Sixth International Conference on Machine Learning and Applications

Feature Extraction Using Random Matrix Theory Approach

Viktoria Rojkova, Mehmed Kantardzic

Department of Computer Engineering and Computer
Science, University of Louisville, Louisville, KY 40292
email: {vbrozh01, mmkant01}@louisville.edu

***Abstract*—Feature extraction involves simplifying the amount of resources required to describe a large set of data accurately. In this paper, we propose to broaden the feature extraction algorithms with Random Matrix Theory methodology. Testing the cross-correlation matrix of variables against the null hypothesis of random correlations, we can derive characteristic parameters of the system, such as boundaries of eigenvalue spectra of random correlations, distribution of eigenvalues and eigenvectors of random correlations, inverse participation ratio and stability of eigenvectors of non-random correlations. We demonstrate the usefullness of these parameters for network traffic application, in particular, for network congestion control and for detection of any changes in the stable traffic dynamics.**

I. Introduction

In machine learning (ML) the eigenvalue-eigenvector decomposition of the cross-correlation matrix C of N variables and of L datapoints is primary tool for high-dimensionality reduction of the data. If the dataset is a complex, large-volume system of variables, it is very likely that subsets of variables are highly correlated with each other. The accuracy and reliability of classification or prediction model suffers great deal, if the highly correlated variables and variables, which are unrelated become mixed in the outcome.

For further steps of ML algorithms, the largest eigenvalue is expected to account for as much of data variability as possible, and each succeeding eigenvalue accounts for as much of the remaining variability as possible. Hence, the extreme eigenvalues are usually the points of the interest in ML domain. Non-extreme eigenvalues capturing lesser variance in the data, represent non-informative for classification "noise" of the system. The boundaries and distribution of the "noise" and as a consequence the spectral statistics of all eigenvalues and the distributions of all eigenvectors are never considered as learnable features of the system.

The Random Matrix Theory (RMT) was developed for studying the complex energy levels of heavy nuclei and it incorporates in part the series of statistical tests against the null hypothesis of random correlations between the variables. The detailed account of the RMT methodology is given in [2], [3], [4], [5], [6], [7]. Given that the tests accept this null hypothesis, the analytically found boundaries of the random correlations, spectral statistics of eigenvalues, distributions of eigenvectors are characteristic features of the system within the time of observation. Clearly, the deviating from null hypothesis eigenvalues and eigenvectors, which include the commonly used extreme eigenvalues, represent non-random features of the system.

In this paper, we explain the RMT methodology, as it was adopted for econometric and network traffic dynamics applications. We illustrate how the feature extraction field can be broadened with the RMT techniques on the example of multivariate network traffic system. The proposed RMT based statistics are able to differentiate the simulated uncongested and congested state of the traffic, thus, it can be used in congestion control. Another vital task in network traffic control is to accurately differentiate the stable, system specific dynamics from any temporal, anomalous changes. We demonstrate that the RMT based statistics captures any changes in system dynamics, as long as the number of elements of the system envolved into the changes is more than the number of significant participants in the most deviating from the RMT predictions eigenvector.

II. RMT Methodology

The RMT was employed in the financial studies of stock correlations [8], [9], communication theory of wireless systems [10], array signal processing [11], bioinformatics studies of protein folding [12]. We are not aware of any work, except for [1], where RMT techniques were applied to the Internet traffic system.

We adopt the methodology used in works on financial time series correlations (see [8], [9] and references therein) and later in [1], which discusses cross-correlations in Internet traffic. In particular, we quantify correlations between N traffic counts time series of L time points, by calculating the traffic rate change of every time series $T \ i = 1, \ldots, N$, over a time scale Δt,

$$G_i\left(t\right) \equiv \ln T_i\left(t + \Delta t\right) - \ln T_i\left(t\right) \tag{1}$$

where $T_i\left(t\right)$ denotes the traffic rate of time series i. This measure is independent from the volume of the traffic exchange and allows capturing the subtle changes in the traffic rate [1]. The normalized traffic rate change is

$$g_i\left(t\right) \equiv \frac{G_i\left(t\right) - \langle G_i\left(t\right)\rangle}{\sigma_i} \tag{2}$$

where $\sigma_i \equiv \sqrt{\langle G_i^2\rangle - \langle G_i\rangle^2}$ is the standard deviation of G_i. The equal-time cross-correlation matrix C can be computed as follows

