Disordered beta thinned ensemble with applications

Rongrong Xie,[1,2,*] Shengfeng Deng,[3,†] Weibing Deng,[1,‡] and Mauricio P. Pato[4,§]

[1] *Key Laboratory of Quark and Lepton Physics (MOE) and Institute of Particle Physics, Central China Normal University, Wuhan 430079, China*

[2] *Abdus Salam International Centre for Theoretical Physics (ICTP), Trieste 34151, Italy*

[3] *Institute of Technical Physics and Materials Science, Center for Energy Research, Budapest 1121, Hungary*

[4] *Instítuto de Física, Universidade de São Paulo Caixa Postal 66318, 05314-970 São Paulo, S.P., Brazil*

(Received 7 August 2021; accepted 17 November 2021; published 30 November 2021)

It recently has been found that methods of the statistical theories of spectra can be a useful tool in the analysis of spectra far from levels of Hamiltonian systems. The purpose of the present study is to deepen this kind of approach by performing a more comprehensive spectral analysis that measures both the local- and long-range statistics. We have found that, as a common feature, spectra of this kind can exhibit a situation in which local statistics are relatively quenched while the long-range ones show large fluctuations. By combining three extensions of the standard random matrix theory (RMT) and considering long spectra, we demonstrate that this phenomenon occurs when disorder and level incompleteness are introduced in an RMT spectrum. Consequently, the long-range statistics follow Taylor's law, suggesting the presence of a fluctuation scaling (FS) mechanism in this kind of spectra. Applications of the combined ensemble are then presented for spectra originate from several very diverse areas, including complex networks, COVID-19 time series, and quantitative linguistics, which demonstrate that short- and long-range statistics reflect the rigid and elastic characteristics of a given spectrum, respectively. These observations may shed some light on spectral data classification.

DOI: 10.1103/PhysRevE.104.054144

I. INTRODUCTION

In the late 1950s, Wigner proposed an ensemble of random matrices as a tool to describe statistical properties in the dense region of the spectra of many-body systems. During the 1960s, the formalism was then fully developed by Wigner, Mehta, and mainly Dyson in a series of seminal papers (see Ref. [1] for a review with preprints). Random matrix theory (RMT) could then be considered as a well-established theory with a body of statistical measures that became known as the Wigner-Dyson statistics [2]. This standard RMT is constituted by three classes of ensembles of Hermitian Gaussian matrices whose elements are real, for the Gaussian orthogonal ensemble (GOE), complex, for the Gaussian unitary ensemble (GUE), and quaternion, for the Gaussian symplectic ensemble (GSE). These classes are labeled by the Dyson index β that gives the number of degree of freedom of the matrix elements, 1, 2, and 4, respectively. A great boost in applications came at the beginning of the 1980s when the link to the manifestations of chaos in quantum systems was set by the Bohigas-Giannoni-Scmit conjecture that states the equivalence between quantum chaos and RMT [3], while, in contrast, regular systems would have the uncorrelated Poisson statistics. The Wigner-Dyson statistics contains two kinds of measures: short-range ones that probe local correlations, for which the most used measurement is the nearest-neighbor distribution (NND), and the long-range ones that probe correlations along the spectrum, for which the number variance (NV) is the most employed quantity (see Appendix A for its connection with the two-point correlation function). Henceforth, the NND and the NV will be the main quantities of interest in this work.

Generally speaking, spectra are points on a line, and their existence are not restricted to Hamiltonian systems as had been extensively treated in classical RMT. For instance, the sequence of prime numbers forms a spectrum. Moreover, spectra can be also constituted in situations in which line and points are considered in a general way [4]. For example, as studied in Ref. [5], if punctuation is removed from a text, then the text becomes a spectrum of blanks. In this case, the NND is the distribution of the distances between neighboring blanks, measured by the number of letters, and it thus gives the distribution of the length of words. The same idea can be applied to Chinese characters by considering that for characters strokes play the same role as letters do for words. For polymers, proteins and DNA are sequences of letters, and by just taking out a given letter, then a spectrum is defined [6]. In addition to the above areas of nonstandard spectra, in this work, we also extend the analysis to spectra extracted from complex networks and from correlation matrices of COVID-19 time series.

Despite the success RMT had enjoyed for the understanding of spectra generated from physical systems in the more than half century of its existence (see the review paper [7] with more than 800 references), spectra far from levels of Hamiltonian systems usually manifest properties, such as large fluctuations in long-range correlations, that cannot be fully accounted for by classical RMT. Therefore, extensions

[*] Corresponding author: emilyxierr@gmail.com

[†] Corresponding author: gitsteven@gmail.com

[‡] Corresponding author: wdeng@mail.ccnu.edu.cn

[§] Corresponding author: mpato@if.usp.br of the original formalism have also been proposed in order to enlarge the range of applications of the random matrix approach. Since complex empirical systems usually possess less symmetries than most Hamiltonian systems, and they are often not closed but subject to external noises, here we are particularly interested in taking into account the effects of three relatively recent RMT generalizations. The first one is the so-called beta ensemble [8] made of tridiagonal matrices, in which the value of the Dyson index β can assume any real positive value in contrast to the values of 1, 2, and 4 it has in the Gaussian ensembles, so that the symmetry mandate can be largely eased. The second generalization, which is crucial for accounting for the large fluctuation in long-range statistics, has been called the disordered ensemble, in which an external source of randomness is introduced that operates concomitantly with the internal Gaussian ones [9]. Finally, the third one is the ensemble constructed by randomly removing a fraction of the eigenvalues from a given spectrum, namely, the thinned ensemble [10,11]. The removal of levels decreases the correlations among the remaining ones so that they show statistics intermediate between Wigner-Dyson and Poisson. In our scheme, it enters as a tool to take into account the incompleteness of the spectrum, that is, to treat the effect of missing levels. As will be seen later, the necessity of this ingredient is actually related to the fluctuation scaling phenomenon. It is our motivation to show that by combining these three RMT extensions, a random matrix model is constructed that can capture special features found in the analysis of spectra that are far from levels of physical systems.

The need for the three RMT extensions is then immediately justified by the data. The three RMT classes, GOE, GUE, and GSE, are associated with symmetries of the physical systems which play no role in the present case turning necessary to consider arbitrary real values of β. We found that the NVs show a parabolic increase for large interval, a super-Poissonian behavior which is a characteristic of disordered ensembles [9] and is characterized by a larger variance than the Poisson distribution. In some cases, the NND and the NV can be fitted only by resorting to the intermediate statistics of the thinned ensemble whose data is a signature of defects. However, the most important feature emerges when the NND and NV data are confronted so that they can behave independently. This means that we are dealing with a special kind of spectra that shows a certain degree of complexity and elasticity, which are typical characteristics of soft matter [12]. In this respect, the local- and long-range correlations, that is the NND and NV, correspondingly are manifestations of the rigid and the elastic aspects of a given spectrum, hence the NND and NV are bound to give some advantages to spectral data classification.

The analogy of spectra to the state of matter is known in the theory of spectra [13]. Actually, the configuration exhibited by the RMT eigenvalues as a consequence of the repulsion among them has already been considered as related to a crystal lattice structure. The picture is that the eigenvalues behave as a picket fence in which they vibrate around fixed points [14]. Here the analogy is extended to show that a new phase appears when spectra are subjected to external sources of randomness.

This paper is then organized as follows. In the next section, we discuss in detail the three RMT extensions. In Appendix A we show how the external source modifies the number variance, and in Appendix B we derive the basic asymptotic expressions for very long spectra. Section III applies the formalism in the analysis of spectra extracted from three different areas regarding complex networks, COVID-19 time series, and literary texts. Finally, we conclude this work in the last section.

II. DISORDERED BETA THINNED ENSEMBLE

Before combining the three RMT extensions, namely, the beta, the disordered, and the thinned ensembles, we give a summary of their main points. One should keep in mind that our main focus would be the short- and long-range statistics, i.e., the NND and the NV, of each ensemble. The main points are then illustrated in Figs. 1 and 2, which would be further employed to interpret the behavior of empirical spectra. Though Sec. II B 1 is useful for obtaining the matrix elements of the disordered ensemble, it is not very relevant to the subsequent discussions, and one may skim through it to just get a rough idea.

A. The beta ensemble

The beta ensemble consists of a family of Hermitian tridiagonal matrices

$$H_\beta = \begin{pmatrix} a_1 & b_1 & & & & \\ b_1 & a_2 & b_2 & & & \\ & \cdot & \cdot & \cdot & & \\ & & \cdot & \cdot & \cdot & \\ & & & \cdot & \cdot & \cdot \\ & & & b_{n-2} & a_{n-1} & b_{n-1} \\ & & & & b_{n-1} & a_n \end{pmatrix}, \tag{1}$$

where the diagonal elements a_i are normally distributed, namely, $N(0, 1)$, while the b_i are distributed according to

$$f_\nu(y) = \frac{2\exp(-y^2)y^{\nu-1}}{\Gamma(\nu/2)}, \tag{2}$$

with $\nu = (n-i)\beta$ and β is a real positive parameter. From this definition, it is found that the joint density distribution of the eigenvalues is given by [8]

$$P_\beta(E_1, E_2, \ldots, E_n) = C_n^\beta \exp\left(-\frac{1}{2}\sum_{k=1}^{n} E_k^2\right)\prod_{j>i}|E_j - E_i|^\beta. \tag{3}$$

As stated in the introduction, the above equation shows that for $\beta = 1, 2,$ and 4 their eigenvalues share all the statistical properties of the RMT Gaussian classes of matrices, that is, they have Wigner-Dyson statistics. For arbitrary values of β, analytic expressions are not yet fully derived. Asymptotically, it can be shown that when $\beta \to 0$ with the matrix size kept fixed, the matrix becomes diagonal, and in this case, the density of eigenvalues is Gaussian and the Poisson statistics follows. On the other hand, when $\beta \to \infty$ fluctuations are suppressed.

For the product $n\beta \gg 1$, the asymptotic density of eigenvalues is the semicircle law

$$\rho_\beta(E) = \frac{1}{\pi\beta}\sqrt{2n\beta - E^2}, \tag{4}$$

and the NND is well described by the Wigner surmise

$$p_\beta(s) = \frac{2B^{\beta+1}s^\beta}{\Gamma[(\beta+1)/2]}\exp[-(Bs)^2], \tag{5}$$

where $B = \Gamma(\frac{\beta+2}{2})/\Gamma(\frac{\beta+1}{2})$ and $s = 2N_\beta(E/2)$ with

$$\begin{aligned} N_\beta(E) &= \int_0^x \rho_\beta(E')\,dE' \\ &= \frac{n}{\pi}\left(\arcsin\frac{E}{\sqrt{2n\beta}} + \frac{E}{\sqrt{2n\beta}}\sqrt{1-\frac{E^2}{2n\beta}}\right). \end{aligned} \tag{6}$$

Equation (5) defines a one-parameter family of functions, whose parameter β can be determined by fitting the data as had been done in Ref. [5]. In addition, the fluctuations in the number of eigenvalues in the interval $(-\frac{\theta}{2}, \frac{\theta}{2})$ is characterized by the number variance $\langle n^2\rangle - \langle n\rangle^2$. For the GOE with $\beta = 1$, we have the number variance expressed in terms of the unfolded interval length $L = \langle n\rangle = 2N_\beta(\theta/2)$ as [2,15]

$$\Sigma^2_{\rm GOE} = \langle n^2\rangle - \langle n\rangle^2 = \frac{2}{\pi^2}\left[\ln(2\pi L) + 1 + \gamma - \frac{\pi^2}{8}\right], \tag{7}$$

where γ denotes the Euler gamma constant.

