Hindawi Publishing Corporation
Journal of Probability and Statistics
Volume 2010, Article ID 375942, 16 pages
doi:10.1155/2010/375942

Research Article

An Effective Method of Monitoring the Large-Scale Traffic Pattern Based on RMT and PCA

Jia Liu,[1] Peng Gao,[1] Jian Yuan,[2] and Xuetao Du[1]

[1] *Research Division, China Mobile Group Design Institute Co. Ltd, Beijing 100080, China*
[2] *Department of Electronic Engineering, Tsinghua University, Beijing 10084, China*

Correspondence should be addressed to Jia Liu, liujia1@cmdi.chinamobile.com

Received 2 September 2009; Revised 17 January 2010; Accepted 12 April 2010

Academic Editor: Chunsheng Ma

Mechanisms to extract the characteristics of network traffic play a significant role in traffic monitoring, offering helpful information for network management and control. In this paper, a method based on Random Matrix Theory (RMT) and Principal Components Analysis (PCA) is proposed for monitoring and analyzing large-scale traffic patterns in the Internet. Besides the analysis of the largest eigenvalue in RMT, useful information is also extracted from small eigenvalues by a method based on PCA. And then an appropriate approach is put forward to select some observation points on the base of the eigen analysis. Finally, some experiments about peer-to-peer traffic pattern recognition and backbone aggregate flow estimation are constructed. The simulation results show that using about 10% of nodes as observation points, our method can monitor and extract key information about Internet traffic patterns.

1. Introduction

Accurate and timely detection and recognition about entire network traffic patterns are of great significance for network operations. However, the present monitoring technologies make it impossible for a clear understanding of the network-wide traffic characteristics, especially in a large-scale network. To address this issue, taking a network-wide view of traffic has been proposed as one of the most important principles in the future network [1]. Recently, there are several active areas in traffic measurement research, including high-rate flow detection [2], traffic engineering [3], anomalies detection [4], and network management [1] and so forth.

Traffic measurement over an entire network is faced with many challenges because of the rapid growth of network size. High-speed links, concurrent flows, and mixed services make it too expensive for routers to trace every traffic flow. Even if this can be achieved, it is impossible to handle such a mass of information.

A sampled traffic flow could be denoted as time series, and the correlation between traffic flows could reveal important information about network traffic pattern. This property is well utilized in Random Matrix Theory (RMT). In the method, a random correlation matrix, which is constructed from mutually uncorrelated time series, is compared against a correlation matrix of measured data, and the results of the related research [5, 6] reveal that almost 98% of the eigenvalues of cross-correlation matrix of measured data agree with RMT predictions, suggesting a considerable degree of randomness in the measured cross correlations, while the deviations between the two matrices convey information about the characteristics of real world. Results from RMT have been used as an analysis tool in some selected areas [6, 7].

Recently the RMT-based method has been used in a network study. In [8], the Renater study firstly uses RMT to analyze cross-correlation among network flows. They find that the largest eigenvalue of the flow cross correlation matrix is approximately 100 times larger than predicted for uncorrelated time series, and the eigenvector component distribution of the largest eigenvalue deviates significantly from the Gaussion distribution predicted by RMT. Further, the Renater study reveals that all components of the eigenvector corresponding to the largest eigenvalue imply their collective contribution to the strong correlation in congestion over the whole network. Since all network flows contribute to the eigenvector, the eigenvector can be viewed as an indicator of spatial-temporal correlation in network congestion.

Differing from the Renater study which is performed in a small-scale network with only 30 routers, we find that RMT is more applicable to large-scale network traffic monitoring and analysis. In [9], an RMT-based approach is proposed to study the pattern shift in Internet traffic caused by distributed denial-of-service attacks with only a few observation points. In our previous work [10], we measured large-scale client-server and peer-to-peer traffic patterns with a few subnets nodes which have the large degree. However, we select the observation points only through several repeated experiments without strong evidence. In this paper, we extend the application scenarios to monitor the traffic patterns of the links and subnets, and analyze the selection of the monitors through Principal Components Analysis (PCA) which is a mathematical tool in common use to analyze multivariate data and dimensionality reduction.

As stated above, we propose a method in the paper, based on the combination of RMT and PCA, to capture the main traffic patterns of large-scale networks with only a few observation points. With complete explanation in the paper, we put forward an effective approach to monitor the large-scale traffic patterns exactly with only about 10% of subnets routers as observation points which is selected by analysis. The rest of this paper is structured as four sections. In Section 2, we give a large-scale network model as a prototype of Internet, which is the basis of our experiment in the following parts. In Section 3, the hidden information is extracted from the covariance matrix of the flow data by the RMT analysis. In Section 4, the meanings of the largest eigenvalues and the small eigenvalues are explained by PCA theory, and guiding principles are proposed and analyzed to select observation points. In Section 5, some experiments are constructed and the simulation results subjected to detailed analysis. Then, we conclude our paper in Section 6.

2. The Model of Networks

In order to verify our ideas, many experiments are constructed to monitor the network behavior. Here, we use a four-tier model as illustrated in Figure 1, including 11 backbone routers

