A New Anomaly Detection Method Based on IGTE and IGFE

Ziyu Wang(✉), Jiahai Yang, and Fuliang Li

Tsinghua National Laboratory for Information Science and Technology (TNList),
Institute for Network Sciences and Cyberspace,
Tsinghua University, Beijing 100084, China
wangziyu11@mails.tsinghua.edu.cn, yang@cernet.edu.cn,
lifuliang207@126.com

Abstract. Network anomalies have been a serious challenge for the Internet nowadays. In this paper, two new metrics, IGTE (Inter-group Traffic Entropy) and IGFE (Inter-group Flow Entropy), are proposed for network anomaly detection. It is observed that IGTE and IGFE are highly correlated and usually change synchronously when no anomaly occurs. However, once anomalies occur, this highly linear correlation would be destroyed. Based on this observation, we propose a linear regression model built upon IGTE and IGFE, to detect the network anomalies. We use both CERNET2 netflow data and synthetic data to validate the regression model and its corresponding detection method. The results show that the regression-based method works well and outperforms the well known wavelet-based detection method.

Keywords: Anomaly detection · Regression · IGTE · IGFE

1 Introduction

Network anomalies have been a serious challenge for the Internet nowadays. There are basically two classes of detection methods. The first class is called misused detection, also known as signature-based detection [10,12,17,19,25]. The primary advantage of misused detection is its high degree of accuracy. However, the misused detection is incapable of detecting emerging anomalies (zero day attacks), whose features are not known in advance. The second class of detection methods is called anomaly detection [1,5,16,23]. Anomaly detection typically derives a normal model of the network data, then computes an "outlier score" for each data point. The normal model is usually derived from different quantities of the network traffic, such as the number of packets, the number of bytes, the number of flows, etc. Outlier score is a measure about the level of "outlierness" of each data point, based on the deviation distance from the normal model. The concept of outlier score is similar to residual which is commonly used in the field of anomaly detection. In this paper, we will treat these two concepts equivalently without distinction. Once certain outlier score exceeds predefined

J. Tian et al. (Eds.): SecureComm 2014, Part II, LNICST 153, pp. 93–109, 2015.
DOI: 10.1007/978-3-319-23802-9_10

threshold, an alarm is triggered. Since anomaly detection only cares about the statistical properties of network traffic rather than specific anomaly features, it is capable of detecting zero day attacks. This capability is the strong advantage of anomaly detection over misuse detection. Hence, anomaly detection has been well studied by researchers in recent years [24,27,30,32].

The wavelet analysis is widely applied in anomaly detection [7,9,11,26,28]. Barford et al. [2] first introduce wavelet techniques into the field of network anomaly detection. They first use wavelets filters to decompose single-link traffic into three parts: low-frequency part, mid-frequency part, high-frequency part, and then they use the local variances of mid-frequency part and high-frequency part to generate a V-signal, then apply thresholding to the V-signal to detect anomalies. The basic idea of the wavelet-based detector is to compare local variance with global variance. However, it ignores the fact that the variance of network traffic is usually proportional to the absolute volume of network traffic. High traffic volume usually corresponds to massive active users in the network. Therefore, large local variance is more likely to be a result from normal network behavior rather than anomalies. Unfortunately, the wavelet-based detector ignores this fact and is prone to generate false positives.

Given the shortcomings of wavelet-based detector, we propose a new anomaly detection method based on two new metrics—IGTE and IGFE. These two metrics are basically entropies summarizing the distribution of the traffic volume and the number of IP flows among different groups. We focus on the relation between IGTE and IGFE rather than the variance, which makes this new method unaffected by the absolute network traffic volume. First, we randomly map the network flows which constitute the network traffic into fixed number of groups. The number of bytes and the number of network flows are calculated for each group. Consequently, we obtain two matrices, which are called Randomly Aggregated Traffic Matrix (RATM) and Randomly Aggregated Flow Matrix (RAFM). It is assumed that the distribution of the traffic volume among different groups and the distribution of the number of flows should resemble each other. Then we calculate two types of entropies based on the columns of RATM and RAFM respectively. These two entropies are called Inter-group Traffic Entropy (IGTE) and Inter-group Flow Entropy (IGFE). It is found that IGTE and IGFE are highly correlated under normal condition, and when anomalies occur, this correlation will be destroyed. Based on this observation, we propose a regression-based detection method. Using CERNET2 Netflow data and synthetic data, we validate that our regression-based detector is capable of achieving high detection rate and low false positive rate.

The main contributions of this paper are: (1) putting forward two new metrics—IGTE and IGFE—which are effective for anomaly detection, (2) validating the highly linear correlation between IGTE and IGFE, (3) proposing a new effective regression-based anomaly detection method built upon IGTE and IGFE, (4) analyzing the shortcomings of wavelets-based detection method [2].

The remainder of this paper is organized as follows. Section 2 presents related work in the field of anomaly detection. In Sect. 3, we introduce the procedure of generating RATM and RAFM. In Sect. 4, we illustrate how to derive two new metrics—IGTE and IGFE—from RATM and RAFM, and show the highly linear correlation between them. We explain the principle and rationale of the regression based detector in Sect. 5. In Sect. 6, we compare the regression-based detector and the famous wavelets-based detector by using both CERNET2 Netflow data and synthetic data. We conclude this work in Sect. 7.