$$C_{ij} \equiv \langle g_i\left(t\right) g_j\left(t\right)\rangle \tag{3}$$

0-7695-3069-9/07 $25.00 © 2007 IEEE
DOI 10.1109/ICMLA.2007.95

The properties of the traffic cross-correlation matrix C have to be compared with those of a random cross-correlation matrix [13]. In matrix notation, the interaction matrix C can be expressed as

$$C = \frac{1}{L} G G^T, \tag{4}$$

where G is $N \times L$ matrix with elements $\{ g_{i\,m} \equiv g_i \left(m \bigtriangleup t \right);\ i = 1, \ldots, N;\ m = 0, \ldots, L-1 \}$, and G^T denotes the transpose of G. Just as was done in [9], we consider a random correlation matrix

$$R = \frac{1}{L} A A^T, \tag{5}$$

where A is $N \times L$ matrix containing N time series of L random elements $a_{i\,m}$ with zero mean and unit variance, which are mutually uncorrelated as a null hypothesis.

Statistical properties of the random matrices R have been known for years in physics literature [2], [6], [3], [4], [5], [7]. In particular, it was shown analytically [14] that, under the restriction of $N \to \infty$, $L \to \infty$ and providing that $Q \equiv L/N (> 1)$ is fixed, the probability density function $P_{rm}\left(\lambda\right)$ of eigenvalues λ of the random matrix R is given by

$$P_{rm}\left(\lambda\right) = \frac{Q}{2\pi} \frac{\sqrt{\left(\lambda_+ - \lambda\right)\left(\lambda - \lambda_-\right)}}{\lambda} \tag{6}$$

where λ_+ and λ_- are maximum and minimum eigenvalues of R, respectively and $\lambda_- \leq \lambda_i \leq \lambda_+$. λ_+ and $\lambda_-$are given analytically by

$$\lambda_{\pm} = 1 + \frac{1}{Q} \pm 2\sqrt{\frac{1}{Q}}. \tag{7}$$

Random matrices display *universal* functional forms for eigenvalues correlations which depend on the general symmetries of the matrix only. First step to test the data for such a universal properties is to find a transformation called "unfolding", which maps the eigenvalues λ_i to new variables, "unfolded eigenvalues" ξ_i, whose distribution is uniform [5], [6], [7]. Unfolding ensures that the distances between eigenvalues are expressed in units of *local* mean eigenvalues spacing [5], and thus facilitates the comparison with analytical results.

The cumulative distribution function of eigenvalues counts the number of eigenvalues in the interval $\lambda_i \leq \lambda$,

$$F\left(\lambda\right) = N \int_{-\infty}^{\lambda} P\left(x\right) dx, \tag{8}$$

where $P\left(x\right)$ denotes the probability density of eigenvalues and N is the total number of eigenvalues. The function $F\left(\lambda\right)$ can be decomposed into an average and a fluctuating part,

$$F\left(\lambda\right) = F_{av}\left(\lambda\right) + F_{fluc}\left(\lambda\right), \tag{9}$$

Since $P_{fluc} \equiv dF_{fluc}\left(\lambda\right)/d\lambda = 0$ on average,

$$P_{rm}\left(\lambda\right) \equiv \frac{dF_{av}\left(\lambda\right)}{d\lambda}, \tag{10}$$

is the averaged eigenvalues density. The dimensionless, unfolded eigenvalues are then given by

$$\xi_i \equiv F_{av}\left(\lambda_i\right). \tag{11}$$

Three known universal properties of GOE matrices (matrices whose elements are distributed according to a Gaussian probability measure) are: (i) the distribution of nearest-neighbor eigenvalues spacing (NNS) $P_{GOE}\left(s\right)$

$$P_{GOE}\left(s\right) = \frac{\pi s}{2} exp\left(-\frac{\pi}{4} s^2\right), \tag{12}$$

(ii) the distribution of next-nearest-neighbor eigenvalues spacing (NNNS), which is according to the theorem due to [4] is identical to the distribution of nearest-neighbor spacing of Gaussian symplectic ensemble (GSE),

$$P_{GSE}\left(s\right) = \frac{2^{18}}{3^6 \pi^3} s^4 exp\left(-\frac{64}{9\pi} s^2\right) \tag{13}$$

and finally (iii) the "number variance" statistics Σ^2, defined as the variance of the number of unfolded eigenvalues in the intervals of length l, around each ξ_i [5], [7], [6].