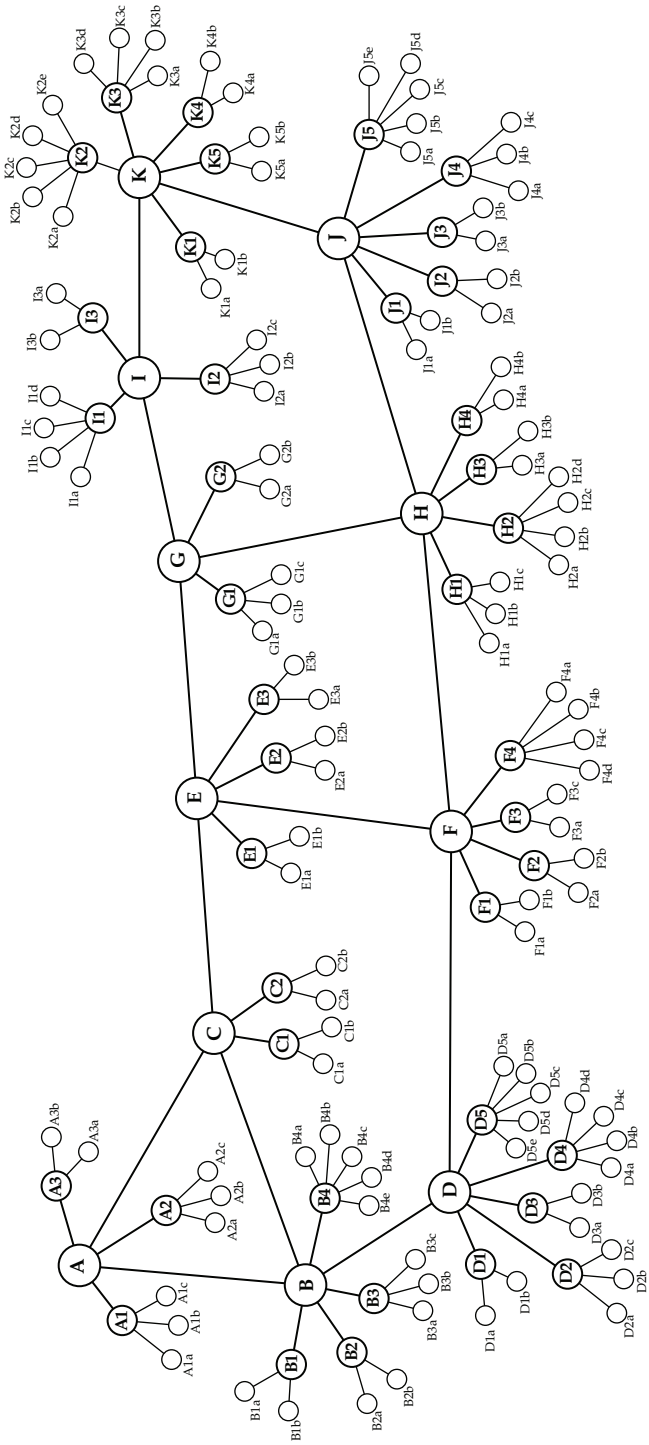

Figure 1: The simulation model with 11 backbone routers, 40 subnet routers, 110 leaf routers, and 22000 hosts.

(A, B, . . . , K), 40 subnet routers (A1, A2, . . . , K4, K5), 110 leaf routers (A1a, A1b, . . . , K5c, K5d), and 22000 hosts.

TCP protocol is applied in our model. As known, modern TCP implementations contain four intertwined algorithms: slow start, congestion avoidance, fast retransmit, and fast recovery. In order to shorten the simulation time, the model works adopting likely Reno TCP, except reducing the congestion window to half of the current window size after receiving one, instead of three, duplicate ACKs.

2.1. Background Traffic

The background traffic is constructed by the traffic between normal hosts.

Our model includes 22,000 sources, each of which represents a client. Each source generates traffic as an ON/OFF process which can provide a convenient model of user behavior. At the beginning of each ON period, a destination receiver is chosen randomly from any leaf router.

To store and forward packets, all routers maintain a queue of limited length, where arriving packets are stored until they can be processed: first in, first out. While small queue lengths lead to many losses during TCP slow start and large queues produce excessive delays, to achieve a reasonable balance, TCP simulations [11, 12] often set lengths of router queue in a range of 10 to 200 packets. We assume that setting maximum queue length (160 packets, in our simulation) within this range would not influence our qualitative findings.

Empirical measurements on the Internet observe a heavy-tailed distribution of file sizes [13]. Here, the Pareto distribution is therefore used to model heavy-tailed characteristics. The Pareto distribution function has the following form:

$$P[X \leq x] = 1 - \left(\frac{k}{x}\right)^{\alpha} \quad (k \leq x), \tag{2.1}$$

where $0 \leq \alpha \leq 2$ is the shape parameter. For our experiment, we select the same shape parameter, $\alpha = 1.5$, for both ON and OFF process; however, different means are choosen. Here, $\lambda_{\text{on}} = 50$ (packets) is selected to represent the preference for small files, as is typically the case with Web page downloads. Empirical observations of OFF periods change dramatically between observations made at night or in the day. Let $\lambda_{\text{off}} = 5000$ (milliseconds) represent the average thinking time before a user requests another file. Aiming to the detection of traffic pattern, we need to simulate background traffic, that is, neither too sparse nor too congested. Because when the network is too lightly loaded, the traffic pattern cannot be observed for the weak correlations among flows. On the other hand, when the network hardly overloaded, the likely-congested phenomenon we expect to observe will be submerged by congestion everywhere.

2.2. P2P Traffic

In this case, a dynamic P2P overlay network is considered as follows. Let $t0 < t1 < t2$, in which the three parameters represent different time. At the beginning, a group of peers distributed under A1, D2, J5, K2, and K3 communicate with each other at the time $t0$. Then the peers of D2 leave the sharing-file system at $t1$, while the ones under H2 join in at $t2$. The experiment

Table 1: Table of delays.

AB	AC	BC	BD	CE	DF	EF	EG	FH	GH	GI	HJ	IK	JK
10	12	12	4	8	20	10	7	9	7	4	7	5	3
BA	CA	CB	DB	EC	FD	FE	GE	HF	HG	IG	JH	KI	KJ
10	12	12	4	8	20	10	7	9	7	4	7	5	3

Table 2: Table of Routes.