2 Related Work

In recent years, lots of researches have been devoted to the field of anomaly detection. Yaacob et al. [29] introduce a new approach through using Auto-Regressive Integrated Moving Average (ARIMA) technique to detect potential attacks in the network. Although they show the capability of ARIMA model of predicting future data, their validation process is rough, and the threshold they choose is heuristic.

Silveira et al. [22] state that when many network flows are multiplexed on a non-saturated link, their volume changes tend to cancel each other out over short timescales, making the average change across flows approximately follows the normal distribution. Based on this observation, they propose the ASTUTE-based anomaly detector. While it is good at detecting anomalies which involve many small IP flows, it fails to detect anomalies caused by a few large IP flows. Besides that, the efficacy of the ASTUTE-based detector highly depends on the stationarity of network traffic. The authors claim that at short timescales (less than a hour), the traffic can be well modeled by stationary processes. However, this conclusion does not always holds for all networks. It performs poorly in those networks in which IP flows are changeable. For example, the CERNET2 Netflow data used in this paper contains many IP flows which emerge and vanish quite suddenly. The ASTUTE-based detector marks almost every point in the data set as anomalous, which is practically impossible in real world. We manually check Netflow data and consult the operators of CERNET2, it turns out that most of the anomalies detected by the ASTUTE-based detector are false positives. Therefore, we do not adopt ASTUTE-based detector as comparison in this paper.

Lakhina et al. [14,15] first apply principal component analysis (PCA) in network-wide anomaly detection. PCA-based detector (also referred to as subspace-based detector) uses the first few principal components to derive normal model from the original link traffic matrix, and then applies thresholding to the residual traffic to detect anomalies. The advantage of the PCA method is its capability of detecting small anomalies distributed over multiple links which are hard to detect in single-link traffic. Since this method is applied to link traffic matrix, it is limited to the network-wide anomaly detection. Besides, there are some inherent weaknesses of PCA based detector. For example, a large anomaly may inadvertently pollute the normal subspace, the effectiveness of PCA is sensitive to the level of aggregation of the traffic measurements, and the false positive rate is sensitive to small differences in the number of principal components in the normal subspace [18,31]. Rubinstein et al. [21] show that the attackers can successfully evade PCA-based detection by only adding moderate amounts of poisoned data. Besides, since all kinds of PCA-based detectors need to operate on the link traffic matrix, it is necessary to collect data from all links in the networks simultaneously. However, it is usually a difficult task for large networks. The lack of scalability limits the application of PCA-based detectors. In this paper, we only focus on anomaly detection for single-link traffic.

3 RATM and RAFM

Before introducing RATM and RAFM, we give the definition of IP flow here for the purpose of illustration. In practice, an IP flow can be defined in multiple ways according to different contexts. In this study, an IP flow is defined as a sequence of packets that share the same five-tuple value (Source IP address, Destination IP address, Source port, Destination port and Protocol type).

We select the five-tuple values of IP flows as key, and hash them into fixed number of groups. The number of groups are selected by the network operators according to the needs. For each group, we calculate the overall traffic volume of the IP flows mapped into it during each time interval, then the RATM is generated. The rows of RATM correspond to different time intervals, the columns correspond to different groups. In detail, the (i, j) entry of RATM corresponds to the traffic volume of group j at time instant i. Similarly, the RAFM is generated by counting the number of IP flows in each group during each time interval. Thus, the (i, j) entry of RAFM corresponds to the number of IP flows in group j at time instant i.

4 Two New Metrics—IGTE and IGFE

Intuitively, for a given group, under normal condition, the more IP flows mapped into the group, the higher traffic volume would be contained in that group. Consequently, the distribution of traffic volume among different groups should resemble to the distribution of number of IP flows. Entropy can be used as a summarization tool for probability distributions from the point of view of information theory [15]. Thus we calculate the entropies for the rows of RATM and RAFM respectively. The entropy for RATM is named Inter Group Traffic Entropy (IGTE), the entropy for RAFM is named Inter Group Flow Entropy (IGFE). The details for calculating IGTE and IGFE are given as follows.

Suppose a $t \times p$ RATM T, where t is the number of time intervals considered, p is the number of groups predefined. For a given row i of T, the definition of IGTE is defined as follows:

$$IGTE_i = -\sum_{j=1}^{p} \left\{ \frac{T(i,j)}{\sum_{j=1}^{p} T(i,j)} \ln \frac{T(i,j)}{\sum_{j=1}^{p} T(i,j)} \right\} \tag{1}$$

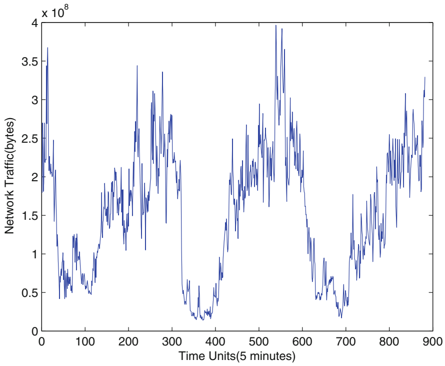

Fig. 1. Network traffic from CERNET2

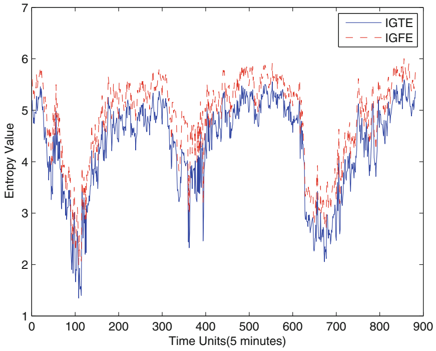

Fig. 2. IGFE series versus IGTE series from CERNET2

where $T(i,j)$ is the (i,j) entry of T, $1 \leq i \leq t$, $1 \leq j \leq p$. Similarly, for a given row i of a RAFM denoted by F, the definition of IGFE is defined as follows:

