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

Sarika Jalan,1 Norbert Solymosi,2 Gábor Vattay,2 and Baowen Li1,3

1Department of Physics and Centre for Computational Science and Engineering, National University of Singapore, 117456, Republic of Singapore
2Department of the Physics of Complex Systems, Eötvös University, H-1117 Pázmány Péter tér 1/A, Budapest, Hungary
3NUS Graduate School for Integrative Sciences and Engineering, 117456, 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 form 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 subsection 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 subsection 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 3 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 \[^{54}\]. 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 \[^{3, 10, 21, 22}\] 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 \(\lambda_i, i = 1, \ldots, N\), where \(N\) is size of the network and \(\lambda_1 < \lambda_2 < \lambda_3 < \cdots < \lambda_N\). In order to get universal properties of the fluctuations of eigenvalues, people usually unfold the eigenvalues by a transformation \(\overline{\lambda}_i = N(\lambda_i)\), where \(N\) is averaged integrated eigenvalue density \[^{10}\]. Since we do not have any analytical form for \(N\), we numerically unfold the spectrum by polynomial curve fitting (for elaborate discussion on unfolding, see Ref. \[^{14}\]). After unfolding, average spacings is unity, independent of the system. Using the unfolded spectra, we calculate spacings as \(s_i = \overline{\lambda}_{i+1} - \overline{\lambda}_i\). NNSD is defined as the probability distribution \((P(s))\) of these \(s_i\)’s. In the case of GOE statistics,

\[
P(s) = \frac{\pi}{2} s \exp\left(-\frac{\pi s^2}{4}\right)
\]

The \(\Delta_3\)-statistic measures the least-square deviation of
the spectral staircase function representing the averaged integrated eigenvalue density $N(\lambda)$ from the best straight line fitting for a finite interval $L$ of the spectrum, i.e.,
\[
\Delta_3(L;x) = \frac{1}{L} \min_{a,b} \int_x^{x+L} \left[ N(\lambda) - a\lambda - b \right]^2 d\lambda
\]
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.,
\[
\Delta_3(L) \sim \frac{1}{\pi^2} \ln L.
\]

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 $u^k_l$ 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 $r = |u^k_l|^2$, in the limit of large matrix dimension, is given by Porter-Thomas distribution \[2\], i.e.,
\[
P(r) = \frac{N}{\sqrt{2\pi r}} \exp \left( -\frac{Nr}{2} \right)
\]
Shannon entropy for the state whose components are described by the above distribution, would be given by in large $N$ limit as \[2\],
\[
H_s \sim -N \int_0^{\infty} r \ln P(r) dr \sim \ln \left( \frac{N}{2} \right).
\]
Additionally, inverse participation ratio (IPR) is also considered to study the RMT features of the eigenvectors. The IPR of eigenvector is defined as
\[
I^k = \sum_{l=1}^{N} |u^k_l|^4
\]
where $u^k_l, l = 1, \ldots, N$ are the components of eigenvector $u^k$. The meaning of $I^k$ is illustrated by two limiting cases: (i) a vector with identical components $u^k_l \equiv 1/\sqrt{N}$ has $I^k = 1/N$, whereas (ii) a vector with one component $u^k_l = 1$ 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 \[2\] 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 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 \[20\]. 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 $<k> \sim 30$. 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-
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 \(\text{[21]}\).

### 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. \(\text{[3]}\), 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 \(\ln(3106/2) \sim 7.35\). Furthermore, RMT predicted value for Shannon entropy of a random vector of dimension \(N = 73\) (corresponding to degenerate part) is \(\ln(73/2) \sim 3.6\). 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. \(\text{[6]}\), 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 \(I_k \approx 0.5\)), \(u^{1148}\) (with \(I_k \approx 0.31\)), \(u^{2257}\) (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 \(u^{1671}\)
Figure 3A shows the degree distribution of the network, revealing a power law distribution with a fat tail. The top five contributing nodes have rather small degree. They are all associated with localized eigenvectors, which are either well below the average degree of the network or around the average degree of the network. Gene expression data shows that these are not the top degree nodes of the network.