$$\Sigma^2\left(l\right) = \left\langle \left[n\left(\xi, l\right) - l\right]^2 \right\rangle_{\xi}, \tag{14}$$

where $n\left(\xi, l\right)$ is the number of the unfolded eigenvalues in the interval $\left[\xi - \frac{l}{2}, \xi + \frac{l}{2}\right]$. The number variance is expressed as follows

$$\Sigma^2\left(l\right) = l - 2\int_0^l \left(l - x\right) Y\left(x\right) dx, \tag{15}$$

where $Y\left(x\right)$ for the GOE case is given by [5]

$$Y\left(x\right) = s^2\left(x\right) + \frac{ds}{dx}\int_x^{\infty} s\left(x'\right) dx', \tag{16}$$

and

$$s\left(x\right) = \frac{sin\left(\pi x\right)}{\pi x}. \tag{17}$$

III. Application of RMT tests to the system of network traffic time series

In this paper, the RMT tests are applied to the averaged traffic count data collected from inter-VLAN backbone routers system of the University of Louisville. The system consists of nine interconnected multi-gigabit backbone routers, over 200 Ethernet segments and over 300 VLAN subnets. We collected the 5 min interval Simple Network Manage Protocol (SNMP) traffic count data for seven days.

We constructed inter-VLAN traffic cross-correlation matrix C with number of time series $N = 497$ and number of observations per series $L = 2015$, ($Q = 4.0625$) so that, $\lambda_+ = 2.23843$ and $\lambda_- = 0.253876$. Our first goal is to compare the eigenvalue distribution $P\left(\lambda\right)$ of C with $P_{rm}\left(\lambda\right)$ [13]. The empirical probability distribution $P\left(\lambda\right)$ is then given by the corresponding histogram. The resulting distribution $P\left(\lambda\right)$is displayed in Figure 1 and compared to the probability distribution $P_{rm}\left(\lambda\right)$ taken from Eq. (6) calculated for the same value of traffic time series parameters ($Q = 4.0625$). The solid curve demonstrates $P_{rm}\left(\lambda\right)$ of Eq.(6). The largest eigenvalue shown in inset has the value $\lambda_{497} = 8.99$. The deviations from the RMT predictions are zoomed on the inset to Figure 1.

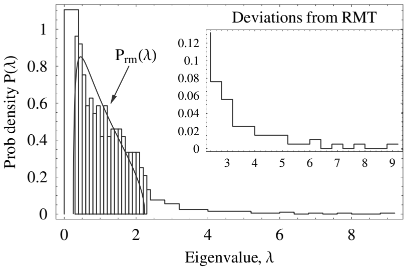

Figure 1. Empirical probability distribution function $P(\lambda)$ for the inter-VLAN traffic cross-correlations matrix C (histogram).

We note the presence of "bulk" (RMT-like) eigenvalues which fall within the bounds $[\lambda_-,\lambda_+]$ for $P_{rm}(\lambda)$, and presence of the eigenvalues which lie outside of the "bulk", representing deviations from the RMT predictions. In particular, largest eigenvalue $\lambda_{497} = 8.99$ for seven days period is approximately four times larger than the RMT upper bound λ_+.

The unfolded eigenvalues ξ_i are obtained by following the phenomenological procedure referred to as Gaussian broadening [15], (see [15], [9], [8]). The first independent RMT test is the comparison of the distribution of the unfolded eigenvalues NNS $P_{nn}(s)$, where $s \equiv \xi_{k+1} - \xi_k$ with $P_{GOE}(s)$ [5], [6], [7]. The agreement between eigenvalues cumulative distribution $F_{av}(\lambda)$ and distribution of unfolded eigenvalues ξ_i is presented in Figure 2, confirmed by D value of Kolmogorov-Smirnov goodness of fit test, which is $< 1/\sqrt{N}$, for $N > 35$ (same value of D is obtained for second and third tests). The empirical probability distribution of unfolded eigenvalues NNS $P_{nn}(s)$ and $P_{GOE}(s)$ are presented in Figure 3. The solid line represents the agreement between empirical probability distribution $P_{nn}(s)$ and the distribution of NNS of the GOE matrices $P_{GOE}(s)$ testifies that the positions of two adjacent empirical unfolded eigenvalues at the distance s are correlated just as the eigenvalues of the GOE matrices.