$$IGFE_i = -\sum_{j=1}^{p} \left\{ \frac{F(i,j)}{\sum_{j=1}^{p} F(i,j)} \ln \frac{F(i,j)}{\sum_{j=1}^{p} F(i,j)} \right\} \tag{2}$$

where $F(i,j)$ is the (i,j) entry of F, $1 \leq i \leq t$, $1 \leq j \leq p$. Therefore, based on RATM and RAFM, we can obtain an IGTE series and an IGFE series respectively. Note that IGTE and IGFE are essentially entropies. If the distribution of traffic volume and that of the number of flows resemble to each other, IGTE and IGFE should be highly correlated. In order to validate this conjecture, we calculate the IGTE and IGFE series from approximately three-day network traffic obtained from CERNET2 (an academic network in China which will be described in detail later) which is shown in Fig. 1, and plot the IGTE and IGFE series in Fig. 2. It is shown that the curve of IGFE series is extremely similar to the curve of IGTE series, which implies that IGFE and IGTE are highly linearly correlated. To verify this conjecture rigorously, we calculate the correlation coefficient between these two series. The result is 0.976, which means IGTE and IGFE are indeed highly linearly correlated. Note that the three-day network traffic may contain anomalies which are not known a priori. Even so, the linear relationship between IGTE and IGFE is strong enough. This observation lays the foundation of our regression based detector.

5 Detection Methods

5.1 Regression-Based Detection

Based on IGTE and IGFE, we propose a new anomaly detection method using linear regression analysis. The goal of regression analysis is to construct mathematical models which describe relationships that may exist between variables [20]. Usually, we are interested in just one variable, i.e. the response variable, and we want to study how it depends on a set of variables which are called explanatory variables.

Let y denote the response variable, $x_1, x_2, \ldots, x_p$ denote the set of explanatory variables. Denote the samples from y as $Y = (y_1, y_2, \ldots, y_n)^T$, the samples from $x_1, x_2, \ldots, x_p$ as $X_e = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix}$. The goal of regression analysis is to obtain the relationship of dependency between y and $x_1, x_2, \ldots, x_p$.

Regression analysis assumes y and $x_1, x_2, \ldots, x_p$ satisfy the following linear regression equation:

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + e \tag{3}$$

where $e \sim N(0, \sigma^2)$, $\sigma, \beta_0, \beta_1, \ldots, \beta_p$ are parameters to be determined. From Eq. (3), the corresponding samples should satisfy the following equation:

$$Y = X\beta + E \tag{4}$$

where $X = (1, X_e)$ is defined as the extended matrix of X_e, $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T, E \sim N(0, \sigma^2 I)$.

Define $Q(\beta) = \sum_{i=1}^{n} \{y_i - (\beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip})\}^2 = ||Y - X\beta||^2$, then $Q(\beta)$ measures the noise of the regression equation. The optimal estimate of β should make $Q(\beta)$ as small as possible. Thus the estimate of β is as follows:

$$\hat{\beta} = (X^T X)^{-1} X^T Y \tag{5}$$

Then the estimate of σ^2 is as follows:

$$s^2 = \frac{1}{n-p-1} Q(\hat{\beta}) \tag{6}$$

We define the normal model of Y as follows:

$$\hat{Y} = X\hat{\beta} \tag{7}$$

Then the estimate of E is

$$\hat{E} = Y - \hat{Y} \tag{8}$$

Note that $\hat{E}$ is the outlier scores, i.e. residuals, of all data points. Intuitively, if the residual of a given data point is close to 0, the data point would be normal, otherwise, the point would be abnormal.

The procedure of detection is as follows. First, we calculate $\hat{E}$ from the samples as described above. For convenience, we denote $\hat{E}$ as $(\hat{e}_1, \hat{e}_2, \ldots, \hat{e}_n)$. For a given sample point i, where $1 \leq i \leq n$, under normal condition, $\hat{e}_i$ should follows the normal distribution $N(0, s^2)$, based on the assumption of Eq. (3). For a given confidence level $1 - \alpha$, if $|\frac{\hat{e}_i}{s}| > z_{\alpha/2}$, where $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of standard normal distribution $N(0, 1)$, the data point i is marked as an anomaly. The meaning of the confidence level is that when a data point is marked as an anomaly, the probability of being a false alarm is α.

Note that the success of the regression model depends greatly on the linear correlation between the response variable and the set of explanatory variables. Given the discussion in Sect. 4, IGTE and IGFE are highly linearly correlated. Therefore, we choose IGTE as the response variable, IGFE as the explanatory variable in this study. Let y denote IGTE and x denote IGFE. The regression equation built upon them is give below.

$$y = \beta_0 + \beta_1 x + e \tag{9}$$

where $e \sim N(0, \sigma^2)$, σ, β_0 and β_1 are parameters to be determined.

