

# Random matrix analysis of localization properties of Gene co-expression network

Sarika Jalan,<sup>1</sup> Norbert Solymosi,<sup>2</sup> Gábor Vattay,<sup>2</sup> and Baowen Li<sup>1,3</sup>

<sup>1</sup>*Department of Physics and Centre for Computational Science and Engineering,  
National University of Singapore, 117456, Republic of Singapore*

<sup>2</sup>*Department of the Physics of Complex Systems, Eötvös University,  
H-1117 Pázmány Péter sétány 1/A, Budapest, Hungary*

<sup>3</sup>*NUS Graduate School for Integrative Sciences and Engineering, 117546, Republic of Singapore*

We analyze gene co-expression network under the random matrix theory framework. The nearest neighbor spacing distribution of the adjacency matrix of this network follows Gaussian orthogonal statistics of random matrix theory (RMT). Spectral rigidity test follows random matrix prediction for a certain range, and deviates after wards. Eigenvector analysis of the network using inverse participation ratio (IPR) suggests that the statistics of bulk of the eigenvalues of network is consistent with those of the real symmetric random matrix, whereas few eigenvalues are localized. Based on these IPR calculations, we can divide eigenvalues in three sets; (A) The non-degenerate part that follows RMT. (B) The non-degenerate part, at both ends and at intermediate eigenvalues, which deviate from RMT and expected to contain information about *important nodes* in the network. (C) The degenerate part with *zero* eigenvalue, which fluctuates around RMT predicted value. We identify nodes corresponding to the dominant modes of the corresponding eigenvectors and analyze their structural properties.

PACS numbers: 89.75.Hc, 64.60.Cn, 89.20.-a

## I. INTRODUCTION

### A. Complex Networks

Gene expression information captured in microarrays data for a variety of environmental and genetic perturbations, in conjunction with other sources such as protein-protein/protein-DNA interaction and operon organization data, promises to yield unprecedented insights into the organization and functioning of biological systems [1, 2]. It has been increasingly realized that dissecting the genetic and chemical circuitry prevents us from further understanding the biological processes as a whole. In order to understand the complexities involved, all reactions and processes should be analyzed together. To this end, network theory will be used. It has been getting fast recognition to study systems which could be defined in terms of units and interactions among them. These studies revealed that the available data from gene co-expression network share some unexpected features with other complex networks as diverse as the Internet routers. In order to understand the behavior of complex systems such as gene co-expression network, several simple models, based on the simple principles and captures some essential features of the system, have been introduced, these models are [3–5].

In this paper, by using network theory and random matrix theory (RMT), we analyze gene co-expression network. First we generate network from the gene co-expression data collected from six brain regions that are metabolically relevant to Alzheimer’s disease [6] by using appropriate threshold, and then study the spectra of this network under the RMT framework. Information about the genes that are preferentially expressed during the course of Alzheimer’s disease could improve our

understanding of the molecular mechanisms involved in the pathogenesis of this common cause of cognitive impairment in senior persons, provide new opportunities in the diagnosis, early detection, and tracking of this disorder, and provide novel targets for the discovery of interventions to treat and prevent this disorder. Information about the genes that are preferentially expressed in relationship to normal neurological aging could provide new information about the molecular mechanisms that are involved in normal age-related cognitive decline and a host of age-related neurological disorders, and they could provide novel targets for the discovery of interventions to mitigate some of these deleterious effects.

Co-expression networks have also been known as relevance networks. The terminology has been introduced by Butte and Kohane [7]. Since then properties of the relevance networks have been extensively studied [8].

The paper is organized as follows: after introductory sub-section on the relevance of network theory and gene co-expression network, we discuss the recent outcome of RMT analysis of complex networks in the following sub-section B. Main goals of our eigenvector analysis are written in the subsection C. Section II describes the important achievements of RMT and explains its various properties we use in our analysis. Section III sheds light on the data and network construction. Section IV presents various numerical results. Section V concludes the paper with a discussion on the relevance of current analysis, as well suggests future directions.

### B. RMT of Network Spectra

Our previous work [9] showed that various vastly studied model networks follow random matrix predictions of

Gaussian orthogonal statistics (GOE) at the level repulsion domain. We demonstrated that nearest neighbor spacing distribution (NNSD) of protein-protein interaction network of budding yeast follows RMT prediction as well [9]. This is a promising result which suggests that these networks can be modeled as a random matrix chosen from an appropriate ensemble. The universal GOE statistics of eigenvalues fluctuations could be understood as some kind of randomness spreading over the protein-protein interaction network and model networks capturing real world properties. Recently, covariance matrix of amino acid displacement has been analyzed under RMT framework [10]. The analysis shows that the bulk of eigenvalues follows universal GOE statistics of RMT. In the present paper, we analyze gene co-expression network [6] under RMT framework. First we calculate nearest neighbor spacing distribution of network spectra, and then perform eigenvector analysis to detect nodes having specific contribution to network.