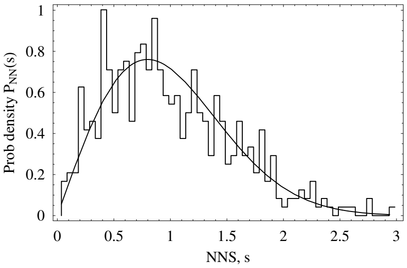

Figure 2. Nearest-neighbor spacing distribution $P_{nn}(s)$ of unfolded eigenvalues ξ_i of cross-correlation matrix C.

Next, we took on the distribution $P_{nnn}(s')$ of NNNS $s' \equiv \xi_{k+2} - \xi_k$ between the unfolded eigenvalues. According to [4] this distribution should fit to the distribution of NNS of the GSE. This correspondence is demonstrated in Figure 4. The solid line shows $P_{GSE}(s)$. Finally, the long-range two-point eigenvalue correlations were tested. It is known [5], [6], [7], that if eigenvalues are uncorrelated we expect the number variance to scale with l, $\Sigma^2 \sim l$. Meanwhile, when the unfolded eigenvalues of C are correlated, Σ^2 approaches constant value, revealing "spectral rigidity" [5], [6], [7]. In Figure 5, the Poissonian number variance is contrasted with the one which was observed. Clearly, eigenvalues belonging to the "bulk" exhibit universal RMT properties. The dashed line corresponds to the case of uncorrelated eigenvalues. These findings show that the system of inter-VLAN traffic has a *universal* part of eigenvalues spectral correlations, shared by broad class of systems, including chaotic and disordered systems, nuclei, atoms and molecules. Thus it can be concluded, that the bulk eigenvalue statistics of the inter-VLAN traffic cross-correlation matrix C are consistent with those of real symmetric random matrix R, given by Eq. (5) [14].

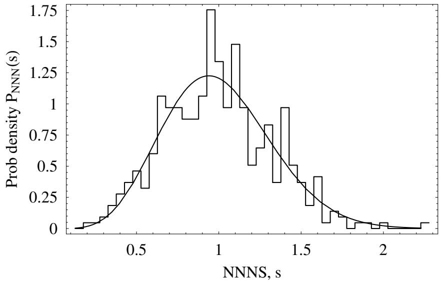

Figure 3. Next-nearest-neighbor eigenvalue spacing distribution $P_{nnn}(s')$.

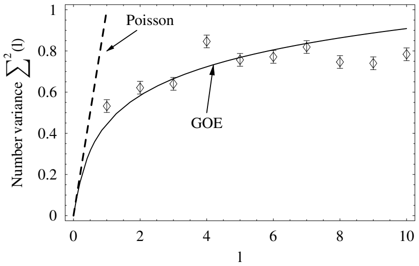

Figure 4. Number variance $\Sigma^2(l)$ calculated from the unfolded eigenvalues ξ_i of C.

IV. Feature Extraction

Feature extraction is a general term for methods of constructing combinations of the *variables* to get around the problems of overfitting while still describing the system with sufficient accuracy. The term *parameter* quantitifies certain relatively constant characteristics of the system and is intermediate in status between a variable and a constant. We propose to use the parameters of the system, which the RMT methodology provides, to describe the system with sufficient degree of accuracy.

A. Inverse participation ratio of eigenvectors

Eigenvectors of inter-VLAN traffic cross-correlation matrix C, determined by $Cu^k = \lambda_k u^k$, where λ_k is $k-th$ eigenvalue. The predictions are that all components participate in the eigenvectors of random interactions, while the number of significant contributors in eigenvectors of meaningful interactions is few. The IPR quantifies the reciprocal of the number of significant components of the eigenvector. For the eigenvector u^k it is defined as

$$I^k \equiv \sum_{l=1}^{N} \left[u_l^k\right]^4, \tag{18}$$

where u_l^k, $l = 1, \ldots, 497$ are components of the eigenvector u^k. The IPR is quite indicative in terms of signaling the number of significant u_k^l, i.e. contributors to the eigenvector of interest. For example, if we have reasons to expect absence of correlations between routers input into the experimental data, $I_k(0)$ should have its value around $1/\sqrt{N}$. Indeed, the eigenvector is normalized, thus $\sum_{l=1}^{N}[u_k^l]^2 = 1$. It has N components, and they are all roughly the same in magnitude (otherwise correlations must be present). Therefore, $u_l^k \simeq 1/\sqrt{N}$, and $I_k(0) \simeq 1/N$ has $I^k = 1/N$. Note, that since N is typically much greater than 1, any finite value of IPR signals *localization* (decrease in in the number of eigenvector contributors).