The details of our regression based detector is summarized in Algorithm 1 .

Note that there is an important auxiliary procedure which is not illustrated in Algorithm 1 due to space limitations. After $Y = (IGTE_1, IGTE_2, \ldots, IGTE_t)^T$ and $X = (1, X_e)$ are calculated, i.e. after step 7, it is necessary to test rigorously whether it is appropriate to build regression equation upon them. In other words, we must test whether the dependence of Y on X is strong enough for the correctness of the regression model. There are two kinds of significance tests for this: F test and t test [20]. Only when the data passes both tests, the corresponding regression model can be considered reasonable.

5.2 Rationale Behind Regression-Based Detection Method

Network traffic consists of IP flows. Anomalies usually change the number of IP flows on the link or the traffic volume of certain IP flows. Some anomalies such as port scans, would generate lots of small IP flow in the network. This leads to large increase in the number of IP flow, which makes the IGFE change dramatically. However, the traffic generated by the anomalies is very small compared to the overall traffic volume on the link, which barely changes the IGTE value. Therefore, the linear correlation between IGTE and IGFE is destroyed, and the regression-based detector generates large residual to trigger alarms.

Some anomalies such as DDoS attacks, would increase the number of IP flows and the traffic volume at the same time. However, the magnitude of traffic volume change is much larger than the number of IP flows. Hence, the degree of

Algorithm 1. Regression based anomaly detector

```
Input: $t \times p$ RATM; $t \times p$ RAFM; $z_{\alpha/2}$;
Output: Time intervals containing anomalies;
for all $i$ such that $1 \le i \le t$ do
    $IGTE_i = -\sum_{j=1}^{p} \left\{ \frac{T(i,j)}{\sum_{j=1}^{p} T(i,j)} \ln \frac{T(i,j)}{\sum_{j=1}^{p} T(i,j)} \right\}$;
    $IGFE_i = -\sum_{j=1}^{p} \left\{ \frac{F(i,j)}{\sum_{j=1}^{p} F(i,j)} \ln \frac{F(i,j)}{\sum_{j=1}^{p} F(i,j)} \right\}$;
end for
$Y = (IGTE_1, IGTE_2, \ldots, IGTE_t)^T$;
$X_e = (IGFE_1, IGFE_2, \ldots, IGFE_t)^T$;
$X = (1, X_e)$;
$\hat{\beta} = (X^T X)^{-1} X^T Y$;
$\hat{E} = (\hat{e}_1, \hat{e}_2, \ldots, \hat{e}_t)^T = Y - X \times \hat{\beta}$;
$Q(\hat{\beta}) = \sum_{i=1}^{n} \{y_i - (\beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip})\}^2$;
$s = \sqrt{\frac{1}{n-p-1} Q(\hat{\beta})}$;
$\hat{E} = \hat{E}/s$;
for $i = 1$ to $t$ do
    if $|e_i| > z_{\alpha/2}$ then
        Output: Time interval i;
    end if
end for
```

change of IGTE is much large than IGFE. It results in the breach of the linear relation between IGTE and IGFE, and the anomalies would be detected by the regression-based detector.

There are also some anomalies which would increase the number of IP flows and decrease the traffic volume. Take Low-rate DDoS attacks [13] for example, the attackers would generate millions of attacking IP flows, which will definitely change the IGFE value. On the other hand, the traffic volume generated by the attacking IP flows is very low on average, since these attacks are performed in the form of pulses. At the same time, the traffic volume of the normal IP flows would be reduced dramatically due to the congestion control mechanism in network. Therefore the overall traffic on the link would decrease dramatically, which would cause the change of IGTE value. Though both IGTE and IGFE change, they change in opposite directions, which destroys the linear relationship between them. Hence, these anomalies can be detected by the regression-based detector.

6 Validation

6.1 Dataset

The data used in this paper is Netflow Records collected from the Second Generation of China Education and Research Network (CERNET2). CERNET2 connects 25 PoPs including Peking University, Tsinghua University, Beijing

University of Aeronautics and Astronautics (Beihang University), University of Science and Technology, etc. The Netflow data is collected from a border router connecting CERNET2 backbone and Beihang University Campus Network. The data collection architecture is shown in Fig. 3. The Netflow V9 protocol [4] is used to collect the data passing through the border router (i.e. Netflow exporter), and transfer the Netflow records to a storage server. The sampling rate is set to 1 : 1000. Lots of information for each IP flow within every five minutes are saved, including the five-tuple value, the total number of bytes and packets, the starting time and finishing time, etc. The average traffic volume in five minutes is about 1.525×10^8 bytes. The average traffic volume of each IP flow is about 985 bytes. The average number of IP flows is about 154730. Note that these numbers are calculated from the sampled data. The numbers should be multiplied by 1000 for the un-sampled data.

