|--------|--------|
| ARIMA | 0.068  | 0.058  | 0.03   | 0.163  | 0.050  | 0.065  | 0.088  |
| RNN   | 0.07   | 0.06   | 0.04   | 0.25   | 0.08   | 0.06   | 0.09   |
| LSTM  | 0.07   | 0.07   | 0.03   | 0.22   | 0.07   | 0.07   | 0.10   |
| SVR   | 0.07   | 0.06   | 0.03   | 0.25   | 0.05   | 0.07   | 0.09   |

From Table 4 and Table 5, we can see that the RMSE of Department4 on the rest day is poor, mainly because of the sudden high frequency and large fluctuation in the latter part; the RMSE of the uncertainty series of flow distribution of each group on other working days or rest days is less than 0.1, which indicates that the forecast results have high reliability. The prediction effect of each prediction model is not very different, and ARIMA model has a slightly better prediction effect. This is because the periodicity of the uncertain sequence of flow distribution is not prominent enough and there are many short-term burst characteristics, so ARIMA model is more suitable for forecasting.

4. Anomaly Detection Experiment

4.1. Scenario Hypothesis and Experimental Data Generation

Two attack scenarios are simulated: DDOS (Distributed Denial of Service) and PROBING (Network Detection Attack). The network flow data on November 25 and 27 of the working day and the rest day are collected under normal conditions, so it is necessary to simulate the anomaly network flow data according to the hypothetical anomaly scenario before anomaly detection, and insert the original data to generate the experimental data of anomaly detection.

Assume that two kinds of attacks occur:

1. Attacks are carried out by multiple hosts, which belong to the same group;
2. The source of DDOS attack is multiple hosts in a group, and the target of DDOS attack is network server.
3. PROBING attacked many hosts in a group, and the target was detectable in the whole network.

The steps of simulating DDOS attacks and PROBING network flows and inserting them into the original normal data to form anomaly detection data are as follows:

1. For DDOS attacks, because DDOS attacks in a real Intranet environment will affect the actual network operating environment and network services, a periodic CBR stream is sent to simulate DDOS attacks using NS2's Constant Bit Rate (CBR) generator. For each attack source, set attack cycle 1 s, attack duration 0.2 s, attack intensity, that is, the amount of traffic used for each attack is 1.5 Mb,
attack number 300, attack duration 300 s. A total of 20 attack sources are simulated, each of which generates 300 s of anomalous network flows for DDOS attacks. The 20 attackers correspond to a random group of 20 network individuals outside the server. The target of the attack is corresponded to the network server in the intranet. The 300 s attacking network flow is inserted into the original 300 s normal network flow of the corresponding 20 network individuals starting from a random time. The time stamp of the attack network flow is added to the original network flow record to form the experimental data for DDOS anomaly detection and verification.

(2) For PROBING, network detection is carried out directly in the real intranet environment, and the network flow generated in the process of network detection is inserted into the network flow data of the date to be detected. Randomly select 20 network individuals from a group other than a server and use the Network Mapper tool to detect the whole intranet with these 20 network individuals as the source of attack for 300 seconds, and record the network flow generated by network detection, i.e. PROBING anomalous network flow. The PROBING anomaly network flow is inserted into the original normal network flow of the corresponding network individuals of the corresponding group starting from a random time of 300 seconds to form the experimental data for PROBING anomaly detection and verification.

DDOS attack network flow is inserted into the network flow data on the working day of November 25, and PROBING attack network flow is inserted into the network flow data on the rest day of November 27, respectively, to form anomaly detection experimental data. In the formation of anomaly detection experimental data, and are randomly generated and stored. If the anomaly detection method proposed in this chapter can correctly detect, and, then the detection method proposed in this chapter is effective. In this experiment, Department 3, 10:44:28 seconds, Department 1, 18:23:56 seconds, that is, the simulated DOS is carried out by the host in department 3 at 10:44:28 seconds on November 15, and the simulated PROBING is carried out by department 1 at 18:23:56 seconds on November 27.

4.2. Test Results Analysis

By using the trained prediction model, the two time series of each group are predicted, and anomalies are judged according to the deviation between the predicted value and the real value. Use the best prediction model in the training data to predict, as shown in Table 6. For the judgment of anomalies, the Theil inequality coefficients of predicted and actual values are used, and the threshold value is 0.2. When the Theil inequality coefficients of both indexes are greater than 0.2, anomalies are considered to occur.

| Table 6 Selection of Prediction Models |
|----------------------------------------|
| Weekday      | n_{ri} | RU_{FlowSize_{ri}} |
|--------------|--------|--------------------|
| Playday      | LSTM   | ARIMA              |

The traffic index and uncertainty index of network flow time series on Nov. 25 are tested. It is found that for Section 3, the sum of Theil inequality coefficients exceeds 0.2 in the 65th time period, i.e. 10:40 to 10:50, as shown in Fig. 1 (a) and (b). The first half of the graph draws the change and forecast of the flow (in megabytes, M) and the uncertainty of the flow distribution of the Department on the third day. The blue is the real curve, the red is the prediction curve, and the corresponding Theil inequality coefficient curve is drawn below. In this period, the predicted flow from Sector 3 is 21639.652M, while the actual flow is 36545.461M and the Theil unequal coefficient is 0.256. It can be considered that there are abnormal outflow flows. The uncertainty of flow distribution is predicted to be 0.344, the actual value is 0.128, and the Theil unequal coefficient is 0.459. The actual value is much lower than the predicted value. Abnormal convergence to a certain purpose, further analysis of the network flow from department 3 during this period, found that a large number of traffic flows to
the server, so it can be judged that DDOS attacks on the server may occur in internal door 3 during this period. The test results are consistent with the simulation scenario.

