SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

September 2014, Vol. 57 092204:1–092204:11
doi: 10.1007/s11432-014-5087-7

Uncovering network traffic anomalies based on their sparse distributions

CHENG GuoZhen*, CHEN HongChang, CHENG DongNian,
ZHANG Zhen & LAN JuLong

National Digital Switching System Engineering and Technological Research Center, Zhengzhou 450002*, China*

Received June 21, 2013; accepted December 14, 2013; published online April 21, 2014

Abstract Characterizing network traffic with higher-dimensional features results in increased complexity of most detectors and classifiers for identifying traffic anomalies. Several key observations from existing studies confirm that network anomalies are typically distributed in a sparse way, with each anomaly essentially characterized by its lower-dimensional features. Based on this important finding, we exploit sparsity in designing a novel detection method for anomalies that ignores redundancies that are dynamically filtered from the feature sets and accurately classifies anomalies. Comparison of our method with three well known techniques shows a 10% improvement in accuracy with an $O(n)$ complexity of the classifier.

Keywords anomaly detection, feature filtering, multi-resolution analysis, sparse distribution

Citation Cheng G Z, Chen H C, Cheng D N, et al. Uncovering network traffic anomalies based on their sparse distributions. Sci China Inf Sci, 2014, 57: 092204(11), doi: 10.1007/s11432-014-5087-7

1 Introduction

Uncovering anomalies in large Internet Service Providers (ISPs) and enterprise networks is challenging. On the one hand, there is a wide variety of such anomalies. Anomalies can come from activities with malicious intent (e.g., DDoS, port scanning), from failures of network components (e.g., link failures, congestion problems), or even from legitimate events such as flash crowds. On the other hand, it is a great number of traffic features representing the various of anomalies that prevent perfect techniques performing in an effective way. Thus traffic cannot be characterized precisely by lower dimensional features, whereas higher dimensional ones costly. Determining a suitable tradeoff by dynamically filtering the "proper" feature subset is an open problem.

A number of techniques have been proposed to detect these anomalies by analyzing network traffic [1]. All of these try to reveal anomalies by detecting deviations from some underlying model of normal traffic.

*Corresponding author (email: guozhencheng@hotmail.com)

© Science China Press and Springer-Verlag Berlin Heidelberg 2014

For a long time, researchers have uncovered anomalies in the network based on traffic volume, for example, number of packets or bytes [1–3]. Nychis et al. [4] reviewed the distribution of traffic volume and degree as a result of entropy. This can hit anomalies caused by much smaller traffic flows, but not the distribution of IP addresses and ports. Lakhina discovered that the distributions of packet features (IP addresses and ports) observed in flow traces reveal both the presence and the structure of a wide range of anomalies [2,3]. Then, using entropy as a summarization tool, they designed a method that is highly sensitive to a wide range of anomalies. As indicated by Ringberg [5], principal component analysis (PCA) is sensitive to large traffic anomalies. Silveira et al. [6–8] introduced Unsupervised Root Cause Analysis (URCA), which isolates anomalous traffic and classifies alarms with minimal manual assistance and high accuracy. The most disastrous limitation is that it is difficult for URCA to converge if initialized badly. However, none of the above studies attach much importance to the correlation among features, or even whether it is necessary for all of these to be detected together.

Recently, special emphasis has been placed on assembling multiple atomic detectors in the field of network traffic anomaly detection. Nyalkalkar et al. [9] compared a promising approach from each of two broad categories (approaches based on spatial correlation and temporal analysis, respectively), namely, entropy-based PCA and hierarchical heavy hitter (HHH) based wavelets. Gao et al. [10] proposed inferring a discriminative model by reaching consensus among multiple atomic anomaly detectors in an unsupervised manner, when there are very few or even no known anomalous events for training. Although the combined detector achieved a 10 to 20% improvement over the base detectors, the researchers were disappointed with the cost.

Traffic features, such as packet and flow counts, form a time series with the following two features: (i) self-similarity [11], i.e., auto-covariance function $R(k)$ at different time scales are fundamentally the same; (ii) long range dependence [12,13], which means that the auto-covariance function decays at a much slower rate than general exponential decay.

When volume anomalies occur in the network, the structure of traffic violates this phenomenon. Thus, multi-resolution analysis is used to uncover these anomalies under these circumstances. it is Typically impossible for large anomalies to present themselves on the network links. On the contrary, network anomalies are normally distributed in a sparse way. Therefore, it comes as no surprise that the integral structure of traffic is usually unchanged even with the existence potential anomalies. For example, when large instantaneous anomalies occur in the network, the anomalies are proportionally lower than normal traffic making them totally disappear as they are drowned by normal traffic noise. From a multi-resolution perspective, we can decompose a traffic series into multi-layers, with each layer is mapped onto one frequency scope.