AB	AC	ABD	ACE	ACEF	ACEG	ACEGH	ACEGI	ACEGHJ	ACEGIK
	BC	BD	BCE	BDF	BCEG	BDFH	BCEGI	BDFHJ	BCEGIK
		CBD	CE	CEF	CEG	CEGH	CEGI	CEGHJ	CEGIK
			DBCE	DF	DBCEG	DFH	DBCEGI	DFHJ	DFHJK
				EF	EG	EGH	EGI	EGHJ	EGIK
					FHG	FH	FHGI	FHJ	FHJK
						GH	GI	GHJ	GIK
							HGI	HJ	HJK
								IKJ	IK
									JK

is designed to simulate a process of P2P sharing-file systems with dynamic peers. The exact time of each joining and leaving peers will be described in Section 5.

2.3. Routes

The parameters of routers are configured so that leaf routers forward at 5000 packets per second (pps), subnet routers forward at 20,000 pps, and backbone routers forward at 160,000 pps, similar to the real Internet routing pace.

The shortest path is selected for each packet, which means the routing is static. The delay between the leaf routers and the corresponding subnet routers is ignored. And the delay between subnet routers is shown in Table 1. As a result, the packets on a connection will take the same route, and no reordering occurs. The path between each leaf routers is given in Table 2.

3. Random Matrix Theory

3.1. Related Works

In the theory of random matrix one is concerned with the following question. Consider a large matrix whose elements are random variables with given probability laws. Then what can one say about the probabilities of a few of its eigenvalues or a few of its eigenvectors? This question is originally of pertinence for the understanding of the statistical behavior of slow neutron resonances [14] in nuclear physics where it was proposed in 1950s and intensively studied by the physicists. Later Random matrix theory (RMT) was developed by Wigner et al. [15, 16]. Then it gained importance in other areas [17].

Data of Internet traffic is time correlated likes financial data [6]. A number of empirical studies have convincingly shown that the temporal dynamics of Internet traffic exhibits long-range dependence [18], which implies existence of nontrivial correlation structure at large

timescales. It can be explained that the TCP congestion-control algorithm exhibits a self-organizing property: when a large number of connections share the Internet, underlying interactions among the connections avoid router congestion simultaneously over varying spatial extent. The flow rate in a network is just like different stocks in financial market while the underlying interactions among the connections seems to be the varying levels of volatility of different stocks. As the eigenvalue deviation from RMT has a significant interpretation, there is no difference for Internet traffic. In 2002, a study of correlations among data flows in Renater [8], based on RMT method, has detected that the largest eigenvalue is approximately 100 times larger than predicted for uncorrelated time series, and the eigenvector component distribution of the largest eigenvalue deviates significantly from the Gaussian distribution predicted by RMT. Furthermore, the Renater study reveals that all components of the eigenvector corresponding to the largest eigenvalue are positive, which implies their collective contribution to the strong correlation in congestion over the whole network. Since all network flows contribute to the eigenvector, the eigenvector can be viewed as an indicator of spatial-temporal correlation in network congestion. This review reveals that congestion emerges from underlying interactions among flows crossing a network in various directions. According to this theory, we [9] successfully monitored stealthy DDoS attacks.

3.2. The Analysis by RMT

Our approach is based on the application of the deviations from RMT. Statistical properties of random matrices such as $\mathbf{X_{K\times M}}$ are specified in [19]. The results promoted that, in the limit $K \to \infty$, $M \to \infty$, $Q \equiv K/M$ (> 1) is fixed. It was shown analytically that the probability density function $P_{rm}(\lambda)$ of eigencalues λ of the random correlation matrix of $\mathbf{X_{K\times M}}$ is given by

$$P_{rm}(\lambda) = \frac{\mathbf{Q}}{2\pi}\frac{\sqrt{(\lambda_+ - \lambda)(\lambda_- - \lambda)}}{\lambda}, \tag{3.1}$$

for λ within the bounds $\lambda_- \le \lambda \le \lambda_+$, where λ_- and λ_+ are the minimum and maximum eigenvalue of $\mathbf{X_{K\times M}}$, respectively, given by $\lambda_\pm = 1 + 1/\mathbf{Q} \pm 1/\sqrt{\mathbf{Q}}$. As RMT method stated above, information related to correlation among time series can be extracted from the eigenvalue out of the range of $\lambda_- \le \lambda \le \lambda_+$.

In the subnet layer of simulation model, there are almost 1600 connections. Each connection is sampled 500 times. As a result, the size of the data matrix is $1600 * 500$. According to (3.1), if $K = 1600$, and $M = 500$, then $Q = 3.2$ and $\lambda_+ = 2.43$. In the experiment, the flow matrix is very different which predicts a finite range of eigenvalues depending on the ratio Q. In the simulation model, λ_+ is of order 100 with many experiments. It suggests that the largest eigenvalues are associated with strong correlation among the network. The result is consistent with the result in Renater [8].

4. The Analysis by PCA

Previously [9, 10], the traffic patterns are monitored by the RMT analysis. In this paper, it provides a extension of the previous work which is described in 3 parts. Fristly, PCA is

introduced in brief. Secondly, the usage of the small eigenvalue is given. Finally, the selection of observation points is solved by PCA.

4.1. The Analysis of the Largest Eigenvalue

The central idea of PCA is to reduce the dimensionality of a data set consisting of a large number of interrelated variables, while retaining as much as possible of the variation present in the data set [20]. Flow data is exactly in line with the conditions of PCA. Thus, we analyze the network-wide traffic by PCA.

A number of mathematical symbols are defined here. Let $\mathbf{X_{K\times M}}$ denote the flow matrix, in which x_{ij} represents the ith flow measured at the jth time interval. $\mathbf{X_{K\times M}}$ can also be expressed as $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_K)'$, in which $\mathbf{x_i}$, a vector of K variables, is denoted as the ith flow measured at M different times. $\mathbf{\Sigma}$ is the covariance matrix of $\mathbf{X_{K\times M}}$.