B. The disordered beta ensemble

Disordered ensembles were defined in [9] by considering random matrices $H_D(\xi)$ which are obtained from a random matrix H of a given ensemble by the relation

$$H_D(\xi) = \frac{\bar\xi}{\xi}H, \tag{8}$$

where ξ is a positive random variable sorted from a distribution $w(\xi)$ with first moment $\bar\xi$. This scheme emerged from a generalization of an ensemble generated, by two independent groups, via using the maximum entropy principle based on the Tsallis entropy [16,17]. It has been labeled "disordered" as an external source of randomness, which is represented by the random parameter ξ, is imposed on the internal randomness. The amplitude of the disorder is then controlled by the localization of the distribution $w(\xi)$ around its average; cf. Appendix A.

From this relation, it follows that, in the case of the above matrix H_β, the joint density distribution of the $2n-1$ matrix elements of the disordered matrix H_D is given by

$$P(H_D) = \int_0^\infty d\xi\, w(\xi)\left(\frac{\xi}{\bar\xi}\right)^{n-1/2} P_\beta\left(\sqrt{\frac{\xi}{\bar\xi}}H_D\right) \tag{9}$$

that, explicitly, gives

$$\begin{aligned} P(H_D) = C_n^\beta \int_0^\infty d\xi\, w(\xi)\left(\frac{\xi}{\bar\xi}\right)^{\gamma_n} \exp\left(-\frac{\xi}{2\bar\xi}\sum_{i=1}^n d_i^2\right) \\ \times \prod_{j=1}^{n-1} c_j^{j\beta-1}\exp\left(-\frac{\xi c_j^2}{\bar\xi}\right), \end{aligned} \tag{10}$$

where $\gamma_n = \frac{n}{2} + \frac{n(n-1)}{4}\beta$ and the diagonal elements are denoted by the letter d and subdiagonal ones by the letter c. Choosing the distribution $w(\xi)$ to be given by (see Refs. [18,19] for other choices)

$$w(\xi) = \frac{1}{\Gamma(\bar\xi)}\exp(-\xi)\xi^{\bar\xi-1}, \tag{11}$$

which follows from the Tsallis entropy formalism, the integral in ξ can be performed, and the expression

$$\begin{aligned} P(H_D) = \frac{2^{n-1}}{(2\pi)^{n/2}\prod_{j=1}^{n-1}\Gamma(j\beta/2)}\left(\frac{1}{\bar\xi}\right)^{\gamma_n} \\ \times \frac{\Gamma(\bar\xi+\gamma_n)\prod_{j=1}^{n-1}c_j^{j\beta-1}}{\Gamma(\bar\xi)\left(1+\frac{1}{2\bar\xi}\sum_{i=1}^n d_i^2 + \frac{1}{\bar\xi}\sum_{j=1}^{n-1}c_j^2\right)^{\bar\xi+\gamma_n}} \end{aligned} \tag{12}$$

is obtained. As a consequence, explicitly the matrix elements are not independent anymore but correlated. As will be seen later, this correlation among the matrix elements further induces a *positive correlation* within eigenvalues so that they may show aggregated behavior at longer ranges, leading to large fluctuations in the long-range statistics. This is in stark contrast to the level-repulsion behavior in the pure beta ensemble. Therefore, the disorder effect will constitute a crucial factor for understanding the large fluctuations manifested in long-range statistics of empirical spectra. We now show how this expression can be used to generate first the matrix elements and after the eigenvalues.

1. Matrix elements

Although Eq. (8), in principle, can be used to generate the disordered matrices, it is also instructive to be able to obtain the elements of the matrices by taking into account the correlations among them. This can be done, for instance, by sorting them in the sequence $d_1 \to c_1 \to d_2 \to \cdots \to c_{n-1} \to d_n$. This means to factorize their distribution $P_D(H_D)$ as

$$\begin{aligned} &P_D(d_1, c_1, d_2, \ldots, c_{n-1}, d_n) \\ &= P(d_1)\frac{P(d_1,c_1)}{P(d_1)}\frac{P(d_1,c_1,d_2)}{P(d_1,c_1)}\cdots\frac{P(d_1,c_1,d_2,\ldots,c_{n-1},d_n)}{P(d_1,c_1,d_2,\ldots,c_{n-1})}, \end{aligned} \tag{13}$$

where each fraction denotes a conditional probability; that is,

$$\frac{P(d_1,c_1)}{P(d_1)} = P(c_1|d_1) \tag{14}$$

is the probability of sorting the value c_1 after the value d_1;

$$\frac{P(d_1,c_1,d_2)}{P(d_1,c_1)} = P(d_2|d_1,c_1) \tag{15}$$

is the probability of sorting the value d_2 after the values d_1, c_1; and so on.

Starting with $n = 1,\ \gamma_1 = \frac{1}{2}$, and

$$P(d_1) = \frac{1}{\sqrt{2\pi\bar\xi}}\frac{\Gamma(\bar\xi+1/2)}{\Gamma(\bar\xi)\left(1+\frac{1}{2\bar\xi}d_1^2\right)^{\bar\xi+1/2}}, \tag{16}$$

such that, by making the substitution of variable

$$d_1 = \pm\sqrt{2\bar\xi}\sqrt{\frac{t_1}{1-t_1}}, \tag{17}$$

it is found that t_1 is sorted from the beta distribution $f(t_1; \frac{1}{2}, \bar{\xi})$ and the signs ± are chosen with equal probability. Proceeding, assuming that all the elements of the diagonal block of the matrix with dimension $(k-1) \times (k-1)$ are already obtained, then the off-diagonal element c_{k-1} of the $k \times k$ block is given by

$$c_{k-1} = \sqrt{\widetilde{Q}_{k-1}\bar{\xi}}\sqrt{\frac{\tilde{t}_{k-1}}{1-\tilde{t}_{k-1}}}, \tag{18}$$

where

$$\widetilde{Q}_{k-1} = 1 + \frac{1}{2\bar{\xi}}\sum_{i=1}^{k-1} d_i^2 + \frac{1}{\bar{\xi}}\sum_{j=1}^{k-2} c_j^2 \tag{19}$$

and $\tilde{t}_{k-1}$ is sorted from the beta distribution $f(\tilde{t}_{k-1}; \frac{k-1}{2}\beta, \bar{\xi} + \gamma_{k-1})$. Next, the diagonal term d_k is obtained as

$$d_k = \pm\sqrt{2Q_k\bar{\xi}}\sqrt{\frac{t_k}{1-t_k}}, \tag{20}$$

where

$$Q_k = 1 + \frac{1}{2\bar{\xi}}\sum_{i=1}^{k-1} d_i^2 + \frac{1}{\bar{\xi}}\sum_{j=1}^{k-1} c_j^2 \tag{21}$$

and t_k is sorted from the beta distribution $f(t_k; \frac{1}{2}, \bar{\xi} + \gamma_k - \frac{1}{2})$. In this way, all the elements of a disordered matrix are determined. We remark that by considering these elements as steps of a random beta process they have an interest in themselves [20,21].

2. The eigenvalues

We start observing that in the denominator of Eq. (12) $\sum_{i=1}^{n} d_i^2 + 2\sum_{j=1}^{n-1} c_j^2 = \mathrm{tr}H_D^2 = \sum_{i=1}^{n} x_i^2$, with $x_i = \sqrt{\frac{\bar{\xi}}{\xi}}E_i$. Besides, as we are dealing with tridiagonal matrices whose subdiagonal elements are positive, they satisfy two important lemmas: the first one states that the Vandermonde determinant is given by

$$\Delta(x) = \prod_{j>i}(x_j - x_i) = \frac{\prod_{i=1}^{n-1} c_i^i}{\prod_{i=1}^{n} q_i}, \tag{22}$$

and the second one states that the Jacobian of the transformation from matrix elements to eigenvalues and eigenvectors is given by

$$J = \frac{\prod_{i=1}^{n-1} c_i}{\prod_{i=1}^{n} q_i}, \tag{23}$$

where q_i are elements of the first row of the eigenvector matrix. Substituting these results in Eq. (12) we derive that eigenvectors and eigenvalues decouple, and we have

$$P_D(x_1, x_2, \dots, x_n) \propto \frac{\prod_{j>i}|x_j - x_i|^{\beta}}{\left(1 + \frac{1}{2\bar{\xi}}\sum_{i=1}^{n} x_i^2\right)^{\bar{\xi}+\gamma_n}} \tag{24}$$

such that if the integral representation of the gamma function is used we can write the normalized expression

$$P_D(x_1, x_2, \dots, x_n) = \int_0^{\infty} d\xi\, w(\xi)\left(\frac{\xi}{\bar{\xi}}\right)^{n/2} P_{\beta}\left(\sqrt{\frac{\xi}{\bar{\xi}}}x_1, \sqrt{\frac{\xi}{\bar{\xi}}}x_2, \dots, \sqrt{\frac{\xi}{\bar{\xi}}}x_n\right). \tag{25}$$

Therefore, as a consequence, the statistical measures of the disordered beta ensemble are obtained by averaging those of the beta ensemble with the distribution $w(\xi)$.