According to the different anomalies inhabiting different frequency scopes, we can separate anomaly-free layers from anomalous ones. Moreover, each of the anomalies is essentially characterized by its lower-dimensional features, irrespective of how many higher-dimensional features there are. In this paper, we introduce a novel approach to filter out the "proper" features from a set with high dimensionality. This approach is based on a mathematical model of a type of sparsity that we call Multi-Resolution Low Rank (MRLR). Then, Based on the MRLR, we develop a dimensionality reduction technique, Recursive Reducing Features (RRF) that reduces redundant features, i.e., most of the features that are not contaminated by anomalies. Then, based on the residual features we proposed a simple yet effective classifier, the Focused Classification Algorithm (FCA). We validate the MRLR using manually analyzed real anomalies as well as synthetic anomaly injection. The validation shows that RRF can accurately filter anomalous flow features, and achieves a 10% improvement in accuracy. FCA whose complexity is $O(n)$, works well in real-time.

The rest of the paper is organized as follows. In Section 2 we provide some background analytical information to justify the techniques adopted. In Section 3 details of the proposed model, i.e., MRLR are given. Section 4 presents the RRF and FCA based on the MRLR, while in Section 5 we describe the traffic traces and anomalies used in the experimental tests, and discuss the results thereof. Finally, in Section 6 we draw some conclusions and discuss future works.

Table 1 Features and the types of anomalies detected by them (H $(\cdot)$ denotes entropy)

Anomaly	Description	Flow features for detecting (F')
Alpha	Unusually high rate of point to point byte transfer	# of packets, # of bytes
DDoS	Distributed denial of service attack against a single victim	H(dest_ip), flow counts, # of src_ip, # of packages
Network scan	Scanning the network for a target port	H(src_ip), H(dest_port), flow counts, # of packages
Port scan	Scanning a host for a vulnerable port	H(src_ip), H(src_port), # of src_port, # of packets
Flash crowd	Unusually large number of accesses for some resource/service	Flow counts, H(dest_ip), H(dest_port)

2 An MRLR model

2.1 Model definition

Let $F = f_1(t), f_2(t), \ldots, f_n(t)$ be the observed feature sets (e.g., aggregated traffic time series for several same protocol words). $F_i(t) \in F$ represents the time series of feature-i. Considering the multi-resolution of traffic, we retain Empirical Mode Decomposition (EMD) to give prominence to anomalies before applying MRLR.

In brief, the EMD method deals with non-stationary and non-linear signal on purpose [14]. In contrast to almost all of the previous methods, this new method is intuitive, direct, and adaptive, and is based on the simple assumption that any signal consists of different simple intrinsic modes of oscillations. Each intrinsic mode, linear or nonlinear, represents a simple oscillation, which is represented by an intrinsic mode function defined as follows: (i) in the whole dataset, the number of extrema and the number of zero-crossings must either be equal or differ at most by one, and (ii) at any point, the average value of the envelope defined by the local maxima and the envelope defined by the local minima is zero.

Typically, for $f_i(t)$ $(i = 1, 2, \ldots, n)$ as input, we get $f_{ij}(t)$ $(i = 1, 2, \ldots, n;\ j = 1, 2, \ldots, p)$ instead after a p-layer EMD. Formally, variable t is discretized to obtain matrix $F(i, j, k)$ $(i = 1, 2, \ldots, n;\ j = 1, 2, \ldots, p;\ k = 1, 2, \ldots, M)$, i.e., traffic feature matrix (TFM), which expresses the k's observed value on scale- j about feature- i, if the length of the observed window is M. This is a multi-way data; an effective way of analyzing multi-way data is to recast it into a simpler, one-way representation. We can "unfold" the multi-way matrix into a single, large matrix by considering each scale as an independent variable. This results in a new, merged matrix of size $N \times M$ $(N = n \times p)$, which contains the ensemble of $N = n \times p$ scales.

The TFM is a very large matrix, i.e., the variables that are used to detect anomalies is high dimensionalities. However, there are two ground truths. First, network anomalies are typically distributed in a sparse way [2,3,15]. Anomalies tend to assemble locally except for network-wide uncontrolled attacks or failures that fortunately occur infrequently. Accordingly, these affect sparse links or several local areas. Second, each of the anomalies is essentially characterized by its lower-dimensional features, as presented in Table 1. From the view of anomaly detection, we have a set of linear constraints on the TFM, i.e.,

$$A(F) = Q, \tag{1}$$

where $A(\cdot)$ is a linear operator, and Q is a $N \times M$ matrix given by

$$Q = \begin{cases} 0, & \text{if } F(i, j) \text{ has no anomaly,} \\ 1, & \text{otherwise.} \end{cases} \tag{2}$$

The operator expresses the information available from anomalies. Note that the sparse presence of anomalies is implicit in (1) by means of the sparse Q; for instance, operator A can filter suspicious variables or rows of the matrix by writing (1) as $F' = Q \cdot *F$, where $\cdot *$ denotes the element-wise product, i.e., $A = C \cdot *B$ means $A(i, j) = B(i, j)\, C(i, j)$.