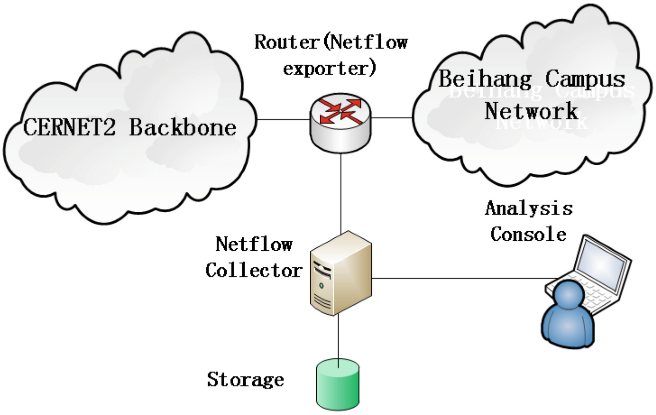

Fig. 3. Data collection architecture

We collected Netflow records from a border router connecting CERNET2 backbone and Beihang University campus network from 21:45 in August 26 to 23:10 in August 29, 2013. The corresponding network traffic is already shown in Fig. 1. We set the number of groups as 1024. Since the Netflow records are stored every five minutes, there are totally 882 time units during the data collection period, then a 882 × 1024 RATM and a 882 × 1024 RAFM are generated respectively.

In this paper, we use the CERNET2 data by two different means. One is to directly apply the detection methods on the CERNET2 data. The other one is to manually inject anomalies into the "cleaned" CERNET2 data, and then apply the detection methods on this synthesized data. The advantage of using CERNET2 data directly is that it can compare the performance of different detection methods in real networks. The advantage of using synthetic data is the capability of obtaining the detection rate and false positive rate by controlling the process of injecting anomalies.

6.2 Validation Using Real World Data

From the 882 × 1024 RATM and RAFM above, an IGTE series and an IGFE series—both of length 882—are obtained. Their curves are already presented in Fig. 2. We choose IGTE as the response variable, denoted by y, and IGFE as the explanatory variable, denoted by x. From Eqs. (5), (6) and (9), we have:

$$y = -0.72631 + 1.09724x + e \tag{10}$$

where $e \sim N(0, 0.048^2)$. Next, we check the significance of regression Eq. (10). Recall that both F test and t test are used in this work for significance test. We use the famous statistical software R [3] to do the tests. We set the confidence level as $1 - \alpha = 1 - 0.05 = 0.95$. The $p - value$ of F test outputted by R is 0.26×10^{-8}, which is much less than $\alpha = 0.05$. $P - value$ is a commonly used metric in hypothesis testing [6]. If the $p - value$ is less than α, the regression model is accepted as valid. The $p - value$ of F test means that Eq. (10) fits the data quite well. The resulting $p - value$ of t test is 0.11×10^{-9}, which is again much less than $\alpha = 0.05$, which also means that Eq. (10) is appropriate for the data. The residual related to Eq. (10) is illustrated in Fig. 4. We check these residual data points according to the detection procedure described in subsect. 5.1, and mark the abnormal points in red circles. There are totally 53 anomalies detected.

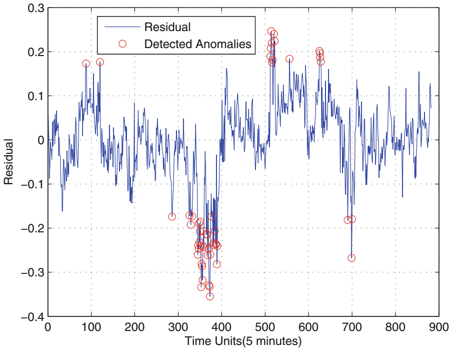

Fig. 4. Anomalies detected by regression-based detector for CERNET2 data (Color figure online)

As a comparison, we apply the wavelets-based detector [2] to the same data set. We set the sliding window length as 12. This window size corresponds to one hour traffic. Thus the output of wavelets-detector, i.e. “deviation scores”, does not contain the first 11 points in CERNET2 data. In other words, the output size of wavelets-detector is $882 - 11 = 871$. The results are shown in Fig. 5, the red circles point out the 60 anomalies detected.

Comparing Figs. 4 and 5, we find that only 3 anomalies are detected commonly by both detection methods. Does that mean that our regression-based

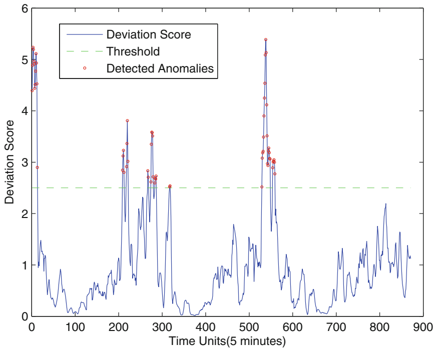

Fig. 5. Anomalies detected by wavelets-based detector for CERNET2 data (Color figure online)

detection method is ineffective? We argue that the wavelets-based detector has its own limits—it ignores the fact that the variance of network traffic is usually proportional to the absolute volume of network traffic (as shown in Fig. 1). Comparing Figs. 1 and 5, it is observed that the anomalies detected by wavelets-based detector coincide with the time intervals in which traffic volume is high. It is known that high traffic volume usually corresponds to massive active users in the network, and the variance of network traffic becomes large accordingly. In other words, large local variance is more likely to be a result from normal network behavior rather than anomalies. Unfortunately, the wavelets-based detector ignore this fact and mark these data points as anomalies arbitrarily. Thus we have reason to believe that most of the alarms triggered by wavelets-based detector are likely to be false alarms. We manually check the Netflow data and also consult the operators who run CERNET2, it turns out that there were no sign of large-scale attacks during the data collecting period, which supports our statement. In the following subsection, we will validate this claim quantitatively and rigorously.