Thus in the RMT regime $n\beta \gg 1$, the one-point function, that is the density, is obtained by integrating all eigenvalues and multiplying by n, giving

$$\rho_{D\beta}(x) = \int_0^{\infty} d\xi\, w(\xi)\left(\frac{\xi}{\bar{\xi}}\right)^{1/2}\rho_{\beta}\left(\sqrt{\frac{\xi}{\bar{\xi}}}x\right) = \frac{1}{\pi\beta}\int_0^{\xi_{\max}} d\xi\, w(\xi)\left(\frac{\xi}{\bar{\xi}}\right)^{1/2}\sqrt{2n\beta - \frac{\xi}{\bar{\xi}}x^2}, \tag{26}$$

where $\xi_{\max} = 2n\beta\bar{\xi}/x^2$. Integrating the density, Eq. (26), from the origin to a value x, the cumulative function is obtained as

$$N_{D\beta}(x) = \frac{n}{2}\left\{1 - \int_0^{\xi_{\max}} d\xi\, w(\xi)\left[1 - \frac{2}{n}N_{\beta}\left(\sqrt{\frac{\xi}{\bar{\xi}}}x\right)\right]\right\}. \tag{27}$$

To measure the short-range spectral fluctuations we define the probability $E(s)$ that the interval $(-\frac{s}{2}, \frac{s}{2})$ is empty. This so-called gap probability is obtained by integrating over all eigenvalues outside the interval $(-\frac{\theta}{2}, \frac{\theta}{2})$ to obtain

$$E_{D\beta}(s) = \int_0^{\infty} d\xi\, w(\xi) E_{\beta}\left[2N_{\beta}\left(\sqrt{\frac{\xi}{\bar{\xi}}}\frac{\theta}{2}\right)\right], \tag{28}$$

with $s = 2N_{D\beta}(\frac{\theta}{2})$. From the gap probability, the NND is obtained by taking the derivatives, $F(s) = \frac{dE}{ds}$ and $p(s) = \frac{d^2E}{ds^2}$, such that $p(s)$ is the distribution and $1 + F(s)$ the probability. Analytic expressions for $E_{\beta}(s)$ are known only for the Gaussian ensemble; in particular, for the case of $\beta = 1$, we use the Wigner surmise

$$E_1(s) = \mathrm{erfc}\left(\sqrt{\frac{\pi}{4}}s\right). \tag{29}$$

The number variance of the disordered ensemble is given by (see the derivation in Appendix A)

$$\Sigma^2_{D\beta}(L) = \langle n^2\rangle - \langle n\rangle^2 = \int_0^\infty d\xi\, w(\xi)\left\{\Sigma^2_\beta\left[2N_\beta\left(\sqrt{\frac{\xi}{\bar{\xi}}}\frac{\theta}{2}\right)\right] + 4N^2_\beta\left(\sqrt{\frac{\xi}{\bar{\xi}}}\frac{\theta}{2}\right)\right\} - L^2, \tag{30}$$

where $L = \langle n\rangle = 2N_{D\beta}(\frac{\theta}{2})$ denotes the unfolded interval. Using the results of Appendix B, we find that asymptotically the number variance takes the simple form of a parabola given by

$$\Sigma^2_{D\beta}(L) \simeq \left[\left(\frac{\overline{\sqrt{\xi}}}{\sqrt{\bar{\xi}}}\right)^{-2} - 1\right]L^2 \simeq \frac{1}{4\bar{\xi}}L^2, \tag{31}$$

where the Σ^2_β term that increases logarithmically, in the presence of the L^2 term, has been neglected. Therefore, in this case the number variance satisfies Taylor's law [22] exhibiting the phenomenon that has more recently been named the fluctuation scaling mechanism [23–25].

On the other hand, considering the gap probability, the first order term in Eq. (B4) can be used such that with $s \simeq \rho(0)\theta$ being replaced in Eq. (28), it becomes

$$E_{D\beta}(s) = \int_0^\infty d\xi\, w(\xi) E_\beta\left(\sqrt{\frac{\xi}{\bar{\xi}}}s\right) \simeq E_\beta(s), \tag{32}$$

as the disorder fluctuations are quenched in the large $\bar{\xi}$ limit.

The above results show that, asymptotically, the two sources of randomness acting on the system affect, differently, the short- and the long-range statistics. The local statistics are described by the expressions of the beta ensemble, while the long-range number variance is dominated by the external source of randomness, similar to systems affected by a strong external driving in which the internal dynamics of the system essentially becomes irrelevant for large L [24]. This important result is illustrated by the numerical simulations exhibited in Fig. 1, which clearly shows the robustness of the local statistics in contrast with the high sensitivity of the long-range one.

It is important to observe that when the value of the parameter $\bar{\xi}$ approaches 0, a strong disorder regime is reached in which the local statistics show a power-law decay [9,17]; also cf. the discussion in Appendix C. It is also noteworthy that, for reasons that will be clear in Sec. II D, Eq. (31) is still not adequate to capture the characteristics of the empirical data studied in Sec. III. Though its parabolic form seems to be in accord with what one observes in the data, some further fine-tuning is still required. One possibility is to additionally incorporate the thinning ensemble.

C. Thinned ensemble

Using the beta ensemble, we are supposing that we are dealing with perfect spectra which, necessarily, is not the case especially for those originating from complex systems that we are interested in. In order to make our approach more comprehensive, we add to the model the formalism introduced in [11] that deals with incompleteness in a sequence being analyzed. This formalism was a natural development of the missing level theory [10] and it consists in constructing from a given spectrum a new one by removing with a probability $1 - f$ levels from it, such that the resulting spectrum has, in average, f levels. In statistics, this construction is denoted as a thinning point process [26], and it has the important aspect of preserving, in the less dense object, properties of the original one. The RMT formalism is based on Fredholm determinants [2], and this determinantal method is preserved by the thinning process. This fact explains the great attention that has recently been attracted by this model [27–31]. In our case, we use it as a sort of an error-correction code.

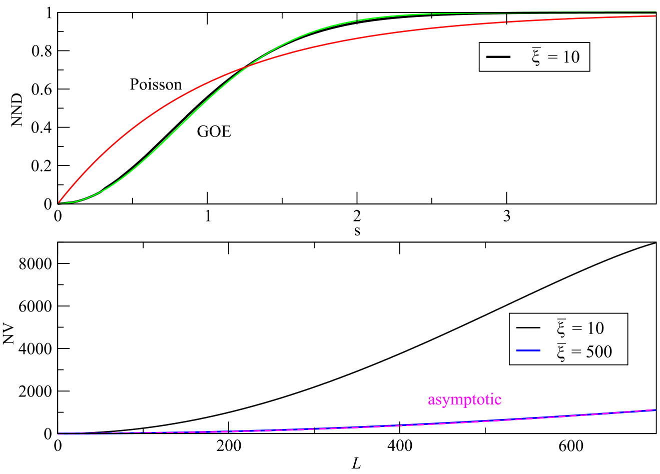


FIG. 1. The effect of disorder on the GOE ($\beta = 1$) cumulative nearest-neighbor distribution and on the number variance. It shows the robustness of the local statistics in contrast to the sensitivity of the long-range one. For comparison, the cumulative NNDs of the Poisson ensemble (red) and the original GOE ensemble (green), as well as the asymptotic NV of the disordered GOE ensemble (magenta), are also displayed.

In [11], it is shown that the thinned spectra have statistics intermediate between RMT and Poisson. Moreover, the RMT formalism analytically also describes this intermediate situation with f playing the role of a parameter that varies from zero to one. In terms of the spacing distribution of the initial spectrum, the NND is given by

$$p(s, f) = \sum_{k=0}^{\infty}(1 - f)^k P\left(k, \frac{s}{f}\right), \tag{33}$$

where the lower case denotes the incomplete quantity and the upper case the original complete one. The $P(k, s)$ are spacing distribution with k levels inside the interval s and the division by f taking into account the contraction of the incomplete spectrum. However, this construction is not very practical for data fitting purposes due to the lack of an exact form. As shown in Fig. 2 (where disorder effects are irrelevant in the NND; see Sec. II D), a simple beta ensemble fitting can be quite handy in cases when $f \sim 1$, especially when data noise

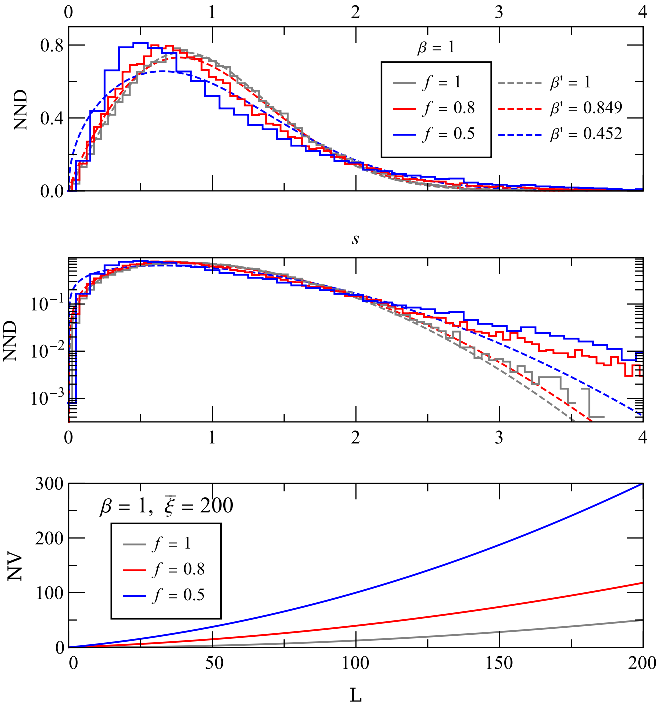


FIG. 2. The effect of removing levels (solid lines) on the nearest-neighbor distribution (top pane), the cumulative nearest-neighbor distribution (middle pane), and the number variance (bottom pane) of the disordered GOE ($\beta = 1,\ \bar{\xi} = 200$), with $f = 1$ (gray), 0.8 (red), and 0.5 (blue), respectively. The nearest-neighbor distribution implies that the thinned spectra still have level repulsion and an exponential decay, while the number variance, asymptotically, shows a parabolic behavior. For $f \neq 1$, the dashed lines correspond to the beta ensemble fitting with the fitting parameter β'. Hence, for $f \sim 1$, the effect of the thinning process on the NND is weak, one may consider a beta ensemble fitting still to be appropriate.

also inevitably render large error margins in the fitting results. Since in (33) all spacing distributions $P(k, s)$ are normalized with first moments $\langle s\rangle_k = k + 1$, it then follows that $p(s, f)$ also is normalized and has the first moment equal to one. This shows the need of rescaling the argument of the functions with the parameter f. Accordingly, for the density we have

$$\rho_\beta(x, f) = f\rho_\beta(x), \tag{34}$$

while for the number variance it can be shown that [11]

$$\sigma^2_\beta(L, f) = (1 - f)L + f^2\Sigma^2_\beta\left(\frac{L}{f}\right). \tag{35}$$

The NND expression implies that the thinned spectra still have level repulsion and an exponential decay as can be seen in Fig. 2. On the other hand, the number variance expression shows an asymptotically straight line Poisson behavior. Finally, we remark that the thinning process has no effect on an uncorrelated Poisson spectrum.

D. The combined ensemble

In practice, we found that the parabolic tails of data NVs can not be delineated by solely considering the beta ensemble and disorder effects. Actually, from an FS point of view, in many real systems, fluctuations of certain count n should scale as $\Sigma^2(n) \sim \langle n\rangle^\alpha$. An FS exponent of value 2 corresponds to totally disordered patterns; an FS exponent of value 1 indicates uncorrelated or short-range correlated patterns (e.g., the Poisson ensemble); and an FS exponent less than 1 or a logarithmic NV [e.g., Eq. (7)] hints at anticorrelated patterns [24]. Due to nontrivial *positive* correlations between the constituent components, real systems are usually characterized by an FS exponent between 1 and 2, the NV of which may be alternatively expressed as a combination of a linear term and a quadratic term [22,24]. By noting the linear term introduced via the thinning process in Eq. (35), it is then natural to also incorporate the thinned ensemble for the general construction. Henceforth, by combining the above three ingredients, we have a disordered randomly incomplete β-spectrum that we will coin as belonging to the *disordered beta thinned ensemble*.