To well detecting the contribution of each flow to the whole network traffic, we should define a variable vector $\mathbf{Y}$ to indicate the network-wide traffic which lies on each vector of $\mathbf{X_{K\times M}}$. In the real network, the concerned parameters of links and nodes are almost the aggregation of flows, such as the receiving rate of each node, and the rate of each link. Then $\mathbf{Y}$ can be described as

$$Y = \mathbf{L}'\mathbf{X}, \tag{4.1}$$

in which the components l_i of $\mathbf{L}$ are the weight of the ith flow, representing the contribution of the whole networks. PCA has proved that the $\mathrm{var}(Y)$, in (4.2), reaches the maximum when $\mathbf{L}$ is the corresponding eigenvectors of the largest eigenvalue:

$$\mathrm{var}(Y) = \mathrm{Var}\left[\mathbf{L}'\mathbf{X}\right] = \mathbf{L}'\mathbf{\Sigma}\mathbf{L}. \tag{4.2}$$

As mentioned above, the element in the vector $\mathbf{L}$ represents the contribution of the corresponding flow, and the variable $\mathbf{Y}$ can be used to monitor the network-wide traffic. Thus, we can define a weight vector $\mathbf{S} = (s_1, s_2, \ldots, s_N)$ of the parameters about which we are concerned. N means the number of parameters, such as the number of subnets or the number of the links. If the pth element of $\mathbf{S}$ is the aggregation of the p_1th, p_2th, $\ldots$, p_mth flow, s_p can be represented as follows:

$$s_p = l_{p1}^2 + l_{p2}^2 + \cdots + l_{pm}^2, \tag{4.3}$$

in which $\mathbf{l}_{pi}$ corresponds to the weight of the p_ith flow. The detail of the weight vector of $\mathbf{S}$ is explained in our previous works [10]. And $\mathbf{S}$ can be used to describe the network traffic pattern, which is simulated in Section 5.

4.2. The Analysis of the Small Eigenvalues

RMT is a method to compare a random correlation matrix, constructed from mutually uncorrelated time series, against a correlation matrix for data under investigation. Deviations

between properties of the two matrices convey information about "genuine" correlations. In this section, we will explain the information extracted from the small eigenvalues.

According to previous analysis, we can argue the problem of small eigenvalues separately into two cases: the zero eigenvalue and the nonzero (small-positive number) eigenvalue.

(i) If the smallest eigenvalues are zero: the zero variance defines an exactly constant linear relationship between the elements of $\mathbf{X_{K\times M}}$. If such relationship exists, then we can infer that one of the elements of $\mathbf{X_{K\times M}}$ is redundant with one zero eigenvalue. When there are q zero eigenvalues, deduced by analogy, q elements are redundant without information lost. As a result, $(K - q)$ variables can be retained in Σ without information losts.

(ii) If the smallest eigenvalues are nonzero (a small-positive eigenvalue), the elements of $\mathbf{X_{K\times M}}$ have a near-exact linear relationship. It reveals that much but not all information is lost when the small eigenvalue is selected. It behaves a linearity with little disturbance.

According to the analysis above, the small eigenvalue reveals the linear or near-linear relationship between the elements of $\mathbf{X_{K\times M}}$. The corresponding eigenvector of the small eigenvalues is like a linear smooth filter here. Although little information is retained with the small eigenvalues, an extraordinary relationship is exhibited between the variables which correspond to a few observation points. In this case, the filtered information of $\mathbf{L}$ is from the variables unobserved and the fluctuation of all observed variables. Thus, when the eigenvectors of the small eigenvalues are used, the element of the vector $\mathbf{S}$ corresponding to the observation points will appear in the spectrum of eigen analysis. Thus, if an observation point failed, the raises of the corresponding node will vanish. Then, the effectiveness monitors can be detected by small eigenvalue analysis. This result will be simulated in Section 5.

4.3. The Selection and Placement of the Observation Points

In this paper, we propose an approach based on RMT and PCA to monitor the traffic pattern of the large-scale networks. In Renater study [8], the result is inferred from complete information of all network connection points. The Renater study is feasible benefiting from the small-scale networks with only 30 routers. However, real networks with hundreds of routers typically have large-scale. As link speeds and the number of flows increase, it is too expensive to arrange the observation points through the whole network, and thus real-time monitoring is difficult to achieve in practice. In this part, we will address the problems of selecting the number and arrangement of the observation points.

Above all, the number of the retained variables should be decided and they can represent most of the variation of flow matrix $\mathbf{X_{K\times M}}$. Using m PCs instead of p variables considerably reduces the dimensionality of the problem when $m \ll p$, but usually the values of all p variables are still needed in order to calculate the PCs, as each PC is likely to be a function of all p variables. It might be preferable if, instead of using m PCs to account for most of m, or perhaps slightly more, of the original variables, to account for most of the variation in $\mathbf{X_{K\times M}}$. Jolliffe [20] proposed that if Σ can be successfully described by only m principal components (PCs), then it will often be true that $\mathbf{X_{K\times M}}$ can be replaced by a subset of m variables, with a relative small loss of information.

There are many rules for deciding the number of PCs retained to account for most of the variation in $\mathbf{X_{K\times M}}$. Here, we select the most intuitive rules called "Cumulative Percentage of Total Variation" [20] to determine the number of observations. In this rule, the definition of "Percentage of variation accounted for by the first m PCs" is given by

$$t_m = 100\frac{\sum_{k=1}^{m}\lambda_k}{\sum_{k=1}^{M}\lambda_k}, \tag{4.4}$$

in which λ_k represents the kth largest eigenvalue, and M represents the dimensionality of $\mathbf{\Sigma}$. Suppose that there are N nodes in the networks. Thus, there are almost $2N(N-1)$ link flows. If m satisfies $t_m \geq 90\%$, the number of sampled link data is almost m.

PCA is probably the best known method for dimensionality reduction. Perhaps the most important problem in PCA is to determine the number of principal components m in a given data set, and to decide which subset or subsets of m variables are best. Moving on to the choice of m variables, the problem is focused on the method for selecting a subset of m variables that preserve most of the variation in the original matrix.