The illustrative example of IPR as a parameter for congestion control is presented further. With the help of OPNET modeler simulation tool, we simulated the network layout with the same number of backbone routers and subnets. We have placed the nodes with high traffic loads in the simulated layout and insured the loss of utilities with the performance statistics provided by OPNET. The congestion of the traffic is defined as the loss of utility to a network user due to high traffic loads [17]. The packet loss ratio of simulated congested and uncongested traffic are presented in Figure 6a and 6b, respectively . The IPR of cross-correlation matrix C versus the position of eigenvalue in spectrum for simulated congested, simulated uncongested, real traffic and control (random matrix) are presented in Figure 7 . The control green line is the IPR of eigenvectors of random cross-correlation matrix R. Blue line is IPR of real traffic. Yellow and red lines are IPR of simulated uncongested and congested traffic correspondingly. As we can see, eigenvectors of real and simulated uncongested traffic (blue and yellow lines) are closer to the control IPR. The IPR of congested traffic (red line) shows the higher localization level. The localization signifies the restrictions in communication or correlation pattern formation. Even though there are still islands of freely communicating network nodes, the number of nodes involved in such communication decreases. The system attempts to keep the balance by dropping the packets, which is testified by packets loss ratio measurement.

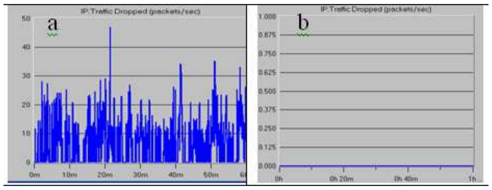

Figure 5. Dropped packets per second, (a) congested traffic and (b) uncongested traffic.

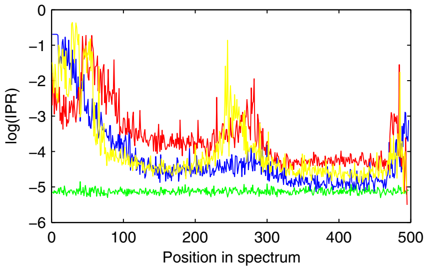

Figure 6. Inverse participation ratio as a function of eigenvalue λ.

B. Stability of inter-vlan traffic interactions - overlap matrix

We assume that the health of inter-VLAN traffic is expressed by stability of its interactions in time. Meanwhile, the temporal critical events or anomalies will cause the temporal instabilities. The "deviating" eigenvalues and eigenvectors provide us with stable in time snapshots of interactions representative of the entire network. Therefore, these eigenvectors judged on the basis of their IPR can serve as monitoring parameters of the system stability.

We expect to observe the stability of inter-VLAN traffic interactions in the period of time used to compute traffic cross-correlation matrix C. To observe the time stability of inter-VLAN meaningful interactions we compute the "overlap matrix" of the deviating eigenvectors for the time period t and deviating eigenvectors for the time period $t + \tau$, where $t = 60h, \tau = \{0h, 3h, 12h, 24h, 36h, 48h\}$.

First, we obtained matrix D from $p = 57$ eigenvectors, which correspond to p eigenvalues outside of the RMT upper bound λ_+. Then we compute the "overlap matrix" $O(t, \tau)$ from $D_A D_B^T$, where O_{ij} is a scalar product of the eigenvector u^i of period A (starting at time $t = t$) with u^j of period B at the time $t = t + \tau$,

$$O_{ij}(t, \tau) \equiv \sum_{k=1}^{N} D_{ik}(t) D_{ik}(t + \tau) \tag{19}$$

The values of $O_{ij}(t, \tau)$ elements at $i = j$, i.e. of diagonal elements of matrix O will be 1, if the matrix $D(t + \tau)$ is identical to the matrix $D(t)$. Clearly, the diagonal of the "overlap matrix" O can serve as an indicator of time stability of p eigenvectors outside of the RMT upper bound λ_+. The gray scale colormap of the "overlap matrices" $O(t = 60h, \tau = \{0h, 3h, 12h, 24h, 36h, 48h\})$ is presented in Figure 8. Black color of grayscale represents $O_{ij} = 1$, white color represents $O_{ij} = 0$. At lag $\tau = 3$ hours the inter-VLAN interactions show the highest degree of stability. For further lags the overall stability decays. As the analysis of deviating eigenvectors content showed, the highly interacting traffic time series are time series of service based VLANs, intended for