According to Eqs. (32) and (33), the local statistics should be robust with respect to disorder but can be altered by thinning effects. In Fig. 2 numerical simulations show that the NND displays a typical intermediate statistics behavior between Poisson and RMT. As a rule of thumb, the thinning effect on the NND is weak when $f \sim 1$ so that the NND may still be fitted using only the beta ensemble curve. But for the number variance data, better results are obtained by using the expression that takes into account both disorder and incompleteness, which gives

$$\Sigma^2_{D\beta T}(L) = \int_0^\infty d\xi\, w(\xi)\left[\sigma^2\left(2N_\beta\left(\sqrt{\frac{\xi}{\bar{\xi}}}\,\frac{\theta}{2}\right), f\right) + \frac{4}{f^2}N^2_\beta\left(\sqrt{\frac{\xi}{\bar{\xi}}}\,\frac{\theta}{2}\right)\right] - \left(\frac{L}{f}\right)^2, \tag{36}$$

which, asymptotically, becomes

$$\Sigma^2_{D\beta T}(L) \simeq (1 - f)L + \left[\left(\frac{\overline{\sqrt{\xi}}}{\sqrt{\bar{\xi}}}\right)^{-2} - 1\right]\left(\frac{L}{f}\right)^2, \tag{37}$$

where $L = 2N_{D\beta}(\frac{\theta}{2})$. Hence, the NV shows a much more sensitive dependence on the thinning parameter f than the NND; cf. Fig. 2. One may then make use of the NV to obtain the value for f and utilize it to fit the NND data when Eq. (33) has to be fully employed if $f \sim 0$.

Therefore, in all cases we have NVs that follow Taylor's law in the super-Poissonian parabola form $aL + bL^2$. This implies the presence of an FS phenomenon. As a matter of fact, here the fluctuation in the scaling can be understood as the manifestation of the breaking of the ergodicity of the ensemble enacted by the introduction of the external randomness [9]. Ergodicity in RMT means that averaging over one large matrix is equivalent to an ensemble average; in other words, individual matrices are equivalent, which, by construction, is not the case of the disordered ensemble. It is interesting to observe that fluctuations are enhanced by the thinning process; cf. Fig. 2. To understand this, note that the super-Poissonian behavior just means that the eigenvalues have a tendency to aggregate at certain values rather than distributing uniformly [22], be it caused by *positive* correlations between the constituent components, disordered inhomogeneity in the system, external driving on the system or criticality, etc. [24], whereas the thinning process simply removes levels uniformly so that aggregations of eigenvalues are further enhanced overall.

III. APPLICATIONS

RMT has been successfully applied to a wide scope of complex systems, such as the stock market [32], brain activities [33], and atmospheric variabilities [34], to name just a few. To demonstrate the versatility of the combined ensemble, in this section we apply the formalism to the analyses of spectra extracted from three different areas, namely, the Laplacian matrix spectra of complex networks, the cross-correlation matrix spectra of COVID-19 time series, and the spectra of blanks in Portuguese and Chinese literary texts. We investigate the NND and NV of all the data. One point we would like to stress is that we are considering spectra for which an average constant density can be assumed. This allows us to rescale the data with respect to the so-called unfolding process, so that the average spacing equals one, as proceeded above to obtain the NV in terms of L, which will be explained further below for our data processing.

A. Complex networks

From transport infrastructures to biological systems, social interactions, neural networks, and the Internet, a varied array of systems are made by a large amount of highly interconnected dynamical units [35,36]. One way to capture the global properties of these systems is to model them as graphs consisting of pairs of nodes connected via links that stand for the interactions between the dynamical units. Traditionally, the Erdős-Rényi random graphs [37] were most comprehensively studied for the investigation of complex network properties. However, growing observations revealed that many real networks behave quite differently from Erdős-Rényi random graphs, and more realistic network models, such as the small-world network [38] and the scale-free network [39], have been proposed.

The spectral analysis of Laplacian matrices of networks is an important tool for extracting the structural properties of complex networks. It turns out that these spectra can also be analyzed under the RMT framework [40–43]. Considering an undirected network with N nodes ($i = 1, \dots, N$), the Laplacian matrix is defined as $\mathcal{L} = K - A$ [44], where $K = \mathrm{diag}(k_i)$ is a diagonal matrix consisting of node degrees k_i and A is the adjacency matrix with elements $A_{ij} = 1$ if nodes i and j are connected and $A_{ij} = 0$ if otherwise. In what follows, we apply RMT to analyze short- and long-range eigenvalue statistics of the Laplacian matrices of model networks and real networks, which are described by the NNDs and the NVs of the eigenvalue spectra.

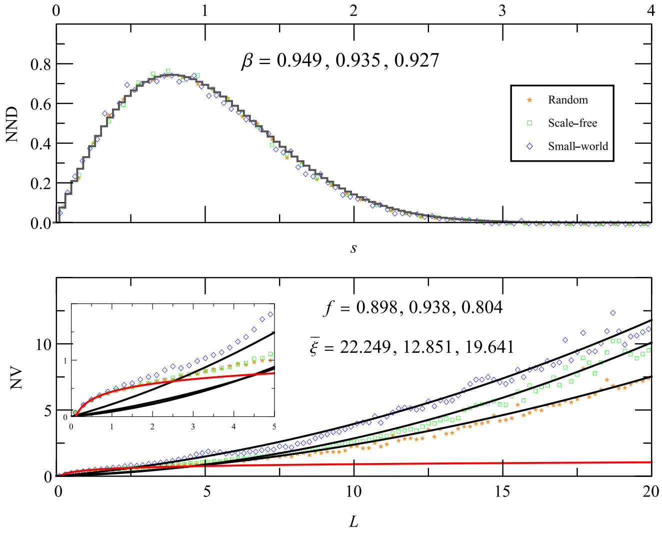


FIG. 3. The NND (top pane) and the NV (bottom pane) of random, scale-free, and small-world networks, fitted with respect to Eqs. (5) and (37). The red curve in the bottom pane marks the GOE NV given by Eq. (7). The fitted parameters are listed, from left to right, with respect to random, scale-free, and small-world networks. The black curve in the top pane shows the averaged NND. The inset of the bottom pane shows that the data NVs follow the universal GOE prediction only for small L.

For model networks, we investigate three well-known model networks: random networks, scale-free networks, and small-world networks. Following Ref. [42], we construct networks of $N = 2000$ nodes and similarly set the connection probability p between pairs of nodes to 0.01 for random networks, the average node degree k to 20 for scale-free networks, and for small-wold networks, the rewiring probability $p = 0.005$ and the average degree $k = 40$. Previously, long-range correlations of network ensembles were also studied via the Δ_3-statistic, which show universal GOE behavior for relatively small L but pick up quadratically for larger L, especially for networks with higher heterogeneity [42,43]. We will show that the NV displays similar behavior but more sensitively depends on external sources of randomness which can be fully captured only by the combined ensemble.

To proceed, we first order the eigenvalues x_i of the network Laplacian matrix. Following the standard procedure [2], these eigenvalues are then unfolded according to $\overline{x}_i = \overline{N}(x_i)$, where $\overline{N}(x) = \int_{x_{\min}}^{x} \rho(x')dx'$ is the cumulative function [cf. Eqs. (6) and (27)], so that the transformed eigenvalues exhibit a uniform spectral density $\rho(\overline{x}_i) = 1$. Since the functional form of $\overline{N}$ usually cannot be deduced, we have resorted to polynomial curve fitting for the cumulative density data. The spacing is then calculated as $s_i = \overline{x}_{i+1} - \overline{x}_i$, and the NND and the NV are defined, following Sec. II, as the probability distribution $p(s)$ of spacing and the variance for the number of unfolded eigenvalues averaged over nonoverlaping intervals of length L, respectively.

Figure 3 shows the NND, the NV, and their fitted curves by using the least square fitting method for the three network ensembles, with the data for each ensemble averaged over 10 realizations of networks. As shown in the top panel of Fig. 3, the NNDs of the three network ensembles all can be fitted with the beta ensemble with $\beta \sim 1$, indicating a universal GOE behavior of the NNDs [40–43]. The NNDs of different networks can be considered as just fluctuating around an average distribution (the black curve), suggesting the robustness of local statistics. This of course just reiterates what previously had been found that many systems follow the universal GOE or GUE behavior for their short-range statistics [42,43,45–48]. In contrast, long-range statistics, measured by either the NV or the Δ_3-statistic, follow the universal GOE or GUE prediction only up to certain values of L [42,43,45–49]. In particular, the NVs are deemed to be sensitive to external randomness. As indicated by the inset of the bottom panel of Fig. 3, the NVs of the three network ensembles follow the GOE statistic only for very small L and quadratically deviate from it as L increases. Even though parabolic-like NVs were observed for disordered matrix ensembles [9], previous studies haven't concluded an explicit explanation for the quadratic tails of long-range statistics observed in many empirical systems. In this work, we found that this behavior of NV can be best accounted for by the disordered beta thinned ensemble; compare the black curves fitted according to Eq. (37) in the bottom pane of Fig. 3. Therefore, external randomness play important roles in long-range statistics. In particular, the presence of a linear term in the fitting form hints the emergence of the fluctuation scaling mechanism that should be ultimately ascribed to the inhomogeneity and finite-size nature of the networks.

From the fitting parameters for the NV data, let us make two observations. First, with $f \sim 1$, the thinning processes are not pronounced for these studied cases, which validates the beta ensemble fitting for NNDs. This can be understood, with respect to the chosen network parameters, by the fact that the studied networks are highly connected to form a large connected component so that the corresponding Laplacian matrices and the eigenvalue spectra are dense. Second, the NVs of the three network ensembles are clearly distinguished by their fitting parameters: the more heterogeneous the network ensemble is, the further its NV deviates from the GOE prediction, which is also indicated by the fitting parameter $\bar{\xi}$. According to Appendix C, larger $\bar{\xi}$ signifies weaker disorder so that the unfolded eigenvalues are also more uniformly distributed. Comparing to the other two kinds of networks, the random network ensemble, being homogeneous by its nature, thus takes the largest $\bar{\xi}$. On the other hand, small-world networks and scale-free networks are characterized by high clustering coefficients and nonvanishing probability for rare hubs [36], respectively, so that the corresponding unfolded eigenvalues are also more aggregated as compared to random networks and hence give rise to larger NV values.