When the number of retained variables is fixed, how to place the monitoring points must be taken into consideration. There are many methods proposed to select "best" retained subsets of $\mathbf{X_{K\times M}}$ [21]. Some of the methods compared, and indeed some of those which performed quite well, are based on PCs. Other methods, including some based on cluster analysis of variables, are applicable to small data sets but not to data sampled from real networks. For example, the data set of the network model is $1600 * 1600$. If the method based on variables' cluster analysis is applied for such a large amount of data, the calculation time and the complexity are insupportable.

Our method is based on the first criterion of McCabe [22], which is described as (4.5):

$$\text{Minimize}\prod_{j=1}^{m^*}\theta_j, \tag{4.5}$$

where θ_j, $j = 1, 2, \ldots, m^*$ are the eigenvalues of the conditional covariance matrix of the m^* deleted variables. For this criterion it is computationally feasible to explore all possible subsets of variables. According to (4.5), we will find the deleted subsets of variables by statistical analysis, then some useful variables are retained. That means the corresponding link of each variable is found, and we can select as the observation points the subset of nodes which are passed by the links most frequently.

In the network model, when we follow "Cumulative Percentage of Total Variation", the resulting set of useful variables is $80 \sim 88$, that is, almost 1500 variables can be deleted. We repeat the calculation 500 times to find the subset nodes whose flow data are often retained. The statistical result is that the times of I1, J5 and K2, and so forth, whose degree are 4, 5, and 5, respectively, is more than other nodes. In the real networks, the problem may be explained simply. Our purpose is to select a few observation points and reserve enough information about $\mathbf{\Sigma}$. Aiming for this purpose, the intuitive idea is to select the nodes with large degree as observation points. That is because nodes with large degree have more information about more flows. Thus, the optimal placement is to lay the observation points on the nodes with large degree. Our experiment result has verified this idea in Section 5.

In conclusion, the selection of the observation points can be described in 5 steps as follows.

Step 1. The number of variables m could be calculated according (4.4).

Step 2. m variables could be selected as (4.5).

Step 3. Repeat Step 2 many times.

Step 4. Identify the most frequent variables in statistical results on Step 3.

Step 5. According to Table 2, identify the most frequent node which are observation points. I have added the steps in the end of Section 4.

5. Simulation Results

In order to verify the method stated in Sections 3 and 4, we do some experiments to monitor the traffic patterns by large and small eigenvalues, with all or few observation points. The experiment to monitor traffic pattern is conducted as follows.

(i) Monitoring peer-to-peer (P2P) traffic by the largest eigenvalue: the experiment is done with all of the routers and then repeated with few of the observation points selected by the method in Section 4.3. The related results are shown in Sections 5.1 and 5.4.

(ii) Monitoring the traffic patterns of link flows by the largest and the smallest eigenvalue: previously, the traffic patterns aggregated in the subnet is monitored. In this part, the link flows can also be monitored by the analysis of Sections 3 and 4. The related results are shown in Section 5.2.

(iii) Monitoring the traffic patterns of the observation points by the smallest eigenvalue: according to the analysis of Section 4.2, the observation points can be monitored by the small eigenvalues. The results are shown in Section 5.3.

In fact, our method of monitoring the traffic patterns can be applied in many scenarios, such as monitoring of mixed traffic pattern and detecting a victim of DDoS attacks, which are discussed, respectively, in our previous work [9, 10].

5.1. Monitoring the P2P Traffic Pattern by the Largest Eigenvalue and Corresponding Eigenvector

In the experiment, P2P traffic is simulated. Hosts as peers are dynamically distributed under A1, D2, H2, J5, K2, and K3. Peers under A1, D2, J5, K2, and K3 start communicating with each other at $t = 125$ time unit with the constant rate 25 packets per time unit. Then at $t = 160$ time unit, the peer groups under D2 leave and that under H2 join in at $t = 200$ time unit with the rate $t = 35$ packets per time unit. The simulation result is shown in Figure 4, which is calculated by the full data set.

In Figure 2, the simulation calculates $\mathbf{S}$ using M data within a moving window. The sampling rate of the data is $T = 0.5$ time unit, and the rate of the moving window is 10 data, that is, the step of the window is 10 data. The first 200 data points are regarded as the initial

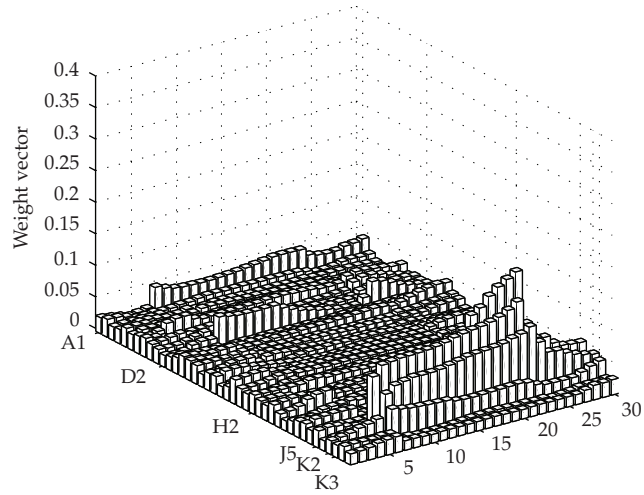

Figure 2: The pattern of P2P traffic, calculated by complete information. For each peer in the group A1, D2, J5, K2, and K3, rate = 25, starting time = 125 ($X = (125 * 2\text{–}200) * 0.1 = 5$). Then peers of D2 leave at $t = 160$ and peers of H2 join in at $t = 200$ ($X = (200 * 2\text{–}200) * 0.1 = 20$) with rate = 35.

data, sliding 10 data points each time unit. So, the x-axis represents how many times the time window slides. Apparently we can calculate corresponding time, for example, if $X = 5$, the corresponding time is $(200 + 5 * 10) * 0.5 = 125$ time unit. The y-axis represents the subnet number, and the z-axis represents the weight vector $\mathbf{S}$ of each subnet flows (which can be decided by (4.3) and the topology). The start and the end of each raises of corresponding subnets is almost the same as what we designed previously. Consequently, our method can extract the dynamically temporal-and-spatial information of the P2P traffic pattern well.