A. Flow Forecasting and Its Theil Inequality Coefficient

B. Flow Distribution Uncertainty Forecasting and Theil Inequality Coefficient

C. Flow forecasting and its Theil inequality coefficient

D. Uncertainty forecasting of flow distribution and its Theil inequality coefficient

Figure 1 Abnormal test results (c) (d) of department 3 activities on November 25 (a) (b) and department 1 activities on November 27

The traffic index and uncertainty index of network flow time series on Nov. 27 are tested. It is found that for Section 1, the sum of Theil inequality coefficients exceeds 0.2 in the 111th time period, i.e. from 18:20 to 18:30, as shown in Fig. 1 (c) and (d). In this period, the predicted flow from Sector 1 is 1243.208 M, while the actual flow is 1972.198, and the Theil inequality coefficient is 0.227. It can be considered that there are abnormal outflow flows. The uncertainty predicted value of flow distribution is 0.366, the actual value is 0.809, and the Theil inequality coefficient is 0.377. It can be considered that the height dispersion of flow distribution is high. Degree is abnormal. Further analysis of the network flow from Section 1 shows that the target of the network flow is evenly distributed among all groups in the whole network, so it can be judged that PROBING may occur in Group 1 during this period. The test results are consistent with the simulated scenarios of the experiment.

5. Concluding Remarks
This paper presents an anomaly detection method for intranet group interaction based on network flow, including the construction of anomaly detection index, prediction of time series data and detection of anomaly sequence. The proposed method is validated by DDOS attack and PROBING anomaly behavior. Experiments show that the proposed method can effectively detect the anomaly in intranet. Abnormal interaction of groups. Future work includes exploring more effective time series indicators to describe Netflow, and exploring other anomaly detection models to achieve better detection results.
6. References

[1] B. Claise. Specification of the IP Flow Information eXport (IPFIX) protocol for the exchange of ip traffic flow information[R]. RFC5101, 2009.1.

[2] C. Callegari, et al., A Methodological Overview on Anomaly Detection[M]. 2013:148-183.

[3] Lin Zhu, Xingyan. Time series data feature extraction method suitable for classification in data mining [J]. Computer system application. 2012 (10): 224-229.

[4] M. Crovella, A. Lakhina, C. Diot. Diagnosing Network-Wide Traffic Anomalies[C]. ACM SIGCOMM. 2004.8:219-230.

[5] P. Chainais, R. Riedi. On non-scale invariant infinitely divisible cascades[J]. IEEE Trans. On Information Theory, 2005,51(3): 1063-1083.

[6] Crown Male, Li Ou, Analysis and Detection of Unauthorized User Intrusion Behavior in Wireless LAN [J]. Computer Applied Research. Vol.28, No.5, 1873-1877, 2011.

[7] Jingtao, Li Jun, abnormal scanning behavior monitoring system based on numerical analysis [J]. Computer system applications. Vol. 35, No. 1, 49-52, 2014.

[8] Wang Yongchao, Xie Yongkai, Research on the Probability of Intranet Worm Infection Estimated by Flow and Data Packet [J]. Computer Applied Research. Vol.27, No.1, 237-239, 2010.

[9] Cai Jiamei, Cai Qixing, Intranet User Behavior Audit System Based on Genetic Neural Network Analysis [J]. Computer System Applications. No. 5, 5-8, 2009.

[10] Jiang Wei, Qin Yajuan, Liu Ying, Research on User Network Behavior Analysis System Model Based on IPFIX [J]. Computer Technology and Development. Vol.21, No.9, 2011.

[11] G. Cormode, F. Korn, S. Muthukrishnan. Finding Hierarchical Heavy Hitters in Data Streams[C], Proc. of VLDB Conference, 2003.

[12] S. Staniford, J. A. Haagland, J. M. Mcalerney. Practical Automated Detection of Stealthy Port Scans[J]. Journal of Computer Security, Vol.10, 105-136, 2002.

[13] S. Saad, I. Traore. Detecting P2P Botnets through Network Behavior Analysis and Machine Learning[C], International Conference on Privacy, Security and Trust, 2011.

[14] C. Estan, S. Savage, G. Varghese. Automatically Inferring Patterns of Resource Consumption in Network Traffic[C], Proc. Of ACM SIGCOMM, 2003.

[15] M. Bakoben, N. Adams, A. Bellotti. Uncertainty Aware Clustering for Behavior in Enterprise Networks[C], IEEE 16th International Conference on Data Mining Workshops, 2016.

[16] K. Xu, Z. L. Zhang, S. Bhattacharyya. Profiling Internet Backbone Traffic: Behavior Models and Applications[C], SIGCOMM, 2005.