As an example for real networks, we consider fungal networks adopted from Ref. [50]. In Fig. 4 four different species of fungal networks are analyzed, with the labels corresponding to (a) *Physarum polycephalum* (Pp), an acellular slime mold that forms networks but is taxonomically distinct from fungi; (b) *Phanerochaete velutina* (Pv), a foraging saprotrophic woodland fungus that forms reasonably dense networks; (c) *Resenicium bicolor* (Rb), a white-rot fungus that forages rapidly with a sparse network that is not very cross-linked; and (d) *Agrocybe gibberosa* (Ag), a foraging saprotrophic fungus that is isolated from garden compost and forms dense networks. More results for a total number of 270 fungal networks are available in [51]. Note that since only one network is considered for each set of data, both the NND and the NV data appear to be a little noisy. According to the top pane of Fig. 4, the NNDs of the data again fluctuate around some averaged beta ensemble distribution which, however, is quite different from that of the GOE. Being taxonomically different from the other three networks for fungi, one naturally expects the Pp network to display some distinctive characteristics. Indeed, the bottom pane of Fig. 4 shows that the NVs follow the disordered beta thinned ensemble statistics, in which the Pp data separates from the data for the three fungal networks with quite evident gaps and the NVs of the three fungal networks are more or less grouped together. It is also striking to observe that Rb, corresponding to the most sparse network, gives a thinner spectrum than the rest ones. Nevertheless, the fungal networks for Pp, Pv, and Ag are still dense enough as indicated by their relatively large f values and are robust against damage [52].

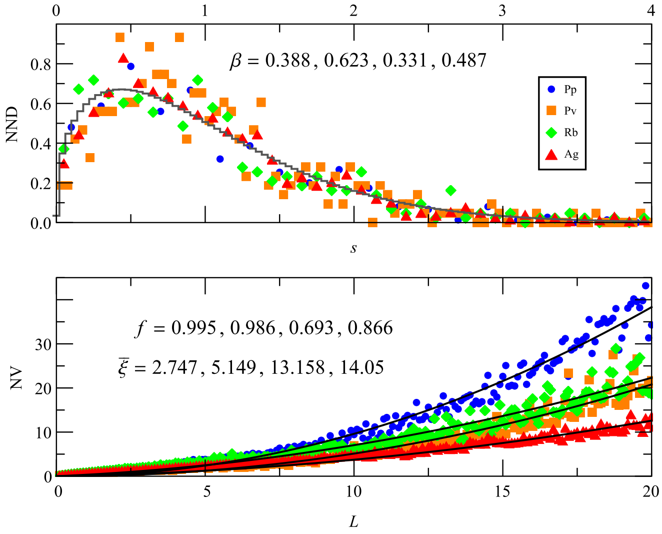


FIG. 4. The NND (top pane) and the NV (bottom pane) of the Pp, Pv, Rb, and Ag fungal networks, fitted with respect to Eqs. (5) and (37). The fitted parameters are listed, from left to right, with respect to the Pp, Pv, Rb, and Ag networks. The black curve in the top pane shows the averaged NND.

The above observations for different networks thus underscore the significance of disorder and thinning processes in RMT analyses. The NNDs belonging to the same subclass of data that appear to be related to certain matrix ensemble are usually rigid with respect to changes in external randomness and hence are mostly dictated by that matrix ensemble. However, this is usually not the entire story for many empirical data, as external randomness may affect each system globally in a distinctive way. Long-range correlations are then crucial to capture these effects and are useful to separate data belonging to the same subclass, which can be of assistance to network classification, especially when it is exploited together with other characteristic features of interested networks.

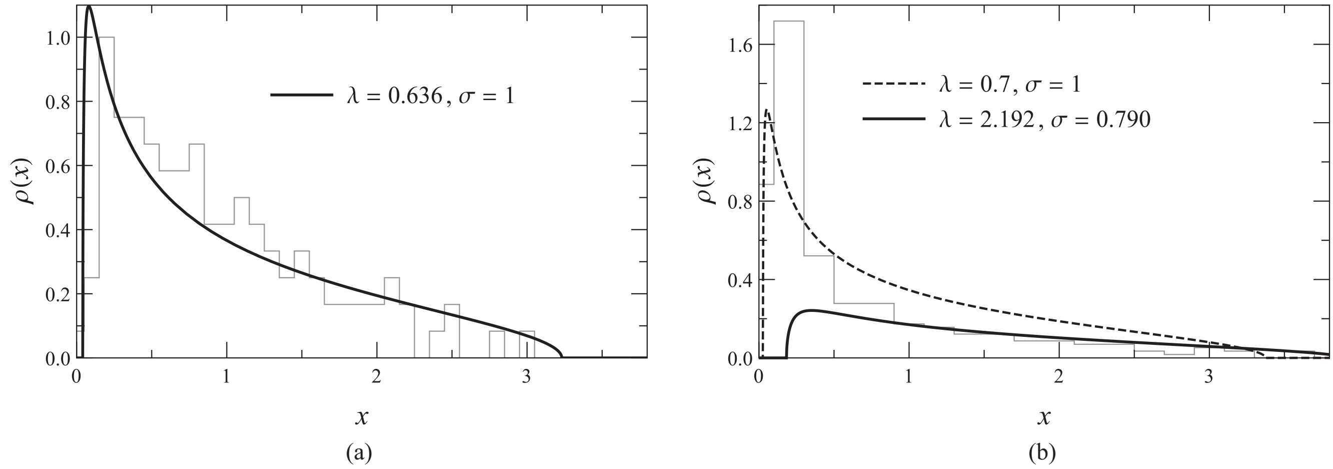


FIG. 5. The eigenvalue distributions of the COVID-19 cross-correlation matrices of (a) global countries and (b) U.S. counties. The solid smooth curves are fitted with respect to the Marčenko-Pastur distribution with asymptotic ratio λ and scale parameter σ. The fitting curve (solid) for U.S. counties in (b) agree only with the tail of the distribution. The dashed smooth curve shows an attempt to fit the initial part of the distribution via manually tuning the parameters.

B. COVID-19 time series

The ongoing pandemic caused by the contagious disease named Coronavirus Disease 2019 (COVID-19) poses the latest threat to global health, which has also triggered a series of related studies via physics concepts and methods, ranging from modeling its spreading dynamics to the study of the role of respiratory droplets [53–58]. In this work, we focus on analyzing the time series of daily new COVID cases in different regions. To this end, by following the RMT analyses of financial time series [32,49,59,60], we study the cross-correlations between daily new cases changes of different countries or of different counties of the United States, where the analyzed data were obtained from the John Hopkins COVID database [61]. In order to draw statistically meaningful conclusions, only data sets with more than 20 000 total cases were considered, corresponding to the data sets of $N = 120$ global countries and $N = 288$ counties in the United States, respectively.

Let us denote $S_i(t)$ the number of daily new cases of country (or county) i on day t, where $i = 1, \dots, N$ and the time t spans over a period of 10 months from May 1, 2020, to March 1, 2021. Following the financial analysis convention, the daily new cases change $G_i(t, \Delta t)$ is calculated with respect to the logarithmic scale as

$$G_i(t, \Delta t) \equiv \ln S_i(t + \Delta t) - \ln S_i(t), \tag{38}$$

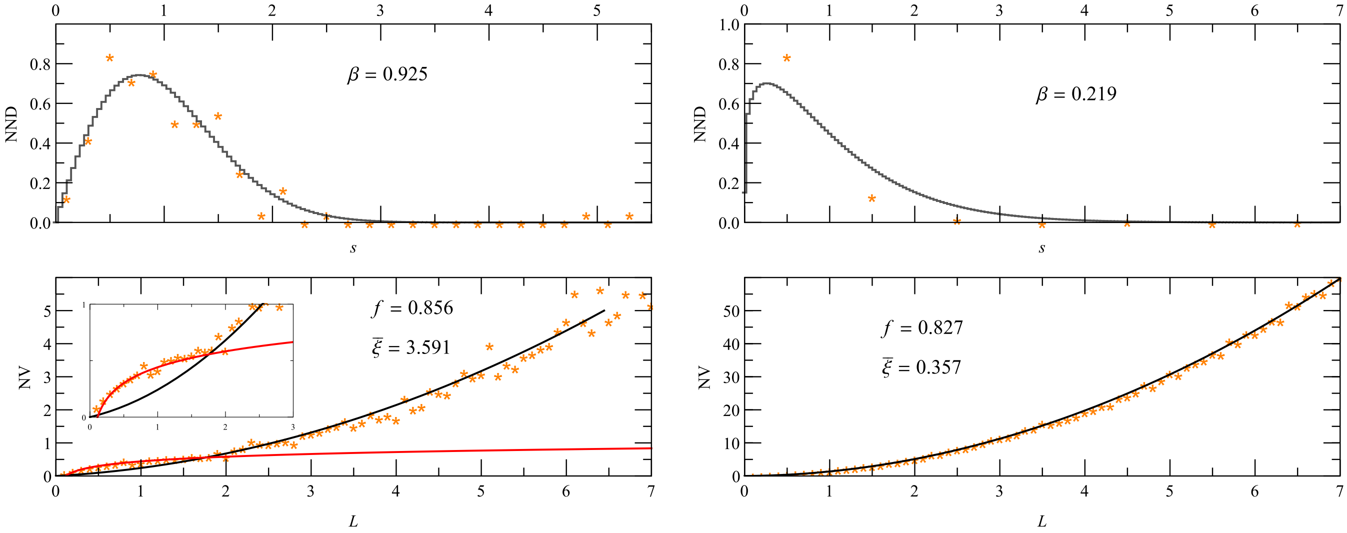
where $\Delta t = 1$ day is the sampling time interval. The time-series correlations between different countries (or counties) are then assembled into the equal-time cross-correlation matrix C with elements [59]

$$C_{ij} \equiv \frac{\langle G_i G_j \rangle - \langle G_i \rangle \langle G_j \rangle}{\sigma_i \sigma_j}, \tag{39}$$

where $\sigma_i \equiv \sqrt{\langle G_i^2 \rangle - \langle G_i \rangle^2}$ is the standard deviation of the daily new cases changes of country (or county) i, and $\langle \cdot \rangle$ denotes the time average over the studied 10-month period. The obtained cross-correlation matrix C then plays the role of a random matrix, validating the application of RMT and the ensuing NND and NV analyses after the eigenvalues are unfolded [59].

As shown in Fig. 5(a), our first observation is that, similar to what had been found in financial time series [32,49,60], the distribution of the bulk eigenvalues of C is close to the Marčenko-Pastur distribution of the Wishart orthogonal ensemble. According to Ref. [32], Wishart matrices are not strictly GOE-type matrices, but belong to a special ensemble called the "chiral" GOE, whose short- and long-range eigenvalue correlations still manifest universal GOE properties. Hence, deviations of the studied data from the Marčenko-Pastur distribution may be suggestive of, in addition to data noise ascribed to the relatively small matrix sizes, external randomness, or even discrepancies from the GOE in eigenvalue correlations. For the U.S. county case [cf. Fig. 5(b)], the eigenvalue distribution data can not be fitted with respect to the Marčenko-Pastur distribution, hinting at a deviation from the GOE behavior [62]. This is immediately confirmed, as shown in Fig. 6(b), by the NND and the NV of the cross-correlation matrix. The fitted β value for the global country data is still close to one, rendering a universal GOE behavior for its NV when L is small [cf. the inset of the bottom pane of Fig. 6(a)]. For larger L, the NV tails of both cases turn out to follow the disordered beta thinned ensemble. It is noticeable that the NV values of the U.S. county data are much larger (with a smaller ξ value) than those of the global country data. To understand this, it is more apt to interpreter it from a fluctuation scaling perspective [24], which just suggests that the U.S. counties are more strongly related by *positive* correlations than between different global countries, so that the corresponding eigenvalues of the U.S. county matrix are more prone to aggregate around certain values. Note again that with $f \sim 1$ for both cases, the thinning effects are also weak and it is valid to fit the NNDs with respect to the beta ensemble.