### C. Important nodes and connections

It is now well known that various real world systems are scale-free network[3]. The scale-free nature of networks suggests that there exist few nodes with very high degrees. Motivated by this finding they suggested that since these nodes are responsible to hold the whole networks and henceforth are the most important ones. Some other analysis (by Newman and others) of real-world networks show that complex networks have community or module structure [11, 12]. Modules are the division of network nodes within which the network connections are dense, but between which they are sparser. The modularity concept assumes that system functionality can be partitioned into a collection of modules and each module performs an identifiable task, separable from the functions of other modules [13]. Analysis of module structure involves *betweenness* measure. Betweenness of an edge is defined as the number of shortest path between pairs of nodes going through the edge. Betweenness studies of real world networks suggests that the nodes connecting the different communities are the most important ones, which has been verified in the metabolic networks by Amaral et. al.[13].

Above description emphasizes on the importance of nodes depending on their position in the network, as these nodes characterize network properties. On the other hand Erdős-Rényi (ER) and Strogatz-Watts (SW) models emphasize on the importance of random connections in the networks. In the ER model any two nodes are connected with probability  $p$ . One of the most interesting property of ER model is the sudden emergence of various global properties, such as, emergence of a giant cluster. As  $p$  increases, while number of nodes in the graph remains constant, the giant cluster emerges through a phase transition [14]. Further, the SW model shows the small world transition with the fine tuning of number of

random connections [15]. Our previous RMT analysis of the spectra of SW model networks [9] show that at the SW transition there is a transition to the *spreading of randomness* in the network characterized by the correlations between nearest eigenvalues. In the current paper we analyze spectra of the gene co-expression network under RMT framework. Particularly we study eigenvectors of the adjacency matrix of this network. The spectra has two parts, one part which follows RMT predictions of universal GOE statistics and other part which does not follow RMT prediction. The eigenvectors deviating from the RMT prediction provide information about the *influential or important nodes* in the network.

## II. RANDOM MATRIX STATISTICS

RMT deals with the statistical properties of matrices with independent random entries. To be self-consistent, we give a brief introduction of the RMT here, and explain various RMT properties of eigenvector components which we will use in our analysis. RMT was initially proposed to explain the statistical properties of nuclear spectra [16]. Later this theory was successful applied in the study of the spectra of different complex systems such as disordered systems, quantum chaotic systems, large complex atoms [17]. Recent studies illustrate the usefulness of RMT in understanding the statistical properties of the empirical cross-correlation matrices appearing in the study of multivariate time series of followings: the price fluctuations in the stock market [18], EEG data of brain [19], variation of various atmospheric parameters [20], etc. Recent analysis of complex networks under RMT framework [9, 10, 21, 22] show that various network models and real world network also follow universal GOE statistics. Furthermore localization of eigenvectors have also been used to analyze various structural and dynamical properties of real and model networks [23, 24].

In the following, we introduce spacing distribution and  $\Delta_3$  statistics of random matrices. We denote the eigenvalues of a network by  $\lambda_i$ ,  $i = 1, \dots, N$ , where  $N$  is size of the network and  $\lambda_1 < \lambda_2 < \lambda_3 < \dots < \lambda_N$ . In order to get universal properties of the fluctuations of eigenvalues, people usually unfold the eigenvalues by a transformation  $\bar{\lambda}_i = \bar{N}(\lambda_i)$ , where  $\bar{N}$  is averaged integrated eigenvalue density [16]. Since we do not have any analytical form for  $\bar{N}$ , we numerically unfold the spectrum by polynomial curve fitting (for elaborate discussion on unfolding, see Ref.[16]). After unfolding, average spacings is *unity*, independent of the system. Using the unfolded spectra, we calculate spacings as  $s_i = \bar{\lambda}_{i+1} - \bar{\lambda}_i$ . NNSD is defined as the probability distribution ( $P(s)$ ) of these  $s_i$ 's. In the case of GOE statistics,

$$P(s) = \frac{\pi}{2}s \exp\left(-\frac{\pi s^2}{4}\right) \quad (1)$$

The  $\Delta_3$ -statistic measures the least-square deviation of

the spectral staircase function representing the averaged integrated eigenvalue density  $\overline{N}(\lambda)$  from the best straight line fitting for a finite interval  $L$  of the spectrum, i.e.,

$$\Delta_3(L; x) = \frac{1}{L} \min_{a,b} \int_x^{x+L} [N(\overline{\lambda}) - a\overline{\lambda} - b]^2 d\overline{\lambda} \quad (2)$$

where  $a$  and  $b$  are obtained from a least-square fit. Average over several choices of  $x$  gives the spectral rigidity  $\Delta_3(L)$ . For the GOE case,  $\Delta_3(L)$  depends *logarithmically* on  $L$ , i.e.,

$$\Delta_3(L) \sim \frac{1}{\pi^2} \ln L. \quad (3)$$

The following sub-section explains the properties of eigenvectors of random matrices.