5.2. Monitoring the Aggregat Flow of Each Link by the Largest Eigenvalue and Corresponding Eigenvector

In order to monitor the link flow, we let all hosts except ones under I1, I2, and I3 send packets to I1 with a constant rate 5 packets per time unit.

Firstly, the largest eigenvalue and the corresponding eigenvectors of $\mathbf{\Sigma}$ are calculated. Secondly, in the light of Table 2, we should determine all the flows through each nodes. At last, the weight vector $\mathbf{S}$ of each link can be calculated according to (4.3).

Based on the static route in Table 2, if all hosts (except hosts distributed under I1 ~ I3) send packets to I1, the links between G and I, E and G, and C and E have the highest-rate flow. Similarly, the links between E and F, that is, EF and FE, have the lowest-rate flow. The result is illustrated in Figure 3. Obviously, the traffic pattern of GI, EG, and CE significantly raised; on the contrary, the traffic patterns of EF and FE seems low and flat.

In order to compare the traffic patterns in Figure 4 with the flows of each link, we use the Matlab *corrcoef* command to calculate the correlation coefficient of $\mathbf{S}$ and the rate of flows. A illustrated in Figure 6, the correlation coefficient is all in the range of $[0.7, 0.9]$. As a result, the traffic patterns of link flows in the backbone can be well monitored.

Conversely, if the traffic patterns are monitored by the small eigenvalue, the results in Figures 5 and 6 reveal much weaker correlation than those in Figures 3 and 4.

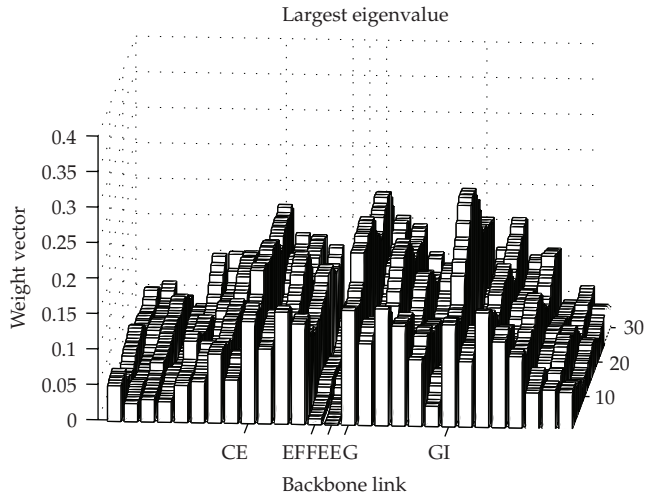

Figure 3: The traffic patterns monitored by the largest eigenvalue.

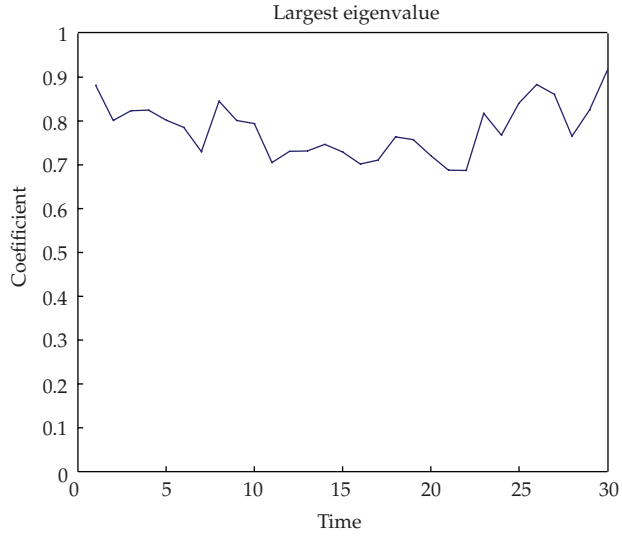

Figure 4: The correlation coefficient of the aggregate flow vector and the corresponding rate.

5.3. Monitoring Observation Points by the Small Eigenvalues

As explained in (4.2), with only a few observation points, when small eigenvalue is selected, the information including the nonobservation points and the fluctuation information of the observation points is filtered by the small eigenvalue. As a result, the information of the observation points will appear in the spectrum of the small eigenvalue. Here, we repeat the experiment in 5.1 with observation points (A3, B1, C1, C2, D1, F4), monitoring the traffic patterns by the second and third smallest eigenvalue, the result is shown in Figure 7. It is clear that the traffic patterns of the observation points raise sharply. As a result, the spectrum of small eigenvalues can be used to monitor the observation points distributed all over the

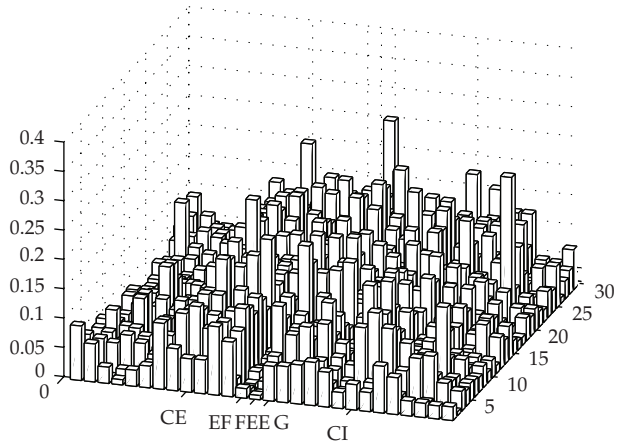

Figure 5: The traffic patterns monitored by the smallest eigenvalue.