In summary, this example demonstrates that even though the cross-correlations of the COVID-19 time series data



(a)

(b)

FIG. 6. The NND and NV of the COVID-19 cross-correlation matrices of (a) global countries and (b) U.S. counties, fitted with respect to Eqs. (5) and (37). The red curve in the bottom pane of (a) marks the GOE NV given by Eq. (7). The U.S. county data show an apparent discrepancy with respect to the GOE.

contain great elasticity, one may still put them into an RMT framework with the combined ensemble, regardless of whether they follow the GOE or not. What is more, as a reflection of the characteristics of each individual data set, the NV may be exploited as an alternative means for exploring regional intracorrelations of different administrative levels in response to a pandemic or to other social or economical activities.

C. Literary texts

A writing system is a process or result of recording a spoken language using a system of visual marks on a surface. There are mainly two types of writing systems: phonographic and logographic. The former includes syllabic writing (e.g., Japanese hiragana) and alphabetic writing (e.g., English, Russian, or Portuguese), while the latter encodes syllables and phonemesa (e.g., Sumerian cuneiforms, Egyptian hieroglyphs, or Chinese characters). In this subsection, we show that the spectra of these two types of texts are in good agreement with the disordered beta thinned ensemble as well, with Portuguese texts and Chinese texts taken as representative examples.

1. Spectra of blanks

In Ref. [5], the spectra of blanks of literary texts of ten languages were analyzed and two language families have been found. The family denoted Poisson-like were fitted with a displaced Poisson distribution as short words did not show a clear statistical behavior. For this reason, we are not considering here spectra from this family and decided to perform a reanalysis of the spectra of Portuguese, a language of the Wigner-like family. In this study, we take the same four Portuguese literary texts as Ref. [5]: *A Filha do Arcediago* by C. C. Branco (1868), *O Primo Basílio* by E. de Queirós (1878), *Os Maias* by E. de Queirós (1888), and *Grande Sertão: Veredas* by J. G. Rosa (1956), which are abbreviated as AFilha, Oprimo, OsMaias, and Grande, respectively. Defining word length as spacing s, the NND is just computed as the distribution $p(s)$ and the NV measures the variance of the number of levels contained in the interval of length L, averaged over all nonoverlapping intervals taken from a spectrum.

The results for NNDs and NVs are presented in Fig. 7. For all the cases in the NND, we find a good agreement with the NND of the beta ensemble, with $\beta < 1$. The NNDs can also be seen as fluctuating around an average distribution,

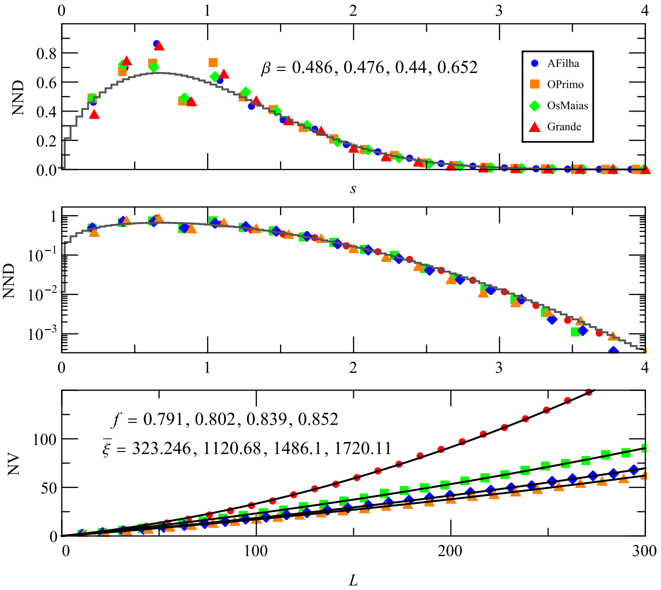


FIG. 7. The NND (top pane), the NND in logarithmic scale (middle pane), and the NV (bottom pane) of four Portuguese texts, fitted with respect to Eqs. (5) and (37). The fitted parameters are listed, from left to right, with respect to the texts: AFilha, Grande, OPrimo, and OsMaias. The black curves in the top and the middle panes show the averaged NND.

which is further confirmed in the log-scale plot. The NV data satisfactorily match the NV of the disordered beta thinned ensemble. We observe that the values of f and $\bar{\xi}$ change in order, and the NV separates the four texts in the order of their publication years, respectively. Note that the validity of the beta ensemble fitting for NNDs is again justified by the weak thinning effects.

2. Chinese texts

The building blocks of the Chinese writing system are Chinese characters—a collection of spatially marked patterns of continuous strokes. In an ideogram language such as Chinese, strokes play a similar role as letters do in alphabetic languages. All Chinese characters are composed of the basic strokes "一", "丨", "丿", "丶", "フ", and their variants [63]. For example, the character "生" is composed of five strokes: "丿", "一", "一", "丨", and "一"; and the character "命" is composed of eight strokes: "丿", " ", "一", "丨", " ", "一", "丨", and " ", where the second, fifth, and eighth strokes are variant forms of "丶" and "フ". In this sense, we can consider the word "生命" (meaning "life") as a spectrum consisting of three levels that are separated by two intervals of lengths 5 and 8. For long texts, we obtain the stoke counts of each character by looking up the *Stroke Table of Unihan Characters* [64]. A Chinese literary text, after discarding blanks and punctuation marks, then produces a long spectrum of levels that can be treated in the same vein as above, where spacing s is now defined as the number of strokes of each character. Below we investigate four Chinese texts: *Dream of the Red Chamber* (红楼梦) by Xueqin Cao (1791), *Water Margin* (水浒传) by Nai'an Shi (14th century), *Ordinary World* (平凡的世界) by Yao Lu (1986), and *Stories of the Sahara* (撒哈拉沙漠的故事) by Mao San (1976), which are denoted as HLM, SHZ, PFD, and SHL according to their Chinese pronunciation initials.

As can be seen from Fig. 8, the NNDs and NVs of the four Chinese texts again show good agreement with the NND of the beta ensemble and the NV of the disordered beta thinned ensemble, respectively. The beta ensemble distribution fitting for the NNDs, which, however, gives $\beta > 1$, in contrast to $\beta < 1$ for the Portuguese case, is also justified by the weak thinning effects.

To demonstrate that the above observations for Portuguese and Chinese texts are not just specific to the selected texts, we further expanded our study to include 467 Portuguese texts and 105 Chinese texts obtained from "Project Gutenberg"; see the data and the fitting results in [51], which further corroborate the above observations. From these examples, it is heuristic to speculate that the short-range statistics is determined by the language and is rather insensitive to the changes of other factors, while long-range statistics shows great diversity from one text to another. Even though it is yet unclear how the fitting parameters are explicitly related to the language, the genre, the writing era, or even the author, etc., of a given text, a combined scrutinizing of both the short- and long-range statistics may provide insightful information that reflects both the rigid and the elastic properties of a text spectrum. Similar to complex networks, such information then may be of great value for text classification, and we hereby advocate for further studies.

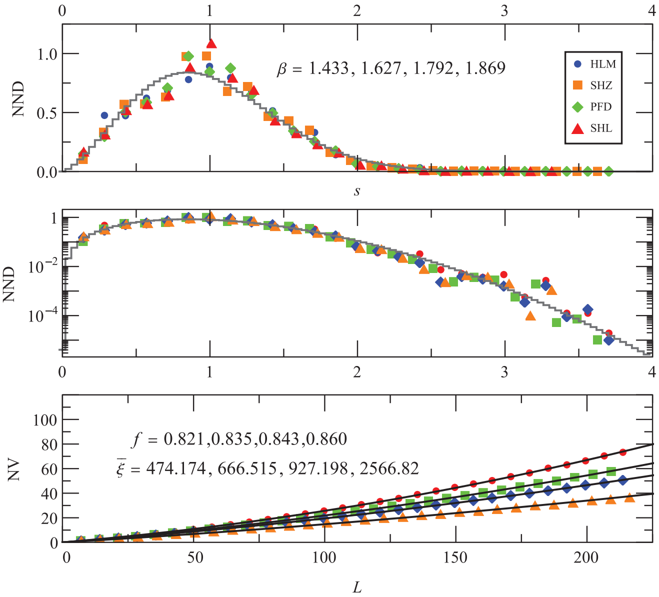


FIG. 8. The NND (top pane), the NND in logarithmic scale (middle pane), and the NV (bottom pane) of four Chinese texts, fitted with respect to Eqs. (5) and (37). The fitted parameters are listed, from left to right, with respect to the texts: HLM, SHZ, PFD, and SHL. The black curves in the top and the middle panes show the averaged NND.

IV. CONCLUSION

In this work, we have introduced the general *disordered beta thinned ensemble* to cope with external randomness in complex systems that are invoked by both disorder and level incompleteness. The model combines three generalizations of the classical RMT ensembles: the beta ensemble, the disordered ensemble, and the thinned ensemble in which statistic measures lie intermediately between Poisson and RMT. The disordered and thinned ensembles were developed as the extensions of standard RMT, but here we show that they also work in the beta case. By making use of the correlations between the matrix elements, we also demonstrated that the matrix elements of a disordered beta ensemble can be alternatively generated in a sequential manner via the beta distribution.

Guided by the main findings for the disordered beta thinned ensemble, we have analyzed spectra from three different areas that are outside the scope of Hamiltonian physical systems. The analyses were done by fitting the results which were obtained by averaging along the empirical spectra ("time" average) with the analytic expressions deduced by performing in the RMT model an ensemble average. We found that all the studied cases are in good agreement with the combined ensemble, meaning that all the three considered ingredients are crucial. Hence we naturally expect that the combined ensemble may find broader applications. From the results, two main aspects are evident: first, the relative independence between the local- (NND) and the long-range (NV) statistics, and second, NVs of the parabolic form $aL + bL^2$, which is a characteristic of the fluctuation scaling (FS) phenomenon. These two kinds of behaviors are not entirely independent. In fact, considering that linear NV is a typical behavior of an uncorrelated spectrum for the extreme case $\beta \to 0$ (Poisson), we can infer that for relative short-range, the points of the spectra appear as an independent sequence, although the distances between neighbors follow a given NND (see the footnote [65]). However, as the range of observation increases, there is a crossover from the linear to the square dependence of the mean in the NV. This can be understood as being caused by the presence of external drives, i.e., disorder and thinning, that create the long-range inhomogeneity responsible for the FS. In particular, for $\beta \sim 1$ as in the cases for model networks and the global COVID-19 time series, this parabolic form of the NV is crucial for understanding systems with a manifestation of a certain extent of the universal GOE behavior in a shorter range, which in turn, however, failed to predict the long-range tail of the NV.