### A. Eigenvector analysis

The distribution of eigenvectors components are studied to obtain system dependent information. Let  $u_l^k$  is the  $l$ th component of  $k$ th eigenvector  $u^k$ . The eigenvector components of a GOE random matrix are Gaussian distributed random variables, for this the distribution of  $r = |u_l^k|^2$ , in the limit of large matrix dimension, is given by Porter-Thomas distribution [25], i.e.,

$$P(r) = \frac{N}{\sqrt{2\pi r}} \exp\left(-\frac{Nr}{2}\right) \quad (4)$$

Shannon entropy for the state whose components are described by the above distribution, would be given by in large  $N$  limit as [25],

$$H_s \sim -N \int_0^\infty r \ln(r) P(r) dr \sim \ln\left(\frac{N}{2}\right). \quad (5)$$

Additionally, inverse participation ratio (IPR) is also considered to study the RMT features of the eigenvectors. The IPR of eigenvector is defined as

$$I^k = \sum_{l=1}^N [u_l^k]^4 \quad (6)$$

where  $u_l^k, l = 1, \dots, N$  are the components of eigenvector  $u^k$ . The meaning of  $I^k$  is illustrated by two limiting cases: (i) a vector with identical components  $u_l^k \equiv 1/\sqrt{N}$  has  $I^k = 1/N$ , whereas (ii) a vector with one component  $u_1^k = 1$  and the remainders zero has  $I^k = 1$ . Thus, the IPR quantifies the reciprocal of the number of eigenvector components that contribute significantly. For a vector with components following distribution (4) has  $I^k \sim 3/N$ .

## III. DATA AND NETWORK CONSTRUCTION

The data-set (GSE5281) was obtained from Gene Expression Omnibus [6]. Liang et al. [2] studied gene expression profiles from laser capture micro dissected neurons in six functionally and anatomically distinct regions

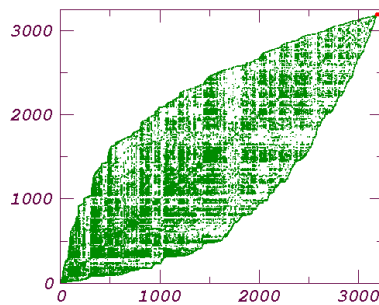


FIG. 1: Adjacency matrix of the largest connected component of the Gene co-expression network with the threshold value of  $\sim 0.89$ . Nodes forming largest connecting cluster are renumbered in the sequential order for a clear visualization.

from clinically and histopathologically normal aged human brains. From these data-sets only 74 normal samples were used to construct the co-expression networks. In the original study the Affymetrix Human Genome U133 Plus 2.0 Array was used. This micro-array contains 54675 oligonucleotids (probesets) representing the expressed human genes for each samples. On the microarray one gene is represented by one or more probesets. Each probeset is built up from 25 mer length oligonucleotides, so called probes [26]. In the present study probesets are the units of observation. For the identification of probesets the Affymetrix IDs were used. The Pearson's product-moment correlation was calculated for each probeset-pair expression level, and those which have value greater than 0.88 are used to construct the gene co-expression network. This network consists of 5000 nodes and 1201480 undirected edges. Nodes represent probeset denoting genes, and edges denote their co-expression levels.

From this weighted network, we construct a sparse binary network as following. We choose the value of threshold being  $r = 0.89$ , if the co-expression strength is greater than  $r$  than the corresponding element in the matrix gets value 1, otherwise 0. Threshold value of  $r = 0.89$  leads to a network with much less number of edges, and results into many disconnected component. Note that choosing the threshold value is a crucial step and different schemes have been proposed to select it [27, 28]. We sort out the nodes and edges forming largest connecting cluster, which is of the size  $N = 3179$  and 46033 connections. The average degree of this network is  $\langle k \rangle \sim 30$ . RMT analysis is done for this biggest component. Fig. 1 shows the adjacency matrix of this component and Fig. 2 is the degree distribution.

## IV. RESULTS

In the following, we present the various RMT results for gene co-expression network constructed above. We calculate the eigenvalues and eigenvectors of the adjacency matrix corresponding to the largest connected net-

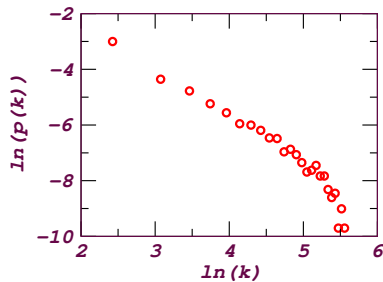


FIG. 2: Degree distribution of the largest connected part of the Gene co-expression network for threshold 0.89.