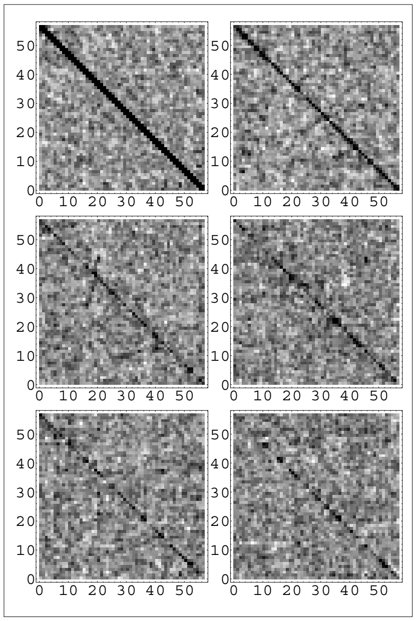

Figure 7. The grayscale of overlap matrix $O(t,\tau)$ at $t = 60h$ and $\tau = \{0h, 3h, 12h, 24h, 36h, 48h\}$.

routing. In fact, we found three types of connections groupings. One group contains connections, which are interlinked on the router. We recognize them as, vlan_X-router incoming traffic connection, vlan_X-router_firewall connection and vlan_X-router outgoing traffic connection. The connections, which are listed as vlan_Y-router1, vlan_Yrouter2, vlan_Y-router3, etc..., are reserved for the same service on every router and comprise another group. Final group of vlan-router connections constituted of connections, which interact due to the routing. Particular network services are evoked at the same time and active for the same period of time, which explains the stability and consequent decay of deviating eigenvectors of traffic interactions.

C. Meshgrid of eigenvector components and spatial-temporal representation of traffic load

At any time the readings from the network nodes give an instantaneous traffic load pattern. This pattern can be viewed as an expansion in terms of eigenvectors of matrix C in the following sense. An eigenvector u_k is a set of different intensities of network-wide traffic load satisfying

$$Cu_k = \lambda_k u_k.$$

Among possible configurations of network-wide traffic load u_k^i is an amount of traffic load on a particular node. Then, ratio u_k^i/G_k is equal to the number of nodes involved in the mutual interaction. For a variance of a traffic load at a given node we get:

‘

$$\sigma_k^2 = \left\langle \left(\sum_{i=1}^{M} \frac{u_k^i}{G_i} \delta G_i \right)^2 \right\rangle = \sum_{i,j=1}^{M} u_k^i u_k^j C_{ij} = u_k^T C u_k. \tag{20}$$

At this point we can employ the result of Esq. (**??**) to realize, that the variance of the traffic load at a given node is specified by the corresponding eigenvalue: $\sigma_k^2 = \lambda_k$. Once again, this is true for a network-wide traffic described by the u_k. By contrast, there is no correlation between two network-wide traffic loads attributed to two eigenvectors u_k and u_l:

$$\left\langle \left(\sum_{i=1}^{M} \frac{u_k^i}{G_i} \delta G_i \right) \left(\sum_{j=1}^{M} \frac{u_l^j}{G_j} \delta G_j \right) \right\rangle = u_k^T C u_l = 0,\, b \neq l.$$

With this in mind, if we meshgrid the eigenvector components against time we will obtain the dynamics of particular network-wide traffic load in space, due to precise location of significant components, and time. The meshgrid of last eigenvector u^{497} components for time period $t + \tau$, where $t = 36$ hours and $\tau = 6n$, $n \in \{0, 1, \ldots, 7\}$ is shown in Figure 10.

D. Network topological representation of the traffic load

Another visualization example is inspired by popular among network practitioners technique - network topological representation as a graph. The network-wide traffic load, expressed by the components of the eigenvector of interest, in our case eigenvectors outside of the RMT boundaries, can be visualized as an indirect graph with active and inactive edges. Active edge corresponds to the traffic time series, which is a significant participant in a given eigenvector (traffic load). The illustration of this technique is presented in Experiments subsection.

V. Experiments

To shed more light on the possibilities of anomaly detection we conducted the experiments to establish spatial-temporal traces of instabilities caused by artificial and temporal increase of the correlation in normal uncongested inter-VLAN traffic. Next, we explored the possibility to distinguish different types of increased temporal correlations.