6.3 Validation Using Synthetic Data

To evaluate the performance of different anomaly detection methods rigorously and quantitatively, we manually inject anomalies into the "cleaned" CERNET2 Netflow data. The detail is as follows: first, we abandon those time intervals which are marked by regression-based detector or wavelets-based detector. The remaining 772 time intervals are considered as "clean" traffic. In other words, we assume that these 772 time intervals contain no anomalies. Since the only two detectors applied on the "clean" traffic are regression-based detector and wavelets-based detector, this assumption makes sense. Then, we manually inject certain number of anomalous IP flows every 22 time intervals. Thus the total number of injected anomalies is 35. In the area of anomaly detection, a general assumption is that the anomalies contained in the data are much less than the normal points. Thus the number of injected anomalies we choose is reasonable.

According to the number of anomalous IP flows injected, we evaluate the performance of detection methods in two cases:

- Anomalies involving a small number of IP flows.
- Anomalies involving many small IP flows.

In the first case, we focus on the impact of the traffic volume of injected anomalies. We inject 11 anomalous IP flows and gradually increase their traffic volume. The true positive rate (detection rate) curves and false positive rate curves are shown in Figs. 6 and 7 respectively. The horizontal coordinates represent the proportion of the anomalous traffic volume in the total traffic volume of the link. The vertical coordinates represent the true positive rates and the false positive rates. The definitions of true positive rate and false positive rate in this paper originate from the introductory document about ROC analysis [8]. It is illustrated that as the anomalous traffic volume increases, the detection rate of regression-based detector rises sharply, and the false positive rate falls quickly. When the anomalous traffic volume reaches 42.8 % of the total traffic volume, the regression-based detector can detect all the injected anomalies while generate no false alarms. It means that the regression-based detector is good at detecting anomalies involving a few large IP flows for which the ASTUTE-based detector performs poor [22]. However, for the wavelets-based detector, the detection rate increases very slowly. Even when the anomalous traffic volume reaches 60 %, which means the order of magnitude of anomalous traffic volume reaches around 10^9 bytes, the detection rate is below 5.8 %. This result is unacceptable for practical application. The performance of wavelet-based detector in the point of view of false positive rate is also poor. As the anomalous traffic increases, the false positive rate is hardly decreasing, and converges to around 2.4 %. In contrast, the false positive rate of the regression-based detector falls down quickly. When the anomalous traffic volume reaches 15 %, no false positive is generated.

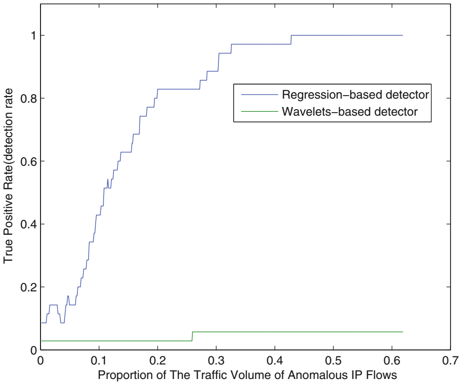

Fig. 6. True positive rate for a small number of anomalous IP flows

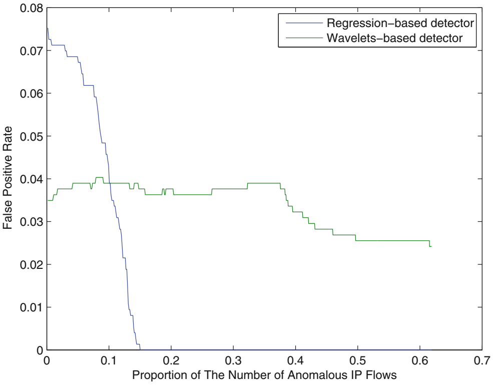

Fig. 7. False positive rate for a small number of anomalous IP flows

In the second case, we focus on the impact of the number of anomalous IP flows. We simulate the scenario of DDoS attacks. We first set the traffic volume of each injected IP flow as 50 bytes. Considering the average traffic volume of each IP flow in CERNET2 is around 985 bytes, the traffic volume per anomalous flow we choose is reasonable and small. Then we gradually increase the number of injected IP flows. The detection rates and false positive rates of the two detectors are shown in Figs. 8 and 9. Note that the horizontal coordinates here represent the proportion of the number of anomalous IP flows in the total number of IP flows in the link. For the regression-based detector, as the number of anomalous flows grows, the detection rate curve rises sharply and the false positive rate curve falls quickly. When the proportion of injected IP flows reaches 41 % of the total number of IP flows, the detect rate reaches 80 %. The false positive rate reaches 0 when the number of anomalous IP flows reaches no more than 4 %. On the other hand, for the wavelets-based detector, the detection rate keeps below 3 % and does not grow with the number of anomalous flows. The false positive rate of the wavelet-based detector keeps around 4 %, which seems acceptable at first. However, when we look deeper into the anomalies marked by the wavelets-based detector, we find that the number of false positives keeps around 34 and the number of true positives keeps close to 1. This observation holds no matter how much the proportion of the number of injected IP flows accounts for. Comparing the amount of false positives and the amount of true positives it detect, the performance is really poor. We also try other values of traffic volume for each injected IP flow, the results are similar.