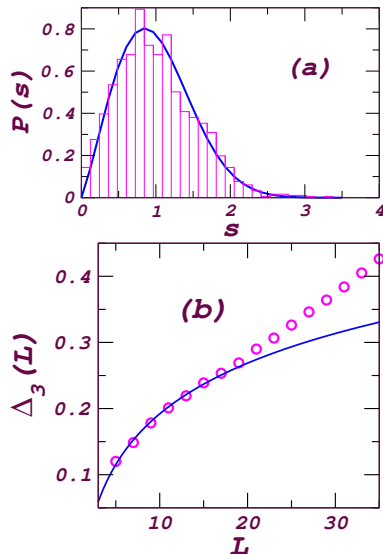


FIG. 3: (Color online) Spacing distribution (a) and  $\Delta_3(L)$  statistics (b) for the eigenvalue spectra of the gene co-expression network. The histogram in (a) corresponds to the numerical values and solid line is GOE prediction (1) of RMT. The circles in (b) are numerical results (2) and the solid curve is GOE prediction (3) of  $\Delta_3$ .

work. Since this is an undirected network, eigenvalues of adjacency matrix are real, and we denote them as  $\lambda_i, i = 1 \dots N$ . Eigenvectors are denoted as  $u^k, k = 1 \dots N$ .

### A. Spacing distribution and $\Delta_3$ analysis

From this spectrum we calculate NNSD  $P(s)$  as described in the section II and  $\Delta_3(L)$  statistic using Eq. (2). Fig. 3(a) shows that NNSD agrees well with the NNSD of GOE matrices (1) with the value of Brody parameter [9, 29]  $\beta \sim 1$ .

Fig. 3(b) plots the  $\Delta_3(L)$  statistics. It can be seen that  $\Delta_3(L)$  statistic agrees well with the GOE statistics up to the value of  $L \sim 25$ , (which is much less than the same for the corresponding random and scale free model networks [9]). According to the RMT, this implies that besides randomness, the network has some specific

features. Note that the points which deviate from GOE statistics ( $L > 20$ ), as shown in the Fig. 3(b) can also be analyzed using deformed GOE statistics as shown in [21].

### B. Eigenvector analysis

Having calculated spacing distribution and  $\Delta_3$  statistics, now we use eigenvector analysis to study the factors responsible for the deviation from RMT. We calculate IPR and entropy for all the eigenvectors. The eigenvectors, whose IPR and entropy deviate from the random matrix predictions, carry the relevant information. The nodes corresponding to the top contributing components of these vectors may be *important nodes* in terms of functionality of the whole network. In the following we present the Eigenvectors analysis results for the gene co-expression network.

Fig. 4(a) shows eigenvalues in the increasing order. Apart from distinguishably seen high eigenvalues towards the end of the spectra, there is a flat part around the zero eigenvalue. Real world networks, in general, are very sparse and are reported to have large number of *zero* eigenvalues [30, 31]. Though for the network we consider here, out of 3179 eigenvalues, only approximately 73 ( $\sim 2.5\%$  of all eigenvalues) are degenerate with the value *zero*. The degeneracy at zero eigenvalue is lesser than many other real world networks [9]. There are nearly 3106 non-degenerate eigenvalues, which could be taken as the effective dimensionality of the network.

We also calculate Shannon entropy for all the eigenvectors using Eq. (5), and compare them with those of the random vectors. Fig. 4(b) shows the entropy as a function of eigen numbers. According to RMT, Shannon entropy of a random vector of dimension  $N = 3106$  is  $\ln(3106/2) \simeq 7.35$ . Furthermore, RMT predicted value for Shannon entropy of a random vector of dimension  $N = 73$  (corresponding to degenerate part) is  $\ln(73/2) \simeq 3.6$ . Based on these calculations, we can divide eigenvalues into three sets; (A) The non-degenerate part that follows RMT. (B) The non-degenerate part, at both ends and at intermediate eigenvalues, which deviate from RMT and expected to contain information about *important nodes* in the network. (C) The degenerate part with *zero* eigenvalue, 1636 to 1708 which fluctuates around RMT predicted value.

Furthermore, we calculate IPR of all the eigenvectors using Eq. (6), and plot in Fig. 4 (c). It shows that IPR of several eigenvalues are localized. For example, vectors corresponding to the 1140 to 1148 eigenvalues have  $I^k \geq 0.1$ , showing that few components contribute more than the other components. Following we enlist some localized eigenvectors corresponding to non-degenerate eigenvalues from set (B):  $u^{1143}$  (with  $I^k \sim 0.5$ ),  $u^{1148}$  (with  $I^k \sim 0.31$ ),  $u^{2257}$  (with  $I^k = 0.25$ ). Some of the localized eigenvectors corresponding to zero eigenvalues are (set (C));  $u^{1636}$  (with  $I^k = 0.1$ ),  $u^{1670}$  and  $u^{1671}$