A sparse matrix is simply a matrix that has only a few non-zero elements. Often anomalies of interest occur only a few elements of the matrix to increase, where the rest remain small. The notion of low rank approximation is associated with these larger elements, because most of its information is carried in the larger elements. Generally, a low rank is similar to sparsity, because the spectrum formed by the singular value of a low rank matrix is sparse [16]. The central premise of normal anomaly detectors is that the presence of anomalies should alter the structure of normal traffic, which is not in conflict with the idea of MRLR, but is instead complementary since the premise is not always valid.

2.2 Consequences of the MRLR model

Given an $N \times M$ matrix A and a positive integer k, we wish to find an $N \times M$ matrix A_k of rank at most k, so as to minimize the Frobenius norm of the matrix difference $X = A - A_k$, defined as

$$\|X\|_F = \sqrt{\sum_{i=1}^{N} \sum_{j=1}^{M} A_{ij}^2}. \tag{3}$$

Thus, the Frobenius norm of X measures the discrepancy between A_k and A, Our goal is to find a matrix A_k that minimizes this discrepancy, while constraining A_k to have rank at most k. If r is the rank of A, clearly $A_r = A$ and the Frobenius norm of the discrepancy is zero in this case. When k is much smaller than r, we refer to A_k as a low-rank approximation.

Singular value decomposition (SVD) can be used to solve the low-rank matrix approximation problem. Let r be the rank of the $N \times M$ matrix A. Then, there is an SVD of A of the form

$$A = U\Sigma V^{\mathrm{T}}, \tag{4}$$

where V^{T} is the transpose of V, and U is an $M \times M$ unitary matrix (i.e., $U^{\mathrm{T}}U = UU^{\mathrm{T}} = \mathrm{I}$), and V is a $N \times M$ unitary matrix (i.e., $V^{\mathrm{T}}V = VV^{\mathrm{T}} = \mathrm{I}$). The eigenvalues $\lambda_1, \ldots, \lambda_r$ of AA^{T} are the same as the eigenvalues of $A^{\mathrm{T}}A$. For $1 \leqslant i \leqslant r$, let $\sigma_i = \sqrt{\lambda_i}$, with $\lambda_i \geqslant \lambda_{i+1}$. Then the $N \times M$ matrix Σ is composed by setting $\Sigma_{ii} = \sigma_i$ for $1 \leqslant i \leqslant r$, and zero otherwise.

To understand the use of the SVDs in matrix approximations, consider A_k which is the best rank-k approximation of A, incurring an error (measured by the Frobenius norm of $A - A_k$) equal to σ_{k+1}. Thus, the larger k is, the smaller is this error. To derive further insight into why the process of truncating the smallest $r - k$ singular values in Σ helps generate a rank-k approximation of with small error, we examine the form of A_k:

$$A_k = U\Sigma_k V^{\mathrm{T}} = \sum_{i=1}^{k} \sigma_i \vec{u}_i \vec{v}_i^{\mathrm{T}}, \tag{5}$$

where $\vec{u}_i$ and $\vec{v}_i$ are the ith columns of U and V, respectively. Thus, since $\vec{u}_i\vec{v}_i^{\mathrm{T}}$ is a rank-1 matrix, we have just expressed A_k as the sum of k rank-1 matrices each weighted by a singular value. As i increases, the contribution of the rank-1 matrix $\vec{u}_i\vec{v}_i^{\mathrm{T}}$ is weighted by a sequence of shrinking singular values σ_i. Based on (3) and (5), truncation of the SVD provides a natural solution to

$$\min \|A - A_k\|_F, \quad \text{s.t.} \quad \mathrm{rank}(A_k) \leqslant k. \tag{6}$$

In terms of the fact that matrix A is exactly a sparse matrix, $k \ll r$, we obtain the low rank approximation of A.

Table 1 illustrates the prevalent anomalies and different kinds of information that are present in each type of traffic feature. Each anomaly can be detected by several features because it changes their relevant behaviors. For example, Alpha tends to be detected by number of bytes or number of packets, while DDoS by H(dest_IP), number of flows, among others. To uncover as many anomalies as possible, the feature set should have high dimensionality (e.g., the total number of feature dimensions in KDD CUP1999 datasets is 41). Suppose v_i represents the unitary absolute deviation of f_i when anomalies occur. Then, the proportion of each deviation v_i is defined as

$$P(f_i) = \frac{v_i}{\sum_{i=1}^{n} v_i}. \tag{7}$$

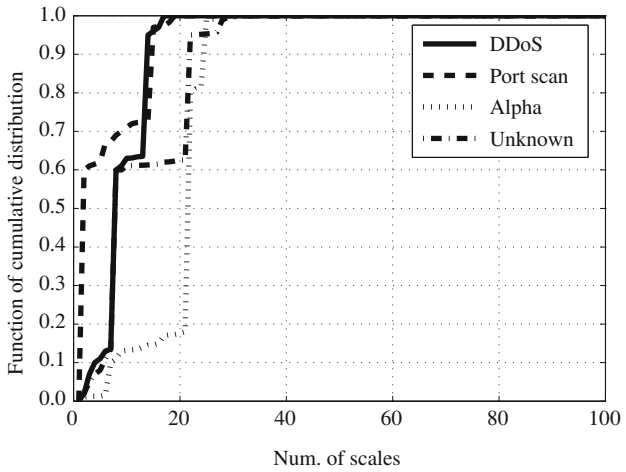

Figure 1 Cumulative distribution of deviation.