From the results for complex networks and literary texts, we conclude that local statistics, measured by the NND, is robust against external randomness and hence can serve as a characteristic of a studied area, while the long-range statistics, measured by the NV, is rather sensitive to external drives and is favorable for capturing distinctions of different cases inside the area. In linguistics cases, on the one hand, the NND results suggest that text spectra can show level repulsion which is simply the consequence of nonvanishing word lengths (or character stroke numbers); on the other hand, the distinctions between the NVs could be associated with the writing style and the genre of a specific text. For complex networks, the NV results clearly unveiled the structural information of the networks encoded in the correlations of the Laplacian matrix elements. From an FS point of view, the NV values constitute a manifestation of inhomogeneity and finite-size effects of the complex networks, suggesting that the unfolded eigenvalues could aggregate around certain values instead of distributing homogeneously. In COVID-19 time series, the extent of inhomogeneity just reveals the positive correlations in the viral spreading patterns among those different regions. Therefore, in this regard, the U.S. counties are more strongly correlated than between different global countries. In summary, the short- and long-range statistics thus reflect the rigid and elastic features of the systems of interest, and they could be of considerable assistance to data classification, especially when they are utilized in conjunction with other classification characteristics of the systems.

Finally, we remark that we have modeled the disorder using the one-parameter distribution, Eq. (11). It is important to mention that other distributions have already been proposed. In Ref. [66], for instance, another one-parameter distribution was proposed, and a more general one, with three parameters, was discussed in Ref. [19]. It would be interesting to investigate if these families can, in some cases, provide more efficient fittings. Furthermore, we also don't rule out other mechanisms—though not yet known to us at this point—that may introduce a linear term in the NV. It would be quite beneficial if such mechanism also permits a more straightforward derivation for a compact analytic NND expression.

ACKNOWLEDGMENTS

R.X. acknowledges a fellowship from the China Scholarship Council (CSC) under Grant CSC No. 202006770018. M.P.P. was supported by Grant No. 307807/2017-7 of the Brazilian agency Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and is a member of the Brazilian National Institute of Science and Technology-Quantum Information (INCT-IQ). This work was supported in part by the National Natural Science Foundation of China (Grants No. 11505071, 11747135, 11905163), the Programme of Introducing Talents of Discipline to Universities under Grant No. B08033, the Fundamental Research Funds for the Central Universities, and the numerical simulations have been performed on the GPU cluster in the Nuclear Science Computing Center at Central China Normal University (NSC3).

APPENDIX A: NUMBER VARIANCE

To calculate the number variance we start with the expression [2]

$$\langle n^2\rangle_G = \int_{-\theta/2}^{\theta/2} dE_1 \int_{-\theta/2}^{\theta/2} dE_2 R_2(E_1, E_2) + \int_{-\theta/2}^{\theta/2} dE\, \rho_G(E), \tag{A1}$$

for the average of the square of the number of eigenvalues in the interval $(-\frac{\theta}{2}, \frac{\theta}{2})$, where $R_2(E_1, E_2)$ is the two-point function. Introducing disorder, this quantity becomes

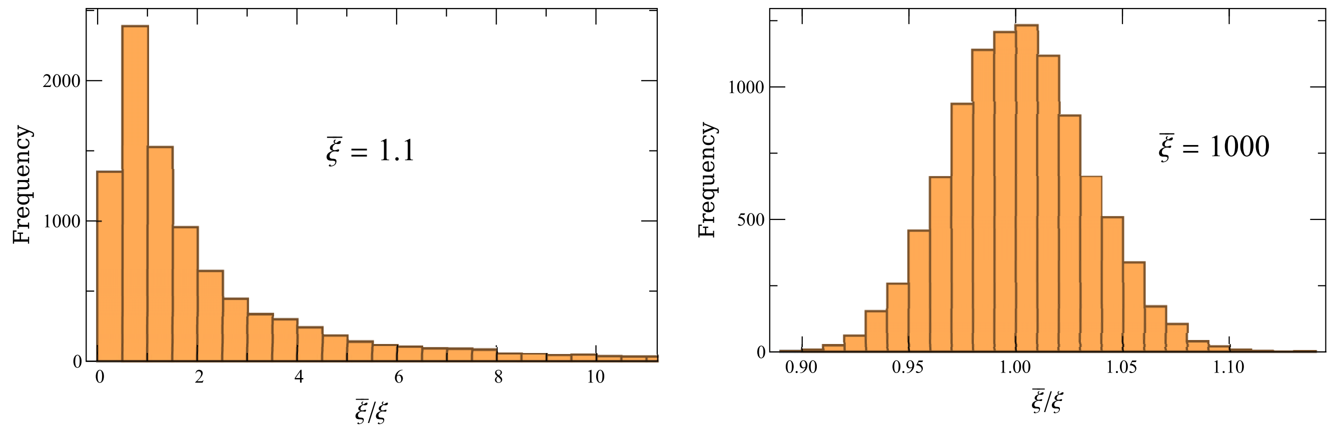
$$\langle n^2\rangle = \int_0^\infty d\xi\, w(\xi)\left[\int_{-\theta/2}^{\theta/2}\int_{-\theta/2}^{\theta/2} dx_1\, dx_2 \frac{\xi}{\bar{\xi}} R_2\left(\sqrt{\frac{\xi}{\bar{\xi}}}x_1, \sqrt{\frac{\xi}{\bar{\xi}}}x_2\right) + \int_{-\theta/2}^{\theta/2} dx \sqrt{\frac{\xi}{\bar{\xi}}}\rho_G\left(\sqrt{\frac{\xi}{\bar{\xi}}}x\right)\right]. \tag{A2}$$

Making in the integrals the substitution of variable $E = \sqrt{\frac{\xi}{\bar{\xi}}}x$, the above expression becomes

$$\langle n^2\rangle = \int_0^\infty d\xi\, w(\xi)\left[\int_{-\sqrt{\frac{\xi}{\bar{\xi}}}\theta/2}^{\sqrt{\frac{\xi}{\bar{\xi}}}\theta/2}\int_{-\sqrt{\frac{\xi}{\bar{\xi}}}\theta/2}^{\sqrt{\frac{\xi}{\bar{\xi}}}\theta/2} dE_1\, dE_2 R_2(E_1, E_2) + \int_{-\sqrt{\frac{\xi}{\bar{\xi}}}\theta/2}^{\sqrt{\frac{\xi}{\bar{\xi}}}\theta/2} dE\, \rho_G(E)\right]. \tag{A3}$$

Changing the variable as $t(E) = \int_0^E dE' \rho_G(E') = N_G(E)$ and using the definition for the two-point cluster function

$$\frac{R(E_1, E_2)}{\rho_G(E_1)\rho_G(E_2)} = 1 - Y(|t_2 - t_1|), \tag{A4}$$



(a)

(b)

FIG. 9. Histograms of $\bar{\xi}/\xi$ for (a) small $\bar{\xi}$ and (b) large $\bar{\xi}$. The distribution of $\bar{\xi}/\xi$ is more concentrated around 1 for large $\bar{\xi}$.

we obtain

$$\langle n^2\rangle = \int_0^\infty d\xi\, w(\xi)\left\{-2\int_0^{N(\sqrt{\frac{\xi}{\bar{\xi}}}\theta/2)} dt\left[1-2N\left(\sqrt{\frac{\xi}{\bar{\xi}}}\frac{\theta}{2}\right)\right]Y(t)+2N_G\left(\sqrt{\frac{\xi}{\bar{\xi}}}\theta/2\right)+4N_G^2\left(\sqrt{\frac{\xi}{\bar{\xi}}}\frac{\theta}{2}\right)\right\}, \tag{A5}$$

where the first two terms inside the curly braces are just the number variance expression of the Gaussian ensemble.

APPENDIX B: ASYMPTOTIC EXPRESSIONS

We are interested in the case of long spectra containing a large number of points. In this situation of very large N, the statistics are measured around the center of the spectra. To be specific, we want to make N to go to infinity, in Eq. (6), keeping the product $\sqrt{N}E$ finite, explicitly

$$N_\beta(E) = \frac{N}{\pi}\left(\arcsin\frac{\sqrt{N}E}{N\sqrt{2\beta}}+\frac{\sqrt{N}E}{N\sqrt{2\beta}}\sqrt{1-\frac{NE^2}{2N^2\beta}}\right)$$
$$\simeq \rho_\beta(0)E. \tag{B1}$$

Assuming now that in Eq. (27) $\xi_{\max}$ is very large and can be replaced by infinity, the disordered cumulative function becomes

$$N_{D\beta}(x) = \int_0^\infty d\xi\, w(\xi)N_\beta\left(\sqrt{\frac{\xi}{\bar{\xi}}}x\right) \simeq \rho_\beta(0)\frac{\overline{\sqrt{\xi}}}{\sqrt{\bar{\xi}}}x, \tag{B2}$$

where (B1) has been used. Therefore, s and L can be approximated in terms of $\rho_\beta(0)\frac{\overline{\sqrt{\xi}}}{\sqrt{\bar{\xi}}}\theta$ in the NND and NV expressions.

Within the same level of approximation we have

$$4\int_0^\infty d\xi\, w(\xi)N_\beta^2\left(\sqrt{\frac{\xi}{\bar{\xi}}}\frac{\theta}{2}\right) \simeq [\rho_\beta(0)\theta]^2 = \left(\frac{\overline{\sqrt{\xi}}}{\sqrt{\bar{\xi}}}\right)^2 L^2. \tag{B3}$$

Finally, if the disorder is defined by the distribution Eq. (11) then we further have

$$\overline{\sqrt{\xi}} = \frac{\Gamma\left(\bar{\xi}+\frac{1}{2}\right)}{\Gamma(\bar{\xi})} \simeq \sqrt{\bar{\xi}}\exp\left(-\frac{1}{8\bar{\xi}}\right), \tag{B4}$$

where the Stirling approximation has been used.

APPENDIX C: THE EFFECT OF THE DISORDER PARAMETER $\bar{\xi}$

For disordered beta ensemble, since $H_D = \bar{\xi}/\xi H_\beta$, the strength of the disorder is determined by the factor $\bar{\xi}/\xi$. As illustrated in Fig. 9, the distribution of $\bar{\xi}/\xi$ is more concentrated around 1 for larger $\bar{\xi}$ [Fig. 9(b)], and the properties of the original ensemble H_β are mostly preserved in the disordered ensemble H_D. Thus this represents a weakly disordered scenario. In contrast, for smaller $\bar{\xi}$ [Fig. 9(a)], the distribution of $\bar{\xi}/\xi$ is more spread so that what are in H_β are bound to be disrupted, introducing a stronger disorder in H_D.

[1] C. S. Porter, *Statistical Theories of Spectra* (Academic Press, New York, 1965).

[2] M. L. Mehta, *Random Matrices*, 3rd ed. (Elsevier, Amsterdam, 2004).

[3] O. Bohigas, M. J. Giannoni, and C. Schmit, Characterization of Chaotic Quantum Spectra and Universality of Level Fluctuation Laws, Phys. Rev. Lett. **52**, 1 (1984).