For both cases, we find that the alarms generated by the wavelets-based detector again coincide with the time intervals with high traffic volume. This observation strongly support our reasoning about the shortcomings of wavelets-based detector—it ignores the fact that large local variances are usually related to the high traffic volume generated by normal users.

In summary, based on the synthetic CERNET2 data, our regression-based detector achieves higher detection rate and lower false positive rate than the

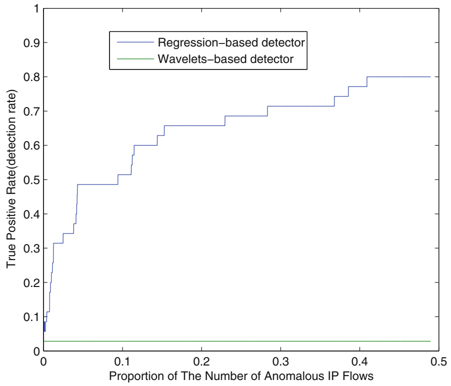

Fig. 8. True positive rate for a large number of anomalous IP flows

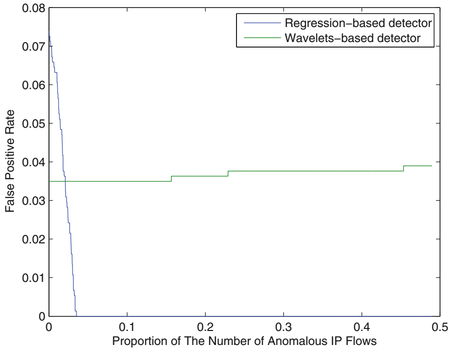

Fig. 9. False positive rate for a large number of anomalous IP flows

wavelets-based detector. Besides, the regression-based detector is good at detecting both anomalies involving a few large IP flows and anomalies involving many small IP flows. Note that we ignore the scenario where anomalies involve many large IP flows on purpose. Because in this case, the volume of the network traffic would change so much that the anomalies can be identified even by the naked eyes. Both the wavelet-based detector and the regression-based detector perform excellently in this case. There is no need to show the experiment results in this case for the sake of brevity.

7 Conclusions and Future Work

In this paper, we propose two new metrics, IGTE and IGFE, for anomaly detection. It is found that IGTE and IGFE are highly linearly correlated. When anomalies occur, this linear correlation will be destroyed. Based on this observation, we propose the regression based detector which is built upon IGTE and IGFE. We validate that the regression based detector can achieve high detection rate and generate very few false positives. We show that the regression-based detector is good at detecting both anomalies involving a few large IP flows and anomalies involving many small IP flows. We compare the regression based detector with the wavelet-based detector, and find that the former outperforms the latter. We analyze the reason for the failure of wavelets-based detector. The wavelets-based detector uses local variance of traffic volume to measure the degree of anomaly. However, large local variance are usually caused by large number of normal users. Thus the wavelets-based detector usually generates too many false positives. We do not deny the possibility that the CERNET2 data used in this paper bias for the regression-based detector while bias against the wavelet-based detector. In the future, we plan to use more data sources to validate the regression-based detector.

Acknowledgments. We are grateful to Lujing Sun for providing us with the Netflow data from CERNET2, and to Kun Wen and Chenxi Li for many helpful discussions. We also thank anonymous reviewers for their constructive comments. This work is supported by the National Basic Research Program of China under Grant No. 2012CB315806, the National Natural Science Foundation of China under Grant No. 61170211, 61202356, 61161140454, Specialized Research Fund for the Doctoral Program of Higher Education under Grant No. 20110002110056, 20130002110058.

References

1. Andrysiak, T., Saganowski, Ł., Choraś, M.: DDoS attacks detection by means of greedy algorithms. In: Choraś, R.S. (ed.) Image Processing and Communications Challenges 4. AISC, vol. 184, pp. 301–308. Springer, Heidelberg (2013)
2. Barford, P., Kline, J., Plonka, D., Ron, A.: A signal analysis of network traffic anomalies. In: Proceedings of the 2nd ACM SIGCOMM Workshop on Internet Measurment, pp. 71–82. ACM (2002)
3. The r project for statistical computing. http://www.r-project.org/
4. Cisco systems netflow services export version 9. http://www.rfc-base.org/rfc-3954.html
5. Brauckhoff, D., Salamatian, K., May, M.: Applying pca for traffic anomaly detection: Problems and solutions. In: INFOCOM 2009, pp. 2866–2870. IEEE (2009)
6. Casella, G., Berger, R.L.: Statistical Inference. Duxbury Press, Belmont (1990)
7. Cong, F., Hautakangas, H., Nieminen, J., Mazhelis, O., Perttunen, M., Riekki, J., Ristaniemi, T.: Applying wavelet packet decomposition and one-class support vector machine on vehicle acceleration traces for road anomaly detection. In: Guo, C., Hou, Z.-G., Zeng, Z. (eds.) ISNN 2013, Part I. LNCS, vol. 7951, pp. 291–299. Springer, Heidelberg (2013)
8. Fawcett, T.: An introduction to ROC analysis. Pattern Recogn. Lett. **27**(8), 861–874 (2006)
9. Guzman, J., Poblete, B.: On-line relevant anomaly detection in the twitter stream: an efficient bursty keyword detection model. In: Proceedings of the ACM SIGKDD Workshop on Outlier Detection and Description, pp. 31–39. ACM (2013)
10. Jamshed, M.A., Lee, J., Moon, S., Yun, I., Kim, D., Lee, S., Yi, Y., Park, K.: Kargus: a highly-scalable software-based intrusion detection system. In: Proceedings of the 2012 ACM Conference on Computer and Communications Security, pp. 317–328. ACM (2012)