Figure 1 depicts the cumulative distribution figure for different $P(f_i)$, which shows deviated proportion as a function of the sequence number of features. The common ground of all curves is that there are several impulses, which dominate the deviations at corresponding features. Nevertheless, this phenomenon validates the sparsity of the anomaly.

3 MRLR-based anomaly detection

This section presents an algorithm that dynamically selects "optimal" features for detection. We also provide empirical evidence that the algorithm in our detector is effective for real traffic.

3.1 Recursive Reducing Features (RRF)

We designed a heuristic greedy algorithm to identify the real offenders in a set of features. Algorithm 1 shows the pseudo code for RRF which takes three parameters: F, the deviation function $V(\cdot)$ and threshold. It outputs A_f whose elements are symptomatic of the underlying anomalies. First, initialize $N_f = F$, $A_f = \emptyset$, where N_f is a candidate of the features. Then, for each element in N_f, if the difference between $V(N_f)$ and $V(N_f|f_i)$ exceeds the threshold, add f_i to A_i, and at the same time delete it from N_f. If $A_f = \emptyset$, RRF deems the threshold to be on the high side, reduce it, and run RRF recursively. If the ratio between count(F) and count(A_f) is less than 10, which could mean that A_f is not a sparse matrix, we consider that this phenomenon violates the sparsity supported by MRLR, and runs recursively until sparsity is satisfied.

3.2 Focused classification algorithm

Having delved into traffic anomalies, we found the ground truth that the fluctuation of each f_i is most often contributed by several classes of anomalies in one anomalous subset C_i. In another, each anomaly emphasizes itself by causing several specific traffic features to fluctuate obviously. Thus, different traffic features can always be explained by several anomalous subsets. When reviewing, one-by-one, the multi-traffic features that simultaneously represent the anomaly we may find some anomalies that appear with high frequency. Based on this, we propose a simple approach for classifying the anomalies, namely, the FCA.

Algorithm 2 gives the pseudo code for the FCA, which takes that has a couple of parameters: the subset of features $\{f_1, f_2, \ldots, f_l\}$ selected by RRF and C^{last}. The concrete steps are listed below:

Step 1: Map F_{sub} to anomaly sets $C.$ and then, compute the intersection C_r among C.

Step 2: If there is at least one element in C_r, C_r is defined by the intersection between C_r and C^{last} that is the intersection C_r obtained in the last iteration, else go to step 3.

Algorithm 1 Pseudo code for the RRF.

Input:

F: the full set of flow features.

N_{last}: N that acquire by last recursion, $\emptyset$ initially. Every recursion makes $A_f + N_{\text{last}}$ to be the full set of flow features.

threshold: deviation threshold.

$V(\bullet)$: the function computing the deviations of features caused by anomaly.

Output:

A_f: the subset of flow features contained anomaly.

$N_f \leftarrow F : \{f_i \,|f_i \in F\}$;
$A_f \leftarrow \emptyset$;
for all $f_i \in N_f$ **do**
  if $\left|V\left(N_f\right) - V\left(N_f \backslash f_i\right)\right| >$ threshold **then**
    $A_f = \text{add}\left(A_f, f_i\right)$;
    $N_f = \text{eliminate}(N_f, f_i)$;
  else
    continue;
  end if
end for
$N_{\text{last}} = \text{add}(N_{\text{last}}, N_f)$;
if $A_f = \emptyset$ **then**
  threshold = reduce(threshold);
  $A_f = \text{RRF}\left(F, N_{\text{last}}, \text{threshold}, V\left(\bullet\right)\right)$;
else
  if element $\left(A_f \cup N_{\text{last}}\right)$/element $\left(A_f\right) < 10$ **then**
    threshold = augment(threshold);
    $F \leftarrow A_f$;
    $A_f = \text{RRF}\left(F, N_{\text{last}}, \text{threshold}, V\left(\bullet\right)\right)$;
  end if
end if
return A_f;

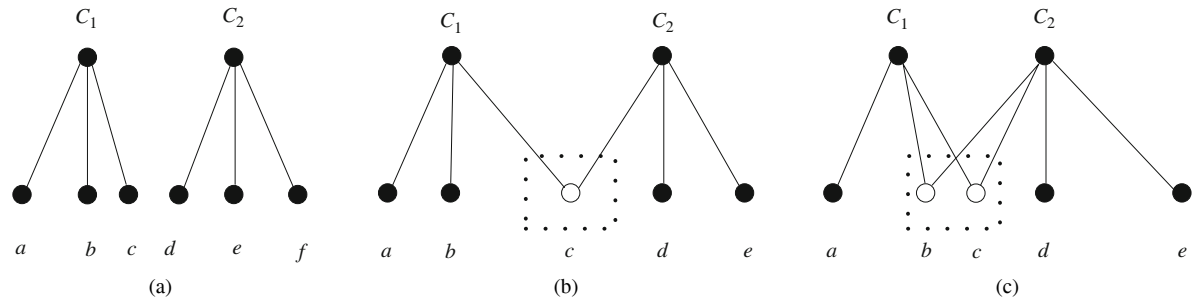

Figure 2 Possible results of FCA. (a) Normal; (b) correct; (c) ambiguous.