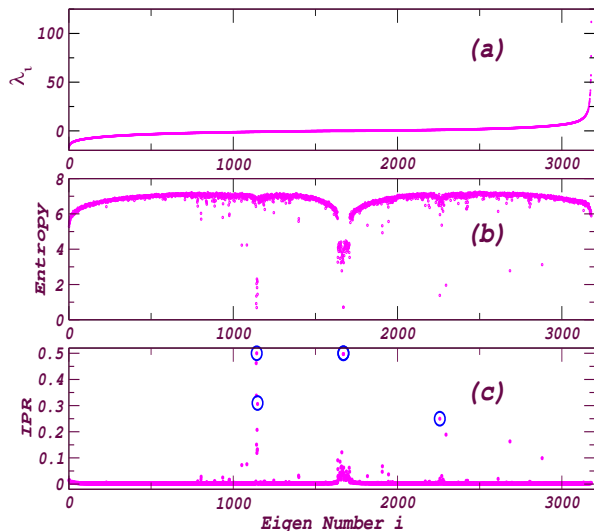


FIG. 4: (Color online) Eigenvalues (a), entropy (b), and IPR (c) as a function of eigen number for the threshold value of 0.89. Open blue circles in (c) correspond to the localized eigenvectors whose top contributing nodes are listed in the Table I

Set B			Set C	
$u^{1143}$	$u^{1148}$	$u^{2257}$	$u^{1670}$	$u^{1671}$
202060_at	227636_at	202916s_at	225921_at	21435x_at
217731s_at	205003_at	226832_at	212635_at	203034s_at
201121s_at	211940x_at	209860s_at	208645s_at	200673_at
221775x_at	224616_at	218175_at	221511x_at	221471_at
229630s_at	222203s_at	221810_at	231896s_at	225950_at

TABLE I: Top five largest contributing nodes in localized eigenvectors for network constructed with the threshold value of 0.89. The nodes are written in the original gene number as given in the datasets [6]

(with  $I^k \sim 0.5$ ). We next analyze the significant contributors of eigenvectors deviating from the RMT predictions. The eigenvector  $u^{1143}$  contains approximately  $1/IPR^{1143} = 20$  significant participants. Table I presents top 5 significant contributors (nodes) corresponding to the localized eigenvector mentioned above. Note that original gene number are written as in the datasets [6]. As shown in the Fig. 2, degree distribution of the connected network analyzed above follows a power law with a fat tail, which means that few nodes are hubs, and carry the whole network. But random matrix analysis of eigenvectors reveals that all the most contributing nodes listed above have rather small degree. They are all almost towards bottom of the power law distribution.

The degree of all the top contributing nodes in the localized eigenvectors are either well below the average degree or around the average degree of the network. Gene, assigned with probeset 202060\_at, (corresponding to the node 2299 in the renumbered network) which is the first top contributing node corresponding to eigenvector  $u^{1143}$ , has a degree 15, the second top contributing node has a

Set B			Set C		
$u^{835}$	$u^{1635}$	$u^{641}$	$u^{1269}$	$u^{1270}$	$u^{1224}$
210338s_at	208666s_at	201121s_at	211733x_at	201494_at	230416_at
210418s_at	224819_at	208667s_at	230869_at	223209s_at	228283_at
202178_at	209460_at	223716_s_at	228045_at	225284_at	238494_at
38398_at	226395_at	224644_at	211733x_at	201494_at	230416_at
213347x_at	201525_at	200626s_at	242317_at	212788x_at	212474_at

TABLE II: Top contributing nodes (genes) in the localized eigenvectors for the threshold value 0.91

degree 17, the third node has a degree 20. Fourth and fifth top contributing nodes have degree 9 each. The top five nodes corresponding to  $u^{1148}$  have degree 21, 14, 7, 17 and 24. Those are corresponding to eigenvector  $u^{2257}$  have degree 1, 1, 6, 3 and 1 respectively. The localized eigenvectors corresponding to set (c) are  $u^{1670}$ ,  $u^{1671}$ , and top five contributing nodes have degree, in sequential order from first to the fifth contributing node (see Table I), 2, 4, 8, 1, 3 and 10, 9, 23, 14, 2 respectively.

Now we change the threshold value to 0.91, this threshold value leads to 25,000 connections in the whole network. This network has largest connected cluster of size 2,439 and number of connections 22546. The average degree of this network is  $\langle k \rangle \sim 20$ . Again we renumber the nodes such that nodes in the connected component take value from 1 to 2,439, and calculate the eigenvalues and eigenvectors of the adjacency matrix corresponding to this largest connected network. From the spectrum  $NNSD$  and  $\Delta_3$  statistics are calculated, and these two show similar GOE statistics as shown in Fig.3 for  $r = 0.89$ .

