Stochastic dynamical model for stock-stock correlations

Wen-Jong Ma and Chin-Kun Hu*
Institute of Physics, Academia Sinica, Nankang, Taipei 11529, Taiwan

Ravindra E. Amritkar
Physical Research Laboratory, Navrangpura, Ahmedabad 380009, India
(Received 2 October 2003; revised manuscript received 1 January 2004; published 4 August 2004)

We propose a model of *coupled random walks* for stock-stock correlations. The walks in the model are coupled via a mechanism that the displacement (price change) of each walk (stock) is activated by the price gradients over some underlying network. We assume that the network has two underlying structures, describing the correlations among the stocks of the whole market and among those within individual groups, respectively, each with a coupling parameter controlling the degree of correlation. The model provides the interpretation of the features displayed in the distribution of the eigenvalues for the correlation matrix of real market on the level of time sequences. We verify that such modeling indeed gives good fitting for the market data of US stocks.

DOI: 10.1103/PhysRevE.70.026101 PACS number(s): 89.65.Gh, 05.40.Fb

In recent decades, the analysis of many physical and social systems has been based on the idea that randomness in the fluctuations is hallmarked by certain prototypes, such as eigenvalue distribution of a random matrix, and any deviation from the latter is a result of the presence of correlations. Such an idea has been used to study the level statistics [1] of locally activated states [2] for electrons in heavy atoms or in solids, atomic vibrations in spatially disordered systems, distribution spectra of eigenvalues for the correlation matrices of stocks [3,4], internet traffic [5], and atmospheric fluctuations [6]. While recent advances [3–6] suggest a robust approach to reveal cross correlations from the data of time sequences, a comprehensive address of the shared dynamics behind these different systems, however, is still lacking. In the present paper, we will address this problem by proposing a stochastic dynamic model for stock-stock correlations [3,4]. Our approach is useful for studying time evolution of other interacting many-body systems subject to random noises.

The nature of fluctuations in financial markets [3,4,7–11] has been of interest to the traders as well as a variety of professionals for more than a century. If such fluctuations could be completely characterized by random walks, as first proposed by Bachelier in 1900 [9], making profit under controlled risks through the transactions in the markets would seem impractical. Correlations in such fluctuations have been demonstrated in recent studies [3,4] of the distribution spectrum of eigenvalues of the cross correlation matrix for the price changes of stocks in real markets. The matrix measures the statistical overlap of the fluctuations [3,4,11,12] in the price changes (the returns) between pairs of stocks. The spectrum from market data [3,4] possesses a bulk of continuously distributed eigenvalues, which is similar to the prototype predicted by the level statistics of the random matrix theory (RMT) [1,4,13] and may be considered as mainly contributed by the randomness. The effects of the correlations [3,4,11] manifest in the eigenvectors of those eigenvalues isolated from the bulk, which include one large eigenvalue λ_M, corresponding to the correlation among all stocks (*market mode*) [4] and several much smaller ones scattered in between λ_M and the bulk component. The patterns in the eigenvectors of these latter eigenvalues were related to the presence of groups of correlated stocks [4,11]. In addition, the bulk component also shows some important deviations from RMT predictions.

The spectrum of eigenvalues and the corresponding eigenvectors are related to the cooperative behavior in the fluctuations of the stock prices, which is not visible in the local information of the individual correlation coefficients composing that matrix [14]. Based on former information, the connections between the deviated eigenvalues and the presence of correlated groups of stocks have been clarified [4] and rewarded with an ansatz [11] which was applied to modeling the real market data [15]. In the model [11,15], the return of each stock is linearly decomposed into two uncorrelated fluctuating parts. The interdependence of stocks within a group is carried by the part which is synchronously shared by all stocks within that group. This formulation refines the conventional multifactor model [8] leading to blocked structures in its correlation matrix [11], which reproduces those spectral features observed in market data [15]. It assumes that each of the isolated eigenvalues from the bulk is contributed by one block of submatrix, with its eigenvector containing only one activated group. This puts a limitation on the model, not able to digest completely the information carried by the eigenvectors for those deviated eigenvalues in the market data, in which the shared activated stocks are present very often among different eigenmodes. Our paper presents a new formulation to include this fact, at the same time, retaining the realization of grouping in its simplest form. Our approach is a general formulation which describes the cooperative activities in the financial fluctuations beyond the statistics of matrices, on the level of time sequences.