Step 3: If there are no elements in C_r, then the element presented most frequently in C is returned.

Figure 2 depicts the possible outputs of FCA. Suppose a–f are known classes of the anomaly. C_1 and C_2 are separately aroused by f_1 and f_2. Figure 2(a) represents $C_r = C_1 \cap C_2 = \emptyset$, which denotes that no anomaly occur. Figure 2(b) represents $C_r = C_1 \cap C_2 = \{c\}$, which suggests that the correct anomaly has been classified. Figure 2(c) represents $C_r = C_1 \cap C_2 = \{b, c\}$, which means that we cannot identify the correct one in b and c.

4 Evaluation

4.1 Datasets

The previous section presented a precise view on MRLR and its RRF. Next, we describe the experiments used to validate it. Evaluating anomaly detectors is notoriously difficult for non-accurate data. In the absence of a ground truth, we use synthetic anomalies and real traffic traces.

Algorithm 2 Focused Classification Algorithm.

Input:

F_{sub}: the subset of features $\{f_1, f_2, \ldots, f_l\}$ selected by RRF.
C^{last}: C_r which latest Output of the FCA, and C initially.

Output:

C_r: the real class that the anomaly affiliated.

Procedure FCA($F_{\text{sub}}, C^{\text{last}}$)
$F_{\text{sub}} \leftarrow C : \{C_1, C_2, \cdots C_l\}$;
$C_r \leftarrow C_1 \cap C_2 \cdots \cap C_l$;
$C_r = \text{intersection}\left(C_r, C^{\text{last}}\right)$;
if $|C_r| = 0$ **then**
  $C_r = \text{max\_support}\,(C)$;
  $\text{intersection}\left(C_r, C^{\text{last}}\right)$;
end if
return C_r;
Procedure intersection(C_r, C^{last})
if elements(C_r) > 1&& $\left|C^{\text{last}}\right| > 0$ **then**
  $C_r = \text{intersection}\left(C_r, C^{\text{last}}\right)$;
end if

Table 2 Datasets used in the experiments

Symbol	Network	Collection time	Time scales (min)	Anomalous density (%)
A	Enterprise network	2010-08	1	< 1
B	CERNET2[a)]	2010-04	1	< 19
C	Abilene[b)]	2007-05	5	< 22

a) CERNET2. http://www.edu.cn/internet_2_1339/index.shtml.
b) The Abilene Observatory Data Collection. http://abilene.internet2.edu/observatory/data-collections.htm.

Table 3 Original variables

Symbol	Description	Symbol	Description
f_1	# of flows identified as a 5-tuple	f_7	# of dest ports emerged
f_2	# of packets	f_8	Entropy of src ip
f_3	# of bytes	f_9	Entropy of dest ip
f_4	# of src IPs emerged	f_{10}	Entropy of src port
f_5	# of dest IPs emerged	f_{11}	Entropy of dest port
f_6	# of src ports emerged		

The datasets used are show in Table 2. Datasets -A was collected from traffic in a single large-scale enterprise, since the network of this enterprise was a virtual private network, we deemed these traffic traces to be approximately anomaly-free. Having inspected synthetic anomalies in it, we signed it. Datasets -B and -C come from the CERNET2 and Abilence networks, respectively, both of which are contaminated by anomalies. We manually marked the anomalies in B and C, and then exposed anomalies utilizing variables as shown in Table 3. Note that the original form of C was a traffic matrix, so we re-aggregated it to variables as Table 3 shows.

For each variable $f_i\,(t)$, after m-layer EMD, we obtained $F\,(i,j)$ $(i = 1, 2, \ldots, N;\ j = 1, 2, \ldots, M)$, where the columns represent variables or features, and the rows represent observations. In general, there are n original variables, with the number of real columns in $F\,(i,j)$ is $N = nm$. For instance, given the 11 features in Table 3, and assuming $m = 10$, the total number of variables is 110 which is a high-dimensional set.

In the next experiment, we used typical anomaly detection algorithms, SVM, Bayes and PCA, for comparison. The RRF algorithm based on MRLR, was deployed between traffic records, with all detection algorithms depicted in Figure 3. Low-dimensional records were first extracted from the original traffic records by RRF. General algorithms could then utilize this redundancy-free data for anomaly detection.