11. Jiang, D., Zhang, P., Xu, Z., Yao, C., Qin, W.: A wavelet-based detection approach to traffic anomalies. In: 2011 Seventh International Conference on Computational Intelligence and Security (CIS), pp. 993–997. IEEE (2011)
12. Jiang, H., Zhang, G., Xie, G., Salamatian, K., Mathy, L.: Scalable high-performance parallel design for network intrusion detection systems on many-core processors. In: Proceedings of the Ninth ACM/IEEE Symposium on Architectures for Networking and Communications Systems, pp. 137–146. IEEE Press (2013)
13. Kuzmanovic, A., Knightly, E.W.: Low-rate tcp-targeted denial of service attacks: the shrew vs. the mice and elephants. In: Proceedings of the 2003 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications, pp. 75–86 (2003)
14. Lakhina, A., Crovella, M., Diot, C.: Diagnosing network-wide traffic anomalies. ACM SIGCOMM Comput. Commun. Rev. **34**, 219–230 (2004). ACM
15. Lakhina, A., Crovella, M., Diot, C.: Mining anomalies using traffic feature distributions. ACM SIGCOMM Comput. Commun. Rev. **35**, 217–228 (2005). ACM
16. Palmieri, F., Fiore, U.: Network anomaly detection through nonlinear analysis. Comput. Secur. **29**(7), 737–755 (2010)
17. Paxson, V.: Bro: a system for detecting network intruders in real-time. Comput. Netw. **31**(23), 2435–2463 (1999)
18. Ringberg, H., Soule, A., Rexford, J., Diot, C.: Sensitivity of PCA for traffic anomaly detection. ACM SIGMETRICS Perform. Eval. Rev. **35**, 109–120 (2007). ACM
19. Roesch, M., et al.: Snort: Lightweight intrusion detection for networks. In: LISA, pp. 229–238 (1999)
20. Ross, S.M.: Introductory statistics. Academic Press (2010)
21. Rubinstein, B.I., Nelson, B., Huang, L., Joseph, A.D., Lau, S.h., Rao, S., Taft, N., Tygar, J.: Antidote: understanding and defending against poisoning of anomaly detectors. In: Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement Conference, pp. 1–14. ACM (2009)
22. Silveira, F., Diot, C., Taft, N., Govindan, R.: Astute: detecting a different class of traffic anomalies. ACM SIGCOMM Comput. Commun. Rev. **40**(4), 267–278 (2010)
23. Simmross-Wattenberg, F., Asensio-Perez, J.I., Casaseca-de-la Higuera, P., Martin-Fernandez, M., Dimitriadis, I.A., Alberola-López, C.: Anomaly detection in network traffic based on statistical inference and alpha-stable modeling. IEEE Trans. Dependable Secure Comput. **8**(4), 494–509 (2011)
24. Soldo, F., Metwally, A.: Traffic anomly detection based on the IP size distribution. In: 2012 Proceedings IEEE INFOCOM, pp. 2005–2013 (2012)
25. Vasiliadis, G., Polychronakis, M., Ioannidis, S.: Midea: a multi-parallel intrusion detection architecture. In: Proceedings of the 18th ACM Conference on Computer and Communications Security, pp. 297–308. ACM (2011)
26. Wang, W., Lu, D., Zhou, X., Zhang, B., Mu, J.: Statistical wavelet-based anomaly detection in big data with compressive sensing. EURASIP J. Wireless Commun. Networking **2013**(269), 1–6 (2013)
27. Winter, P., Lampesberger, H., Zeilinger, M., Hermann, E.: On detecting abrupt changes in network entropy time series. In: De Decker, B., Lapon, J., Naessens, V., Uhl, A. (eds.) CMS 2011. LNCS, vol. 7025, pp. 194–205. Springer, Heidelberg (2011)
28. Wu, J., Cui, Z., Shi, Y., Su, D.: Traffic flow anomaly detection based on wavelet denoising and support vector regression. J. Algorithms Comput. Technol. **7**(2), 209–226 (2013)

29. Yaacob, A.H., Tan, I.K., Chien, S.F., Tan, H.K.: Arima based network anomaly detection. In: Second International Conference on Communication Software and Networks. ICCSN 2010, pp. 205–209. IEEE (2010)
30. Zhang, B., Yang, J., Wu, J., Qin, D., Gao, L.: Mcst: Anomaly detection using feature stability for packet-level traffic. In: 2011 13th Asia-Pacific Network Operations and Management Symposium (APNOMS), pp. 1–8. IEEE (2011)
31. Zhang, B., Yang, J., Wu, J., Qin, D., Gao, L.: Pca-subspace method is it good enough for network-wide anomaly detection. In: 2012 IEEE Network Operations and Management Symposium (NOMS), pp. 359–367. IEEE (2012)
32. Zhang, B., Yang, J., Wu, J., Wang, Z.: Mbst: detecting packet-level traffic anomalies by feature stability. Comput. J. **56**(10), 1176–1188 (2013)