Experiment 1

We selected the traffic counts time series representing the components of the RMT eigenvector and increased the correlation between these series for three hour period. The proposed monitoring parameters show the dependence of system stability on the number of temporarily correlated time series (see Figure 9) . Presented in Figure 9a, left to right are eigenvalue distribution, IPR of eigenvectors and the overlap matrix of deviating eigenvectors. The same parameters with induced temporal correlation between ten and twenty time series are shown on Figures 9b and 9c correspondingly. One can conclude that increased temporal correlation between ten time series does not affect system stability. Meanwhile, when the number of temporarily correlated time series reaches the number of significant participants of eigenvector u^{497} of largest eigenvalues (~ 22), the system becomes visibly unstable. The largest eigenvalue changes from 10 to 12, the tail of inverse participation ratio plot is extended and the diagonal of overlap matrix disappears In addition, we visualize in Figure 10 the system instability during the increase of correlation between twenty time series with spatial-temporal representation of eigenvector u^{497} . In Figure 10a the spatial-

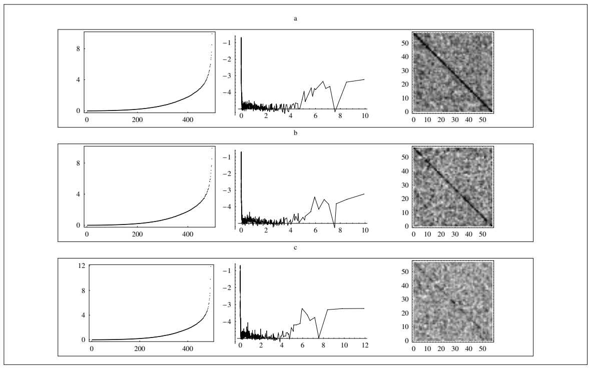

Figure 8. Eigenvalues distribution, IPR and overlap matrix of deviating eigenvectors.

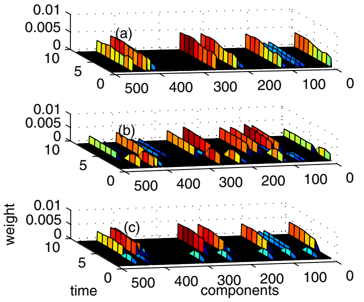

Figure 9. (a) The weights of components of u^{497} plotted for time period from 36 to 84 hours of uninterrupted traffic with 6 hours interval. (b) The weights of components of u^{497} plotted with respect to the same time period, with induced three hours correlation. (c) The weights of components of u^{496} plotted with respect to the same time period, with induced three hours correlation.

temporal pattern of u^{497} captures precise locations of system-specific interactions of uninterrupted traffic for 84 hours of observation.

The abrupt change of this pattern in Figure 10b indicates the starting point of induced correlation between twenty traffic time series usually interacting in a random fashion. The weights and locations of significant components of eigenvector U^{496} are suppressed and replaced by the weights and locations of significant components of eigenvector u^{497} when the interruption ends. Thus, we are able to observe the end point of the induced correlations in Figure 10c, which represents weights of components of eigenvector u^{496} plotted with respect to the same time intervals. With this setup we are able to locate the anomaly in time and space. Translated to network topological representation, the behavior of eigenvectors u^{497} and u^{496} during our manipulations with inter-VLAN traffic may be monitored with the following graphs (see Figure 11) .

Experiment 2

In the previous experiment we injected just one type of increased correlation among time series. Now, the two and three different types of induced correlations produce different spatial-temporal patterns on eigenvector u^{497} components (see Figure 12) . Time series for increased correlation are obtained in the same way as in Experiment 1. We increased the correlation between series by inducing elements from distributions of sine function and quadratic function, respectively for three hours. In Figure 12a, one type of three hours correlation is induced among ten traffic time series and another type of correlation among other ten time series. Three different types of three hours correlations are induced among twenty traffic time series in Figure 12b. The sorted in decreasing order content of significant components shows that time series tend to tend to group according to the type of correlation they are involved in.

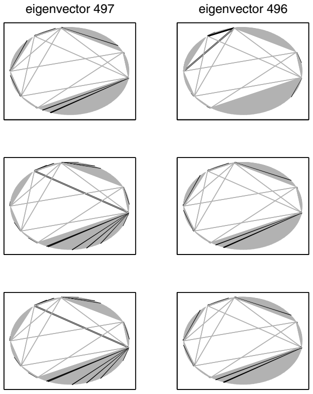

Figure 10. Left column - behavior of u^{497} during time period from 48h to 60h with 6h time window, induced correlation starts at 54h and lasts for 3h. Right column - behavior of u^{496} in same conditions.