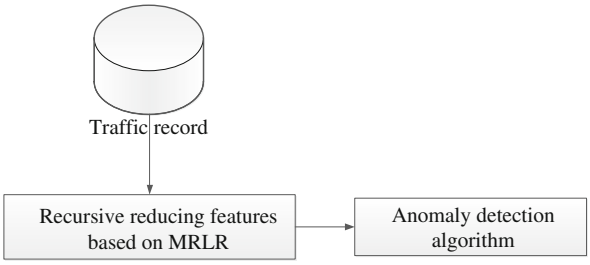

Figure 3 Deployment of anomaly detection algorithms with RRF.

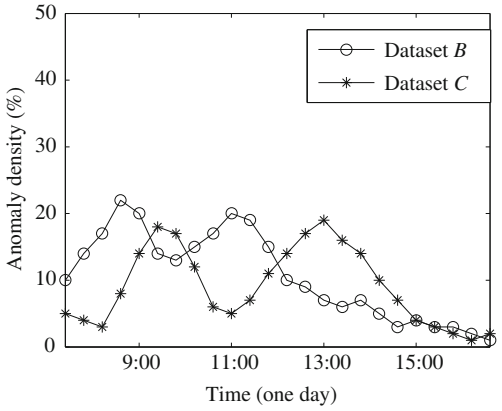

Figure 4 Anomalous density of datasets -B and -C over 24 h period.

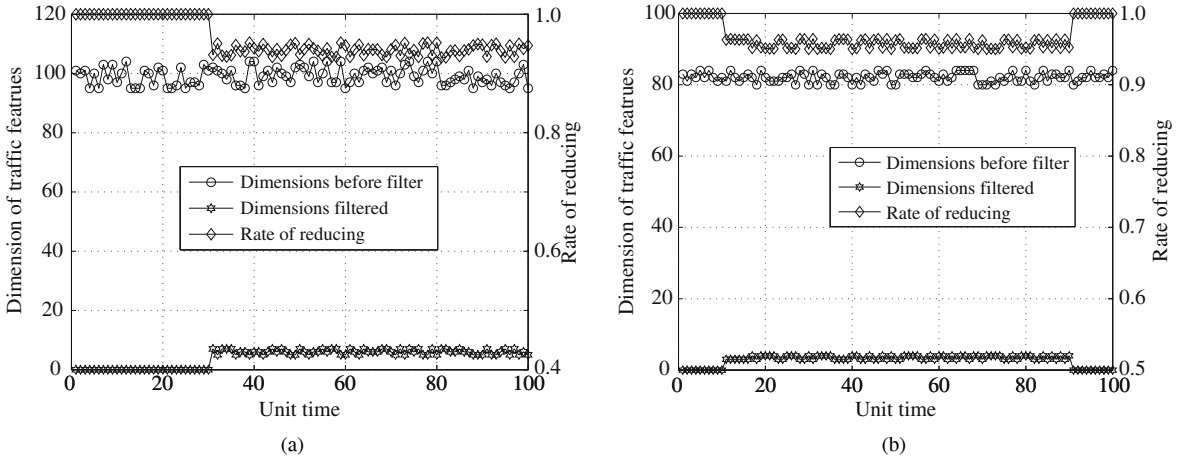

Figure 5 Variable reduction of RRF. (a) DDoS; (b) Network scan.

4.2 Experimental process

We firstly give the following notions before experiment:

Definition 1 (Reduction Ratio)**.** the ratio $(n-l)/n$, where n is the number of original variables and l is the number of variables selected by RRF.

Definition 2 (Anomalous Density)**.** the ratio k/m, where m is the total number of links in the network, k is the number of suspicious links.

Furthermore, we provide empirical evidence about the assumptions with respect to the sparse distribution of anomalies in our detector. Therefore we review the anomalous density of datasets -B and -C as shown in Figure 4. Both have a lower anomalous density than 30% which means that the distribution is also sparse in real traffic traces.

To gain a clearer understanding of the nature of the suspicious variables detected using RRF. We manually inspected each of the 350 anomalies. And then we extract heavy-hitter variables. Figure 5 shows the real original variables and the heavy-hitter ones exposed by RRF. Find that the reduction ratio is lower than 10% which is a typically perfect result.

Although Figure 5 represents the pure effect of RRF on reducing variables, we do not know whether the selected variables are actual heavy-hitters. Using the above injection, we are able to plot the receiver operating characteristic (ROC) curves, which show detection rates as a function of false positive rates. The ROC plot in Figure 6 shows the relative performance of RRF for two common injected anomalies. The curves show that the detection rate is greater than 90% with a false positive rate of 10%.

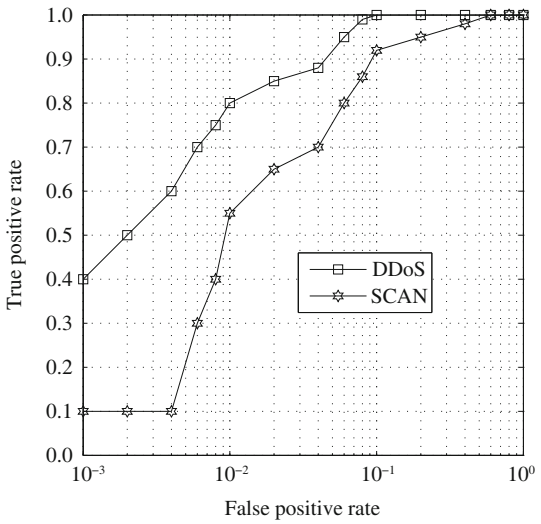

Figure 6 RRF ROC curves for two common anomalies.