[4] P. Carpena, P. A. Bernaola-Galván, C. Carretero-Campos, and A. V. Coronado, Probability distribution of intersymbol distances in random symbolic sequences: Applications to improving detection of keywords in texts and of amino acid clustering in proteins, Phys. Rev. E **94**, 052302 (2016).

[5] W. Deng and M. P. Pato, Approaching word length via level spectra, Physica A **481**, 167 (2017).

[6] https://github.com/xiephysics/Generalized_Poisson_ensemble.

[7] T. Guhr, A. Müller-Groeling, and H. A. Weidenmüller, Random matrix theories in quantum physics: Common concepts, Phys. Rep. **299**, 189 (1998).

[8] I. Dumitriu and A. Edelman, Matrix models for beta ensembles, J. Math. Phys. **43**, 5830 (2002).

[9] O. Bohigas, J. X. de Carvalho, and M. P. Pato, Disordered ensemble of random matrices, Phys. Rev. E **77**, 011122 (2008).

[10] O. Bohigas and M. P. Pato, Missing levels in correlated spectral, Phys. Lett. B **595**, 171 (2004).

[11] O. Bohigas and M. P. Pato, Randomly incomplete spectra and intermediate statistics, Phys. Rev. E **74**, 036212 (2006).

[12] P. G. de Gennes, Soft matter, edited by G. Ekspong, *Nobel Lectures, Physics 1991-1995* (World Scientific Publishing Co., Singapore, 1997).

[13] O. Bohigas, Random matrix theories and chaotic dynamics, in *Chaos et Physique Quantique (Chaos and Quantum Physics), Proceedings of the Les Houches Summer School, Session LII (1989)*, edited by M. J. Giannoni, A. Voros, and J. Zinn-Justin (North-Holland, Amsterdam, 1991).

[14] O. Bohigas and M. P. Pato, Decomposition of spectral density in individual contributions, Phys. A: Math. Theor. **43**, 365001 (2010).

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

[16] F. Toscano, R. O. Vallejos, and C. Tsallis, Random matrix ensembles from nonextensive entropy, Phys. Rev. E **69**, 066131 (2004).

[17] A. C. Bertuola, O. Bohigas, and M. P. Pato, Family of generalised random matrix ensembles, Phys. Rev. E **70**, 065102(R) (2004).

[18] K. A. Muttalib and J. R. Klauder, Family of solvable generalised random-matrix ensembles with unitary symmetry, Phys. Rev. E **71**, 055101(R) (2005); A. Y. Abul-Magd, Nonextensive random matrix theory approach to mixed regular-chaotic dynamics, *ibid.* **71**, 066207 (2005).

[19] O. Bohigas and M. P. Pato, Hyberbolic disordered ensemble of random matrices, Phys. Rev. E **84**, 031121 (2011).

[20] M. P. Pato, Disordered random matrices, J. Phys.: Conf. Ser. **604**, 012015 (2015).

[21] M. P. Pato, Disordered random walks, Braz. J. Phys. **51**, 238 (2021).

[22] L. R. Taylor, Aggregation, variance and the mean, Nature (London) **189**, 732 (1961).

[23] Z. Eisler and J. Kertész, Scaling theory of temporal correlations and size dependent fluctuations in the traded value of stocks, Phys. Rev. E **73**, 046109 (2006).

[24] Z. Eisler, I. Bartos and J. Kertász, Fluctuation scaling in complex systems: Taylor's law and beyond, Adv. Phys. **57**, 89 (2008).

[25] M. A. de Menezes and A.-L. Barabási, Fluctuations in Network Dynamics, Phys. Rev. Lett. **92**, 028701 (2004).

[26] D. J. Daley and D. Vere-Jones, *An Introduction to the Theory of Point Processes Vol. II General Theory and Structure*, 2nd ed., Probability and Its Applications (Springer, New York, 2008).

[27] J. Chen, K. Zhang, and J. Zhong, Identifying functional modules of diffuse large B-cell Lymphoma gene co-expression networks by hierarchical clustering method based on random matrix theory, Nano Biomed. Eng. **3**, 57 (2011).

[28] P. Deift, Some open problems in random matrix theory and the theory of integrable systems I. Symmetry, integrability and geometry: Methods and applications, SIGMA **13**, 016 (2017).

[29] T. Berggren and M. Duits, Mesoscopic fluctuations for the thinned Circular Unitary Ensemble, Math. Phys. Anal. Geom. **20**, 19 (2017).

[30] A. Grabsch, S. N. Majumdar, and C. Texier, Truncated linear statistics associated with the eigenvalues of random matrices II. Partial sums over proper time delays for chaotic quantum dots, J. Stat. Phys. **167**, 1452 (2017).

[31] T. Bothner and R. Buckingham, Large deformations of the Tracy-Widom distribution I: Non-oscillatory asymptotics, Commun. Math. Phys. **359**, 223 (2018).

[32] V. Plerou, P. Gopikrishnan, B. Rosenow, L. A. Amaral, T. Guhr, and H. E. Stanley, Random matrix approach to cross correlations in financial data, Phys. Rev. E **65**, 066126 (2002).

[33] P. Šeba, Random Matrix Analysis of Human EEG Data, Phys. Rev. Lett. **91**, 198104 (2003).

[34] M. S. Santhanam and P. K. Patra, Statistics of atmospheric correlations, Phys. Rev. E **64**, 016102 (2001).

[35] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D. U. Hwang, Complex networks: Structure and dynamics, Phys. Rep. **424**, 175 (2006).

[36] M. E. J. Newman, *Networks*, 2nd ed. (Oxford University Press, Oxford, 2018).

[37] P. Erdős and A. Rnéyi, On the evolution of random graphs, Publ. Math. Inst. Hung. Acad. Sci. **5**, 17 (1960).

[38] D. J. Watts and S. H. Strogatz, Collective dynamics of 'small-world' networks, Nature (London) **393**, 440 (1998).

[39] A. L. Barábsi and R. Albert, Emergence of scaling in random networks, Science **286**, 509 (1999).

[40] J. N. Bandyopadhyay and S. Jalan, Universality in complex networks: Random matrix analysis, Phys. Rev. E **76**, 026109 (2007).

[41] S. Jalan and J. N. Bandyopadhyay, Random matrix analysis of complex networks, Phys. Rev. E **76**, 046107 (2007).

[42] S. Jalan and J. N. Bandyopadhyay, Random matrix analysis of network Laplacians, Physica A **387**, 667 (2008).

[43] S. Jalan and J. N. Bandyopadhyay, Randomness of random networks: A random matrix analysis, Europhys. Lett. **87**, 48010 (2009).

[44] F. R. K. Chung, *Spectral Graph Theory*, CBMS Regional Conference Series in Mathematics **92** (American Mathematical Society, Providence, RI, 1997).

[45] C. Sarkar and S. Jalan, Social patterns revealed through random matrix theory, Europhys. Lett. **108**, 48003 (2014).

[46] R. Wang, Z. Z. Zhang, J. Ma, Y. Yang, P. Lin, and Y. Wu, Spectral properties of the temporal evolution of brain network structure, Chaos **25**, 123112 (2015).

[47] G. S. Matharoo and J. A. Hashmi, Spontaneous back-pain alters randomness in functional connections in large scale brain networks: A random matrix perspective, Physica A **541**, 123321 (2020).

[48] M. Krbálek and P. Seba, The statistical properties of the city transport in Cuernavaca (Mexico) and random matrix ensembles, J. Phys. A: Math. Theor. **33**, L229 (2000).

[49] A. Singh and D. Xu, Random matrix application to correlations amongst the volatility of assets, Quant. Finance **16**, 69 (2016).

[50] S. H. Lee, M. D. Fricker, and M. A. Porter, Mesoscale analyses of fungal networks as an approach for quantifying phenotypic traits, J. Complex Netw. **5**, 145 (2017).

[51] More analyzed data for fungal networks and texts, our codes, and the fitting results are available at https://github.com/xiephysics/Disordered_beta_thinned_ensemble_with_applications.

[52] D. P. Bebber, J. Hynes, P. R. Darrah, L. Boddy, and M. D. Fricker, Biological solutions to transport network design, Proc. R. Soc. B **274**, 2307 (2007).

[53] R. Mittal, R. Ni, and J. H. Seo, The flow physics of COVID-19, J. Fluid Mech. **894**, F2 (2020).

[54] M. Perc, M. N. Gorišek, M. Slavinec, and A. Stožer, Forecasting COVID-19, Front. Phys. **8**, 127 (2020).

[55] A. Vespignani, H. Tian, C. Dye, J. O. Lloyd-Smith, R. M. Eggo, M. Shrestha, S. V. Scarpino, B. Gutierrez, M. U. Kraemer, J. Wu, and K. Leung, Modelling COVID-19, Nat. Rev. Phys. **2**, 279 (2020).

[56] J. R. Gog, How you can help with COVID-19 modelling, Nat. Rev. Phys. **2**, 274 (2020).

[57] C. Tsallis and U. Tirnakli, Predicting COVID-19 peaks around the world, Front. Phys. **8**, 217 (2020).

[58] S. Chaudhuri, S. Basu, P. Kabi, V. R. Unni, and A. Saha, Modeling the role of respiratory droplets in Covid-19 type pandemics, Phys. Fluids **32**, 063309 (2020).

[59] V. Plerou, P. Gopikrishnan, B. Rosenow, L. A. Amaral, and H. E. Stanley, Universal and Nonuniversal Properties of Cross Correlations in Financial Time Series, Phys. Rev. Lett. **83**, 1471 (1999).

[60] H. K. Pharasi, K. Sharma, A. Chakraborti, and T. H. Seligman, Complex market dynamics in the light of random matrix theory, in *New Perspectives and Challenges in Econophysics and Sociophysics*, edited by F. Abergel, B. K. Chakrabarti, A. Chakraborti, N. Deo, and K. Sharma (Springer, Berlin, 2019), pp. 13–34.

[61] Center for Systems Science and Engineering, John Hopkins University, https://github.com/CSSEGISandData/COVID-19, COVID-19 Data Repository, by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University.

[62] M. Kieburg and A. Monteleone, Local tail statistics of heavy-tailed random matrix ensembles with unitary invariance, J. Phys. A: Math. Theor. **54**, 325201 (2021).

[63] Y. Haralambous, Seeking meaning in a space made out of strokes, radicals, characters and compounds, arXiv:1104.4321.

[64] Romanization and Radical/Stroke Table of Unihan Characters, by HKUST Library for the HKIUG Unicode Task Force, https://hkiug-archive.lib.hku.hk/unicode/hkiug-roman-radical-stroke-1.0.html.

[65] To illustrate this point, we observe that if a spectrum is constructed by generating a sequence of points from the Wigner surmise, the NND of course will be Wigner but the NV is $L/4$.

[66] A. Y. Abul-Magd, G. Akemann, and P. Vivo, Superstatistical generalizations of Wishart-Laguerre ensembles of random matrices, J. Phys. A: Math. Theor. **42**, 175207 (2009).