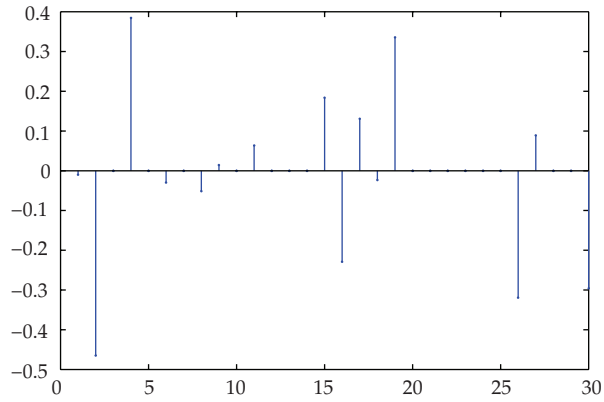

Figure 6: The correlation coefficient of the aggregate flow vector and the corresponding rate.

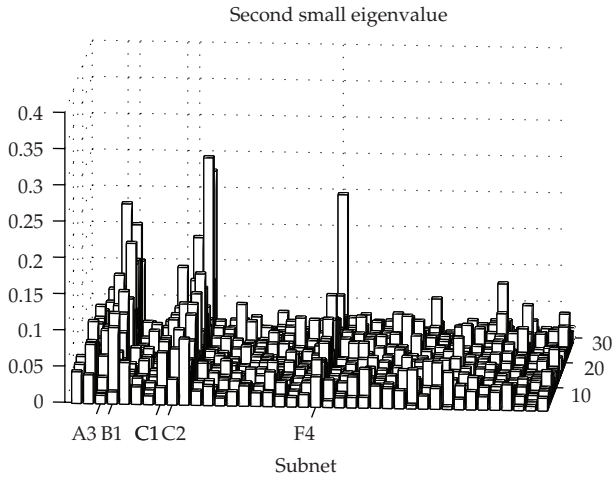

Figure 7: The spectrum of the second small eigenvalue with observation points A3, B1, C1, C2, D1, F4.

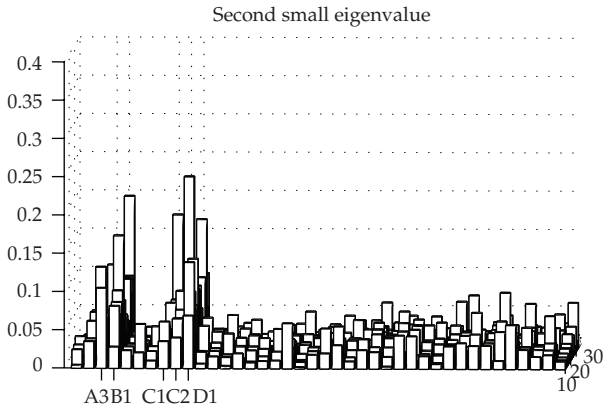

Figure 8: The spectrum of the second small eigenvalue with observation points A3, B1, C1, C2, D1.

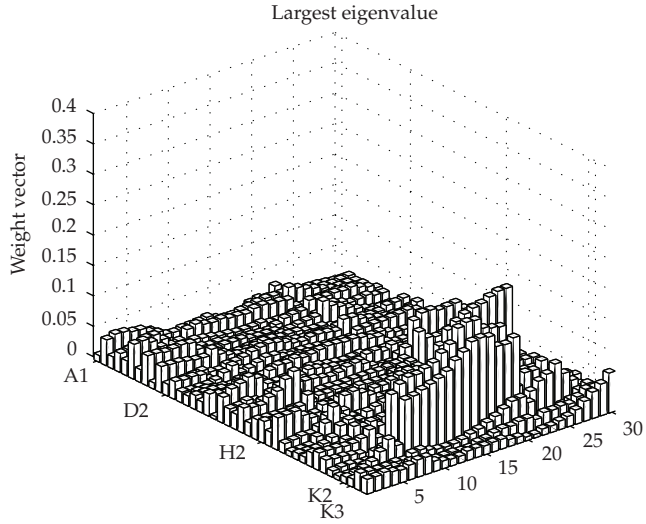

Figure 9: The pattern of P2P traffic, monitored by A3, B1, C1, C2, and D1. For each peer in the group A1, D2, J5, K2, and K3, rate = 25, starting time = 125. Then peers of D2 leave at t = 160 and peers of H2 join in at t = 200 with rate = 35.

networks. For example, if the observation point F4 failed, we can be informed by the spectrum shown in Figure 8, in which the raise of F4 disappeared.

5.4. Selection of the Observation Points

According to the analysis in (4.3), it is optimal to select the nodes with large degree as observation points. With this purpose, the experiment in 5.1 is redone with only a few observation points. As illustrated in Figure 9, selecting A3, B1, C1, C2, and D1 as observation points, some information of A1 and D2 is lost compared with Figure 2. With B4, D5, J5 and D2 selected, shown in Figure 10, the traffic pattern is similar to that in Figure 4 with little information lost. The analysis in (4.3) is verified.

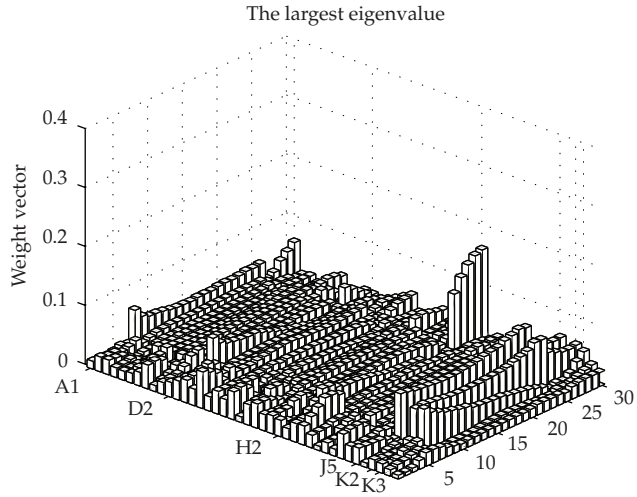

Figure 10: The pattern of P2P traffic, monitored by B4, D5, J5, and D2. For each peer in the group A1, D2, J5, K2, and K3, rate = 25, starting time = 125. Then peers of D2 leave at $t = 160$ and peers of H2 join in at $t = 200$ with rate = 35.