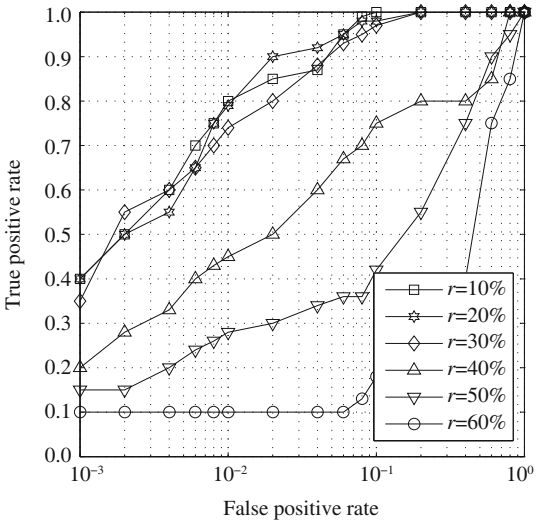

Figure 7 RRF ROC couves for different r.

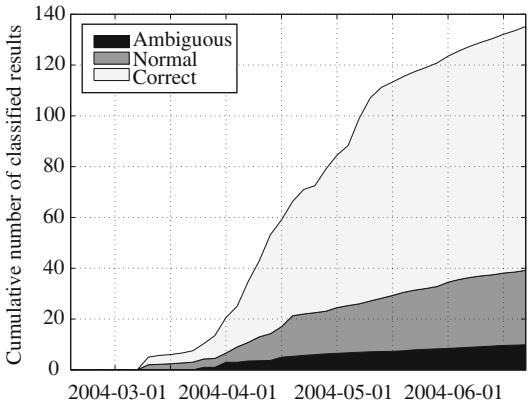

Figure 8 Performance of FCA.

The premise of MRLR is that the observed network has a low anomalous density r which is the daily truth. For maturity, we also observed the ability of MRLR with different r. Figure 7 shows six different behaviors in the relative performance of RRF when increasing the anomalous density in the network from 10% to 60% at intervals of 10%. The performance of RRF decreases with an increase in r. when r is less than 30%, RRF functions in a finer state, yet evidently deteriorates thereafter.

Figure 8 shows the proportion of three outcomes of FCA. There is little evidence of ambiguous phenomena in the results, the dataset used has a low anomalous density and almost no intercross between anomalies.

In Figure 9, we plot the ROC curves for the MRLR-based anomaly classifiers, SVM, Bayes, as well as PCA+Entropy. Whenever the SVM and Bayes classifiers employ the MRLR-based algorithm RRF to filter the meta-data, they both show greater improvement. Although the FCA classifier has slightly lower performance compared to SVM and Bayes, it is greater than that of PCA+Entropy, and has low complexity, about $O(n)$, where n denotes the dimensions of the meta-data.

Finally, Table 4 compares the time cost for FCA, SVM and Bayes, before and after applying MRLR as well as for PCA+Entropy. The execution environment for the emulator is an Intel(R) Pentium(R) Dual E2140 1.6 GHz processor and 1 GB memory. We can see that MRLR+FCA requires less time than SVM, Bayes and PCA+Entropy. In addition, the cost of SVM and Bayes decreases by approximately 30% and 25% respectively, with the application of MRLR.