Table I presents the 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 shown in the dataset [6]. The eigenvector $u_{1143}$ contains approximately 2343 gene contributions, which could be taken as the effective dimensionality of the network. According to RMT, Shannon entropy of a random vector of dimension $N$ is $\ln(N)$, and the value $\frac{\ln(2343)}{N}$ is 0.38. Significant contributors of eigenvectors deviating from the RMT predictions lead to $\Delta_k = 1, 2, 8, 1, 3$ and 10, 9, 23, 14, 2 respectively.

Table II presents top 5 significant contributors (nodes) corresponding to the localized eigenvectors. The top contributing nodes are not the ones with very high degree. For two different threshold values Tables I and II show similar GOE statistics as shown in Figure 3B for $r = 0.89$.

Fig. 3B 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 2343 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 $\ln(2343)$, and the value $\frac{\ln(2343)}{N}$ is 0.41.

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

| Set B | | Set C | |
|-------|---|---|---|
| $u_{1143}$ | $u_{1148}$ | $u_{1143}$ | $u_{1148}$ |
| 220760 | 227763 | 217763 | 217315 |
| 211238 | 217550 | 217550 | 217550 |
| 201218 | 201218 | 201218 | 201218 |
| 225921 | 214356 | 214356 | 214356 |
| 214356 | 214356 | 214356 | 214356 |

The degree distribution of the largest component at this threshold follows a power law as well, revealing the scale-free 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 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. I) 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 [5]. 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 [8] 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-
work, 'hub-links' (links connecting hubs), bridges (inter-

| Probeset | Gene name |
|----------|-----------|
| 227636_at | Ctr9, Pafl/RNA polymerase II |
| 202916_at | family with sequence similarity 20, member B |
| 214351_at | ribosomal protein L13 |
| 217731_at | integral membrane protein 2B |
| 205003_at | dedicator of cytokinesis 4 |
| 226852_at | n ninein (GSK3B interacting protein) |
| 212635_at | ribosomal protein L27a |
| 209860_at | RAB15, member RAS onocogene family |
| 208645_at | lysosomal protein transmembrane 4 alpha |
| 221775_at | ribosomal protein L22 |
| 224616_at | dynein, cytoplasmic 1 |
| 218175_at | coiled-coil domain containing 92 |
| 221511_at | cell cycle progression 1 |
| 221471_at | serine incorporator 3 |
| 229360_at | Wilms tumor 1 associated protein |
| 222203_at | retinol dehydrogenase 14 |
| 221810_at | RAB15, member RAS oncogene family |
| 231896_at | density-regulated protein |

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

Appendix

Tables III and IV correspond to probesets identifiers from tables I and II respectively. First column of these tables are probeset identifiers (Affymetrix 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 probeset. 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 Affymetrix ID as given in the table I and II [20].
| Probeset   | Gene name                                                   |
|-----------|-------------------------------------------------------------|
| 210338s_at| heat shock 70kDa protein 8                                  |
| 208660s_at| suppression of tumorigenicity 13                            |
| 201121s_at| progesterone receptor membrane component 1                  |
| 211733x_at| sterol carrier protein 2                                   |
| 201494s_at| prolylcarboxypeptidase                                      |
| 230416_at | -                                                           |
| 210418s_at| isocitrate dehydrogenase 3 (NAD+)                           |
| 224819_at | transcription elongation factor A (SH)                      |
| 208667s_at| suppression of tumorigenicity 13                            |
| 230869_at | family with sequence similarity 155                        |
| 223209s_at| selenoprotein S                                             |
| 228283_at | COX assembly mitochondrial protein homolog                  |
| 201780_at | 4-aminobutyrate aminotransferase                            |
| 223716s_at| zinc finger, RAN-binding domain                             |
| 228045_at | -                                                           |
| 225241_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                                   |
| 201494s_at| prolylcarboxypeptidase                                      |
| 230416_at | -                                                           |
| 213347x_at| ribosomal protein S4, X-linked                              |
| 201535_at | ubiquitin-like 3                                            |
| 200626s_at| -                                                           |
| 242317_at | HIG1 hypoxia inducible domain family                        |
| 21278s_at | ferritin, light polypeptide                                 |
| 212474_at | AVI9 homolog (S. cerevisiae)                                |

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