Fig. 5 plots eigenvalues (a), entropy (b) and IPR (c) as a function of eigen number. Entropy and IPR are calculated using Eq. (5) and (6) respectively. Out of 2,439 eigenvalues, approximately 96 are degenerate with the value zero. It means that there are nearly 2343 non-degenerate eigenvalues, which could be taken as the effective dimensionality of the network. According to RMT, Shannon entropy of a random vector of dimension  $N = 2343$  is  $\ln(2343/2) \simeq 7.0$ . On the other hand, RMT predicted value for Shannon entropy for degenerate eigenvectors is  $\ln(96/2) \simeq 3.9$ . Based on these calculations, again we can divide eigenvalues in three sets (A), (B) and (C). Localized eigenvectors corresponding to non-degenerate part are:  $u^{835}$  (IPR=0.41),  $u^{1635}$  (IPR=0.3),  $u^{641}$  (IPR=0.3),  $u^{840}$  and  $u^{841}$  (with  $\lambda = 1$ , IPR=0.195 and 0.24) Localized eigenstates corresponding to zero eigenvalues (set (c)) are:  $u^{1269}$  (IPR=0.38),  $u^{1270}$  (IPR=0.37),  $u^{1224}$  (IPR=0.28). Significant contributors in localized eigenvectors are written in Table II.

The degree distribution of the largest component at this threshold follows a power law as well, revealing the scalefree nature of this component. Increasing threshold preserves scalefree property of the network. Some nodes are hubs which carry the whole network and enjoy the structural importance. Again we find that the top contributing nodes are not the ones with very high degree. For two different threshold values Tables I and

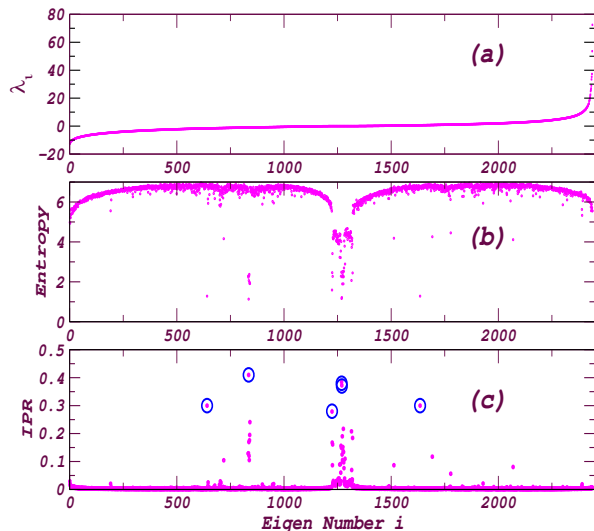


FIG. 5: (Color online) Same as Fig.4 but for threshold value of 0.91. Open blue circles correspond to localized eigenvectors whose top five contributing nodes are presented in the Table II.

II show the largest contributing co-expressing genes in the corresponding localized eigenvectors. We find that choosing threshold is very important for the analysis of Gene co-expression networks, as we can see that top five largest contributing nodes differ entirely (except one) as threshold value is changed. This suggests that, though the gross structure of whole network (Fig. 1) and scale-free property, remains unchanged, value of threshold has a strong effect on the network leading to entirely different sets (except few) of largest contributing nodes for two different threshold values. Appendix enlists the genenames corresponding to the probesets identifiers as given in I and II.

## V. CONCLUSIONS AND DISCUSSIONS

Using RMT, we have analyzed gene co-expression network constructed by applying two different threshold values to the data obtained from six brain regions that are metabolically relevant to Alzheimer's disease [6]. The NNSD of adjacency matrix of the largest connecting component of the network follows universal GOE statistics (with  $\beta \sim 1$ ). This universality adds one more feature, based on the spectral correlations, to the gene co-expression network which is common with different model networks [9] proposed to capture various structural properties of real world networks.

The NNSD gives information about the short range correlations among the eigenvalues. To probe the long range correlations we have studied spectral rigidity via  $\Delta_3(L)$  statistics. This analysis shows that the gene co-expression network considered here follows RMT prediction of GOE for very long range of  $L$ . Beyond this value

of  $L$  deviation in the spectral rigidity is seen, indicating a possible breakdown of universality. This means the network under consideration has *sufficient* randomness which may due to *robustness of the systems*, with regularity which may be to *perform some functional task*. Mixture of random connections and regular structure have been emphasized at various places, for instance information processing in the brain is considered to be random connections among different modular structure [32].

Deviation from the universal RMT predictions identify system-specific, non-random properties of system under consideration, might provide clues about important interactions. To extract these system dependent information we have performed eigenvector analysis. This analysis reveals that there are some eigenvectors which are highly localized. The component  $l$  of a given eigenvector relates to the contribution of node (corresponding gene)  $l$  to that eigenvector. Hence, the distribution of the components contains information about the number of genes contributing to a specific eigenvector. Inverse participation ratio IPR, as defined in Eq. (6), distinguishes between one eigenvector with approximately equal components and another with a small number of large components. According to the RMT predictions, the largest contributing nodes (genes) in the localized eigenvectors may have important function, or important functional relations among them.