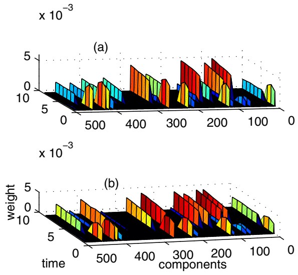

Figure 11. (a) The weights of components of u^{497} plotted for time period from 36 to 84 hours with 6 hours interval, two different types of induced correlations. (b) The weights of components of u^{497} plotted with respect to the same time period, three different types of induced correlations.

VI. Conclusion

The feature extraction in general is about the reduction of the number of variables, which classify the different states of the system with sufficient accuracy and generalize to the new samples of the system's data. Meanwhile, the parameters represent the certain relatively constant characteristics of the system. When evaluating the function over a domain or determining the response of the system over a period of time, the independent variables are modulated, while the parameters are held constant. The function or system may then be reevaluated or reprocessed with different parameters, to give a function or system with different behavior. Hence, parameters play the role of classifiers of the different states of the system. Thus, finding the parameters of the system fits the feature extraction goal. The RMT methodology provides the analytically defined parameters of the system.

ACKNOWLEDGMENT

This research was partially supported by grant from the US Department of Treasury through subcontract from the University of Kentucky. The authors thank Igor Rozhkov for consulting on the RMT methodology.

REFERENCES

[1] M. Barthelemy, B. Gondran and E. Guichard, Large scale cross-correlations in internet traffic, arXiv:cond0mat/0206185 vol **2** 3 Dec 2002.

[2] E.P. Wigner, On a class of analytic functions from the quantum theory of collisions, Ann. Math. **53**, 36 (1951), Proc. Cambridge Philos. Soc. **47**, 790 (1951).

[3] F. Dyson, Statistical theory of the energy levels of complex systems, J. Math. Phys. **3**, 140 (1962).

[4] F. Dyson and M.L. Mehta, Statistical theory of the energy levels of complex systems, J. Math. Phys. **4**, 701, 713 (1963).

[5] M.L Mehta, Random matrices (Academic Press, Boston, 1991).

[6] T.A. Brody, J.Flores, J.B. French, P.A. Mello, A. Pandey, and S.S.M. Wong, Random-matrix physics: spectrum and strength fluctuations, Rev. Mod. Phys. **53**, 385 - 479, issue **3**, July 1981.

[7] T. Guhr, A. Muller-Groeling, and H.A. Weidenmuller, Random matrix theories in quantum physics: common concepts, Phys. Rep. **299**, 190 (1998).

[8] S. Sharifi, M. Crane, A. Shamaie and H. Ruskin, Random matrix portfolio optimization: a stability approach, Physica A **335** (2004) 629-643.

[9] V. Plerou, P. Gopikrishnan, B. Rosenow, L. A. Nunes Amaral, T. Guhr, and H.E. Stanley, Random matrix theory approach to cross correlations in financial data, Phys. Rev. E, vol **65**, 066126, 27 June 2002.

[10] A. Tulino and S. Verdu, Random matrix theory and wireless communications, Communications and Information theory, vol **1**, issue **1**, June 2004, 1 - 182.

[11] D. Tse, Multiuser receivers, random matrices and free probability, Proceedings of 37th Ann. Allerton Conf., Monticello, IL, September 1999.

[12] A. Zee, Random matrix theory and RNA folding, Acta Physica Polonica B, vol **36**, No **9**, June 2005.

[13] L. Laloux, P. Cizeau, J.-P. Bouchaud, and M. Potters, Noise dressing of financial correlation matrices, Phys. Rev. Lett. **83**, August 1999, 1467-1470.

[14] A.M. Sengupta and P.P. Mitra, Distributions of singular values for some random matrices, arXiv:cond-mat/9709283 vol **1** 25 September 1997.

[15] H. Bruus and J.-C. Angles d'Auriac, Energy level statistics of two-dimensional Hubbard model at low filling, arXiv:cond-mat/9610142 vol **1** 18 October 1996.

[16] H.-J. Stockman, Quantum Chaos: an introduction, 1999.

[17] K. Srinivasan, Congestion Control in Computer Networks, EEECS Department University of California, Berkley, Technical Report No. UCB/CSD-91-649, 1991.