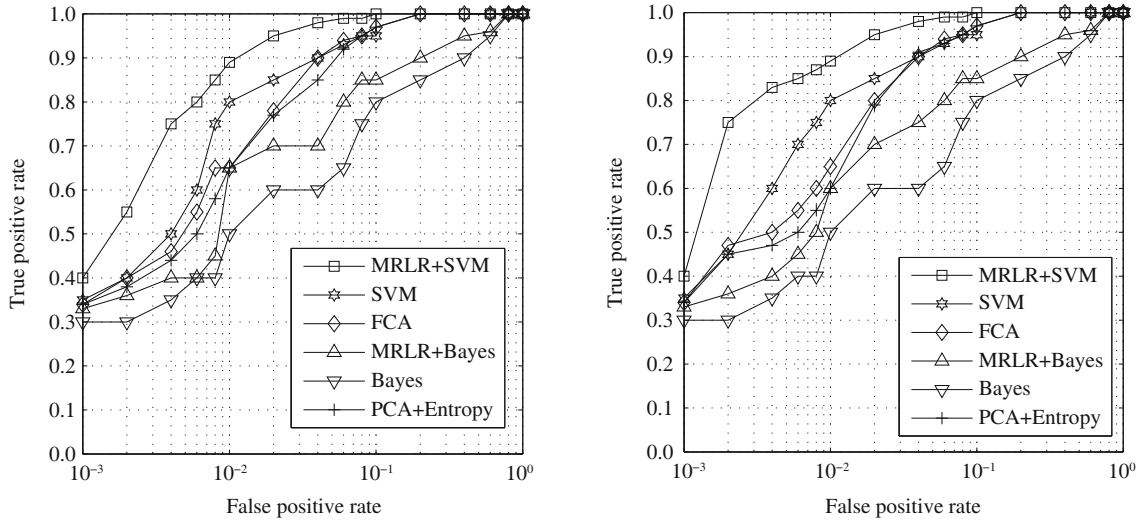

Figure 9 ROC curves comparing MRLR-based classifiers with SVM, and Bayes for different datasets. (a) Using dataset B; (b) using dataset C.

Table 4 Comparison of execution times for FCA, SVM, Bayes and PCA + Entropy

Algorithm	Dataset B	Dataset C
Bayes	95	93
SVM	102	98
MRLR+Bayes	69	72
MRLR+SVM	72	68
MRLR+FCA	29	43
PCA+entropy	75	97

5 Conclusion and future research

This paper proposed an inductive dimensionality theory (i.e., MRLR) based on the sparse distribution of anomalies. Under the supervision of MRLR, we designed an algorithm (RRF) that can dynamically reduce features. We validated MRLR using manually analyzed real traffic anomalies as well as synthetic injected anomalies. Our validation shows that MRLR can accurately filter anomalous flow features, and reduce the dimensions thereof to less than 10%. Our future work will focus on extending MRLR to classify various anomalies.

Acknowledgements

This work was supported by National Basic Research Program of China (973) (Grant No. 2012CB315901), National Natural Science Foundation of China (Grant Nos. 61372121, 61309020, 61309019), National High-tech R&D Program of China (863) (Grant Nos. 2011AA01A103, 2011AA010605, 2011AA01A101), and National Science and Technology Support Program (Grant Nos. 2012BAH02B01, 2012BAH02B03).

References

1 Barford P, Kline J, Plonka D. A signal analysis of network traffic anomalies. In: Proceedings of the 2nd ACM SIGCOMM Workshop on Internet measurment, Marseille, 2002. 71–82
2 Lakhina A, Crovella M, Diot C. Diagnosing network-wide traffic anomalies. In: Proceedings of the Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications, Portland, 2004. 219–230
3 Lakhina A, Crovella M, Diot C. Mining anomalies using traffic feature distributions. In: Proceedings of the Conference

on Applications, Technologies, Architectures, and Protocols for Computer Communications, Philadelphia, 2005. 217–228
4 Nychis G, Sekar V, Andersen D G, et al. An empirical evaluation of entropy-based anomaly detection. In: Proceedings of the 8th ACM SIGCOMM Conference on Internet Measurement, Vouliagmeni, 2008. 151–156
5 Ringberg H, Soule A, Rexford J, et al. Sensitivity of PCA for traffic anomaly detection. In: Proceedings of the ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems, San Diego, 2007. 109–120
6 Silveira F, Diot C. URCA: pulling out anomalies by their root causes. In: Proceedings of IEEE INFOCOM, San Diego, 2010. 1–9
7 Silveira F, Diot C, Taft N, et al. ASTUTE: detecting a different class of traffic anomalies. In: Proceedings of the ACM SIGCOMM Conference, New delhi, 2010. 267–278
8 Silveira F, Diot C, Taft N, et al. Detecting Correlated Anomalous Flows. Thomson, Technical Report CR-PRL-2009-02-0001, 2009
9 Nyalkalkar K, Sinhay S, Bailey M, et al. A comparative study of two network-based anomaly detection methods. In: Proceedings of IEEE INFOCOM, Shanghai, 2011. 176–180
10 Gao J, Fanj W, Turaga D, et al. Consensus extraction from heterogeneous detectors to improve performance over network traffic anomaly detection. In: Proceedings of IEEE INFOCOM, Shanghai, 2011. 181–185
11 Paxson V, Floyd S. Wide-area traffic: the failure of poisson modeling. IEEE/ACM Trans Netw, 1995, 1: 226–244
12 Leland W E, Taqqu M S, Willinger W, et al. On the self-similar nature of Ethernet traffic. IEEE/ACM Trans Netw, 1994, 2: 1–15
13 Klivansky S, Mukherjee A, Song C. On long-range dependence in NSFNET traffic. Technical Report, Georgia Institute of Technology. 1995
14 Huang N E, Shen Z, Long S R, et al. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. Proc Roy Soc London Ser A, 1998, A454: 903–995
15 Zhang Y, Roughan M, Willinger W, et al. Spatio-temporal compressive sensing and Internet traffic matrices. In: Proceedings of SIGCOMM, Barcelona, 2009. 267–279
16 Xu X D, Zhu S R, Sun Y M. Anomaly detection algorithm based on fractal characteristics of large-scale network traffic. J Commun China, 2009, 30: 43–53