The largest connected component is scale-free indicating the structural importance of few nodes (hubs). Eigenvector analysis shows that top contributing nodes in the localized eigenvectors have relatively low degrees. Note that genes which are hubs or those which connect different communities are also important, as shown by several earlier studies in the network framework [5, 13], but the aim of the present work is look for the important genes beyond these structural measures. Changing the value of threshold, while keeping the scale-free structure of network same, has drastic impact on the localization property of eigenvectors. All most all the top contributing nodes differ for two different threshold value, indicating impact on the global properties of the underlying network.

Last, we discuss here the importance of the analysis and future implications of the results presented in the paper. Several studies have shown that the development of multi-target drugs might give better results than the traditional methods targeting a single protein. Single target-design might not always give satisfactory results, as there might be a backup system, which replaces the function of the inhibited target protein. By using multi-target drugs one can decrease the functionality of entire protein cascades producing more effective results. For example, studies have shown that aging is strongly linked with age-related diseases, and they share a common signaling network. Signaling hubs of the age-related protein-protein interaction subnetwork may be good candidates for age-related drug-targets. Multi-target drugs attacking hubs of the protein-protein interaction net-

Probeset	Gene name
202060_at	Ctr9, PafI/RNA polymerase II
227636_at	-
202916s_at	family with sequence similarity 20, member B
225921_at	ninein (GSK3B interacting protein)
214351x_at	ribosomal protein L13
217731s_at	integral membrane protein 2B
205003_at	dedicator of cytokinesis 4
226832_at	-
212635_at	transportin 1
203034s_at	ribosomal protein L27a
201121s_at	progesterone receptor membrane component 1
211940x_at	-
209860s_at	annexin A7
208645s_at	ribosomal protein S14
200673_at	lysosomal protein transmembrane 4 alpha
221775x_at	ribosomal protein L22
224616_at	dynein, cytoplasmic 1
218175_at	coiled-coil domain containing 92
221511x_at	cell cycle progression 1
221471_at	serine incorporator 3
229630s_at	Wilms tumor 1 associated protein
222203s_at	retinol dehydrogenase 14
221810_at	RAB15, member RAS oncogene family
231896s_at	density-regulated protein
225950_at	-

TABLE III: Genenames corresponding to the probesets for the threshold value 0.89

work, 'hub-links' (links connecting hubs), bridges (inter-

modular links having high 'betweenness centrality') or nodes in the overlap of numerous network modules, might give better results [33, 34]. Similarly, targeting genes corresponding to the largest contributing nodes in localized eigenvectors may lead to important effect as well. Future investigations are sought in order to know the functionality of these genes corresponding to the top contributing nodes in the localized eigenvectors, which could be then used for such multi-target drug designs.

## Appendix

Tables III and IV correspond to probesets identifiers from tables I and II respectively. First column of these tables are probeset identifiers (Affymetric ID) and second column dictates the corresponding genenames. However, the he function of some transcripts is not known yet, and some of them has no gene name. The value '-' in the gene name column indicates that information is not available. Note that there are many reasons for probesets without detailed annotation. We know the sequence on microarray for each probesets. On the chip we get all expressed genes, but we do not have secure info for all the gene functions. As the knowledge is growing with the latest available technologies, this gap is decreasing with time. One sure information for the probeset is the Affymetric ID as given in the table I and II [26].