6. Conclusion

In this paper, a method based on RMT and PCA is proposed for monitoring and analyzing network-wide traffic patterns in the Internet. Our work makes three contributions: (1) we theoretically explain the information implied by the large/small eigenvalues of correlation matrices by PCA method, which is instructive for traffic measurement; (2) Using no more than 10% of subnet routers as observation points, key information about the Internet traffic pattern can be monitored and extracted by our method; (3) we apply RMT and PCA to capture various traffic patterns.

Acknowledgments

The authors would like to thank anonymous reviewers for their helpful comments. This work is supported by the National High Technology Research and Development Program (no. 2007AA012430).

References

[1] A. Greenberg, G. Hjalmtysson, D. A. Maltz, et al., "A clean slate 4D approach to network control and management," *ACM SIGCOMM Computer Communication Review*, vol. 35, no. 5, pp. 41–54, 2005.

[2] F. Hao, M. Kodialam, T. V. Lakshman, and S. Mohanty, "Fast, memory efficicent flow rate estimation using runs," *IEEE/ACM Transactions on Networking*, vol. 15, no. 6, pp. 1467–1477, 2007.

[3] G. Junwei and Z. Guofeng, "Traffic measurement for traffic engineering in IP/MPLS based telecom core networks," *The Journal China Universities of Posts and Telecommunications*, vol. 9, no. 2, 2002.

[4] R. Kawahara, T. Mori, N. Kamiyama, S. Harada, and S. Asano, "A study on detecting network anomalies using sampled flow statistics," in *International Symposium on Applications and the Internet Workshops (SAINT '07)*, Hiroshima, Japan, January 2007.

[5] L. Laloux, P. Cizeau, J.-P. Bouchaud, and M. Potters, "Noise dressing of financial correlation matrices," *Physical Review Letters*, vol. 83, no. 7, pp. 1467–1470, 1999.

[6] V. Plerou, P. Gopikrishnan, B. Rosenow, L. A. N. Amaral, and H. E. Stanley, "Universal and nonuniversal properties of cross correlaitons in financial time series," *Physics Review Letter*, vol. 83, no. 7, pp. 1471–1474, 1999.
[7] F. J. Dyson, "Statistical theory of the energy levels of complex systems. I," *Journal of Mathematical Physics*, vol. 3, pp. 140–156, 1962.
[8] M. Barthélemy, B. Gondran, and E. Guichard, "Large scale cross-correlations in Internet traffic," *Physical Review E*, vol. 66, no. 5, Article ID 056110, pp. 1–7, 2002.
[9] J. Yuan and K. Mills, "Monitoring the macroscopic effect of DDoS flooding attacks," *IEEE Transactions on Dependable and Secure Computing*, vol. 2, no. 4, pp. 324–335, 2005.
[10] J. Liu, W. Zhang, J. Yuan, D. Jin, and L. Zeng, "Monitoring the spatial-temporal effect of internet traffic based on random matrix theory," in *Proceedings of the 33rd Conference on Local Computer Networks (LCN '08)*, pp. 258–265, Montreal, AB, Canada, October 2008.
[11] M. Claypool, R. Kinicki, M. Li, J. Nichols, and H. Wu, "Inferring queue sizes in access networks by active measurement," in *Proceedings of the Passive and Active Measurement (PAM '08)*, April 2004.
[12] M. Weigle, K. Jeffay, and F. D. Smith, "Quantifying the effects of recent protocol improvements to standards-track TCP," in *Proceedings of the 11th IEEE/ACM Int'l Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication*, 2003.
[13] A. Feldmann, A. C. Gilbert, W. Willinger, and T. G. Kurtz, "The changing nature of network traffic: scaling phenomena," *Computer Communication Review*, vol. 28, no. 2, pp. 5–29, 1998.
[14] E. P. Wigner, "On the statistical distribution of the widths and spacings of nuclear resonance levels," *Mathematical Proceedings of the Cambridge Philosophical Society*, pp. 790–798, 1951.
[15] E. P. Wigner, "On the statistical distribution of the widths and spacings of nuclear resonance levels," *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 47, pp. 790–798, 1951.
[16] E. P. Wigner, "Results and theory of resonance absorption," in *Proceedings of the Conference on Neutron Physics by Time-of-Flight*, pp. 59–70, Gatlinburg, Tenn, USA, 1956.
[17] V. Plerou, P. Gopikrishnan, B. Rosenow, L. A.N. Amaral, and H. E. Stanley, "Universal and nonuniversal properties of cross correlations in financial time series," *Physical Review Letters*, vol. 83, no. 7, pp. 1471–1474, 1999.
[18] A. Feldmann, A. C. Gilbert, P. Huang, and W. Willinger, "Dynamics of IP traffic: a study of the role of variability and the impact of control," in *Proceedings of the ACM SIGCOMM Conference*, pp. 301–313, Boston, Mass, USA, 1999.
[19] A. M. Sengupta and P. P. Mitra, "Distributed of singular values for some random matrices," *Physics Review E*, vol. 60, no. 3, pp. 3389–3392, 1999.
[20] I. T. Jolliffe, *Principal Component Analysis*, Springer Series in Statistics, Springer, New York, NY, USA, 2nd edition, 2002.
[21] J. Cadima, J. O. Cerdeira, and M. Minhoto, "Computational aspects of algorithms for variable selection in the context of principal components," *Computational Statistics and Data Analysis*, vol. 47, no. 2, pp. 225–236, 2004.
[22] G. P. McCabe, "Principal variables," *Technometrics*, vol. 26, no. 2, pp. 137–144, 1984.