- 
- [1] S. Maslov and K. Sneppen, *Science* **296**, 910 (2002); H. Hishigaki *et. al. Yeast* **18** 523 (2001).
- [2] W. S. Liang *et. al.*, *Physiol Genomics* **28** (3), 311 (2007); W. S. Liang *et. al.*, *Proc Natl Acad Sci U S A*, **105** 4441 (2008).
- [3] A.-L. Barabási and R. Albert, *Science* **286**, 509 (1999).
- [4] S. H. Strogatz, *Nature* **410**, 268 (2001).
- [5] R. Albert and A.-L. Barabási, *Rev. Mod. Phys.* **74**, 47 (2002) ; S. Boccaletti *et. al.*, *Phys. Rep.* **424**, 175 (2006).
- [6] <http://www.ncbi.nlm.nih.gov/geo/>
- [7] A. J. Butte and I. S. Kohane, *Proc AMIA Symp.* **25** 711 (1999).
- [8] For a recent review see: Diogo F. T. Veiga, B. Dutta and Gábor Balázsi, *Mol. BioSyst.* **6**, 469 (2010).
- [9] J. N. Bandyopadhyay and S. Jalan, *Phys. Rev. E*, **76**, 026109 (2007); S. Jalan and J. N. Bandyopadhyay, *Phys. Rev. E*, **76**, 046107 (2007).
- [10] R. Potestio, F. Caccioli and P. Vivo, *Phys. Rev. Lett.* **103** 268101 (2009).
- [11] M. Girvan and M. E. J. Newman, *Proc. Natl. Acad. Sci. USA* **99**, 7821 (2002); M. E. J. Newman, *Social Networks* **27**, 39 (2005); M. E. J. Newman, *Proc. Natl. Acad. Sci. USA* **103**, 8577 (2006)
- [12] M. J. Krawczyk, *Phys. Rev. E* **77** 065701 (R) (2008); G. Palla *et. al.* *Nature* **435**, 814 (2005); M. E. J. Newman, *Phys. Rev. E* **70**, 056131 (2004); A. Arenas, A. Fernandez and S. Gomez, *New J. Phys.* **10** 053039 (2008).
- [13] R. Guimerá and L. A. N. Amaral, *Nature* **433**, 895 (2005).
- [14] Béla Bollobás, *Random Graphs* (Second edition, Cambridge Univ. Press, 2001).
- [15] D. J. Watts and S. H. Strogatz, *Nature* **440**, 393 (1998).
- [16] M. L. Mehta, *Random Matrices*, 2nd ed. (Academic Press, New York, 1991).
- [17] T. Guhr *et. al.*, *Phys. Rep.* **299**, 189 (1998).
- [18] V. Pleron *et. al.*, *Phys. Rev. Lett.* **83**, 1471 (1999).
- [19] P. Seba, *Phys. Rev. Lett.* **91**, 198104 (2003).
- [20] M. S. Santhanam and P. K. Patra, *Phys. Rev. E* **64**, 016102 (2001).
- [21] J. X. de Carvalho, S. Jalan, M. S. Hussein *Phys. Rev. E* **79** 056222 (2009).
- [22] S. Jalan, *Phys. Rev. E* **80**, 046101 (2009).
- [23] P. N. McGraw and M. Menzinger, *Phys. Rev. E* **77** 031102 (2008)
- [24] G. Zhu, H. Yang, C. Yin and B. Li, *Phys. Rev. E* **77**, 066113 (2008).
- [25] K. Zyczkowski, chapter in *Quantum chaos*, edited by H. A. Cerdeira, R. Ramaswami, M. C. Gutzwiller and G. Casati, (World Scientific, 1991).
- [26] H. Göhlmann and W. Talloen, *Gene Expression Studies Using Affymetrix Microarrays*, (Chapman and Hall/CRC 2009).
- [27] X. Zhou, M. C. Kao and W. H. Wong, *Proc. Natl. Acad. Sci. USA* **99**, 12783 (2002); D. Smet, F. Mathys *et. al.*, *Bioinformatics* **18** 735 (2002).
- [28] F. Luo *et. al.*, *Bioinformatics* **8** 299 (2007).

Probeset	Gene name
210338s_at	heat shock 70kDa protein 8
208666s_at	suppression of tumorigenicity 13
201121s_at	progesterone receptor membrane component 1
211733x_at	sterol carrier protein 2
201494_at	prolylcarboxypeptidase
230416_at	-
210418s_at	isocitrate dehydrogenase 3 (NAD+)
224819_at	transcription elongation factor A (SII)
208667s_at	suppression of tumorigenicity 13
230869_at	family with sequence similarity 155
223209s_at	selenoprotein S
228283_at	COX assembly mitochondrial protein homolog
202178_at	protein kinase C, zeta
209460_at	4-aminobutyrate aminotransferase
223716s_at	zinc finger, RAN-binding domain
228045_at	-
225284_at	DnaJ (Hsp40) homolog, subfamily C
238494_at	TNF receptor-associated factor 3
38398_at	MAP-kinase activating death domain
226395_at	hook homolog 3 (Drosophila)
224644_at	-
211733x_at	sterol carrier protein 2
201494_at	prolylcarboxypeptidase
230416_at	-
213347x_at	ribosomal protein S4, X-linked
201535_at	ubiquitin-like 3
200626s_at	martin 3
242317_at	HIG1 hypoxia inducible domain family
212788x_at	ferritin, light polypeptide
212474_at	AVL9 homolog (S. cerevisiae)

TABLE IV: Genenames corresponding to the probesets for the threshold value 0.91

- [29] T. A. Brody, *Lett. Nuovo Cimento* **7**, 482 (1973).
- [30] S. N. Dorogovtsev *et. al.*, *Phys. Rev. E* **68**, 046109 (2003).
- [31] M. A. M. de Aguiar and Y. Bar-Yam, *Phys. Rev. E* **71**, 016106 (2005).
- [32] J. D. Cohen and F. Tong, *Science*, **293**, 2405 (2001).
- [33] P. Csemely, V. Ágoston and S. Pongor, *Trends in Pharmacol Sci* **26** 178 (2005); T. Korcsmáros *et. al.*, *Exp Op Drug Discovery* **2** 1 (2007); GR Zimmermann, Lehár and CT Keith, *Drug Discovery Today* **12** 34 (2007); M. Antal, C. Böde and P. Csemely, *Curr Prot Pept Sci* **10** 161 (2009); P. Csemely, *Trends Biochem Sci* **33** 569 (2008).
- [34] G. I. Simkó, *et. al.* *Genome Medicine* **1**, 90 (2009).