*Electronic address: huck@phys.sinica.edu.tw

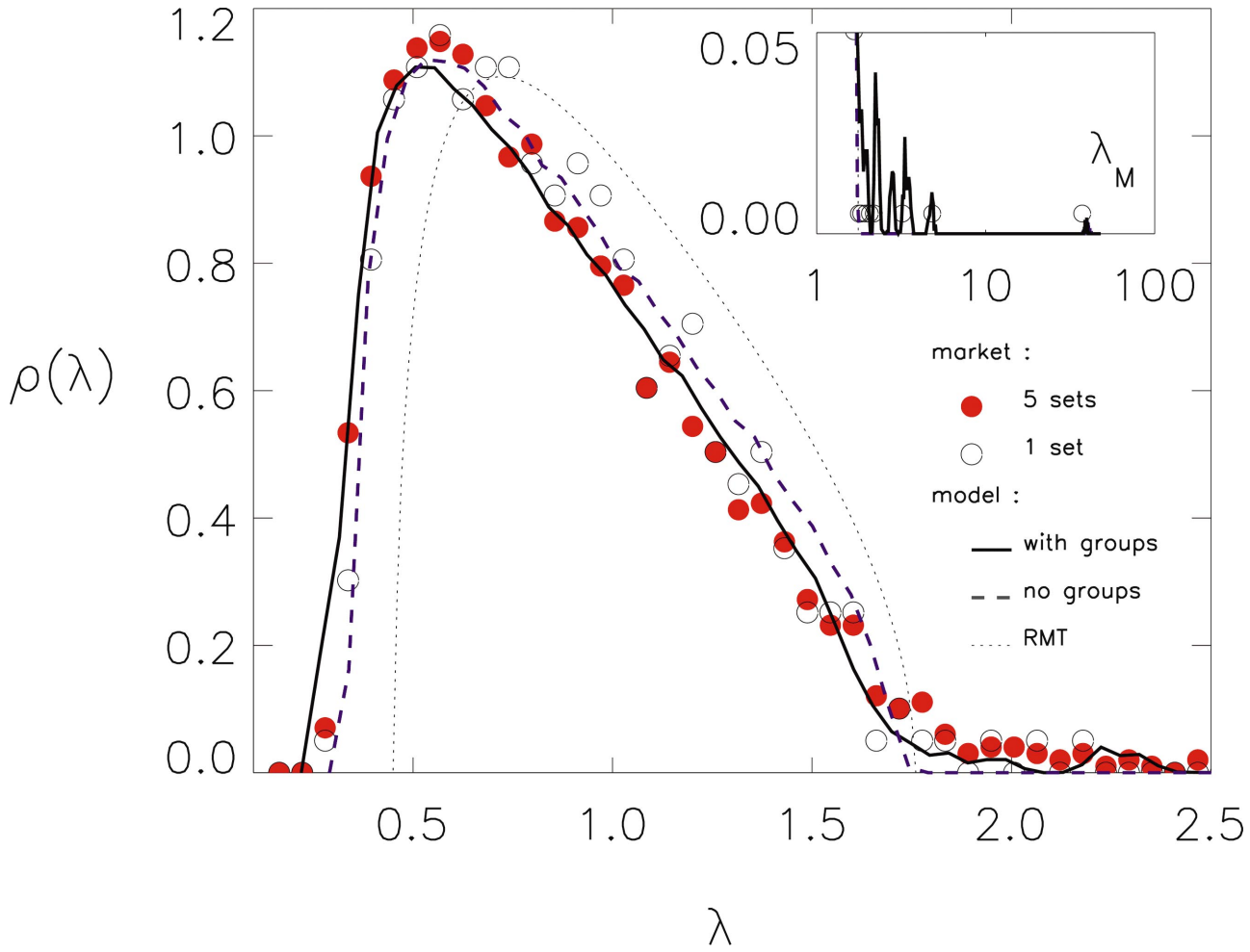

FIG. 1. (Color) The simulated eigenvalue distribution $\rho(\lambda)$ with $\epsilon_g \neq 0$ (thick solid line) and $\epsilon_g = 0$ (dashed line) with $N = 345$ stocks and $T = 3250$ steps, in comparison with the market data (open and filled circles, see text) and the prediction of the random matrix theory [13] (dotted line). The inset shows large isolated eigenvalues, including λ_M. The fit (thick solid line) in the figure is obtained with $\epsilon_M = 0.794$, 8 major groups, each occupied by more than 10 stocks, and 17 small groups. The number of stocks n_g in each group and the corresponding couplings ϵ_g are listed in Table I. The simulated data are obtained by averaging over an ensemble of 106 matrices.

We formulate a stochastic dynamic model, called *coupled random walks*, for the stock-stock correlations based on the assumption that the changes in the prices of the stocks are due to the adjustment to eliminate the difference between the expected price for each individual stock and that for a collection of related stocks. The correlations are imposed as the underlying network over which the price-balancing processes are carried out. We assume two underlying structures built on the network. One includes the stocks of the whole market, over which the process underscores the presence of the market mode [4]. The other one realizes the formation of groups, which accounts for the presence of the other deviated eigenvalues. The partitioning of groups and the strengths of the correlations over each of the two structures are determined to reflect the degree and the extent of comovement for the stocks, indicated by the location of the largest eigenvalue λ_M and by the information contained in the eigenvectors corresponding to those deviated eigenvalues (see latter text). The comparison between the spectra of our model to be introduced below and the US stocks under such analysis is presented in Fig. 1 which shows very good agreement.

I. STOCHASTIC DYNAMIC MODEL

Our model consists of a system of N walks (corresponding to N stocks), and let $x_i(t)$, $i = 1, 2, \ldots, N$, denote the position (price) of the ith walk (stock) at time t. A random walk without correlations can be written as

$$x_i(t+1) = x_i(t) + \xi_i(t),$$

where $\xi_i(t)$ is a random noise. We introduce correlations by expressing the position at time $t+1$ by

$$x_i(t+1) = (1 - \epsilon_M - \epsilon_g) f_i(x_i(t)) + \frac{\epsilon_M}{N} \sum_{j=1}^{N} f_j(x_j(t)) + \frac{\epsilon_g}{n_{g\,k \in g}} \sum f_k(x_k(t)),$$

$$\text{if } i \in g \text{ and } g \in G, \quad (\epsilon_M + \epsilon_g \leqslant 1.0), \qquad (1)$$

where ϵ_M and ϵ_g are coupling constants (subscripts "M" and "g" denote "market" and "group g," respectively) and G is a

class of nonoverlapping subsets of the system of N stocks. We choose the function $f_i(x_i(t))$ as

$$f_i(x_i(t)) = x_i(t) + \xi_i(t). \tag{2}$$

The second term in Eq. (1) corresponds to a coupling to the market mean field determined by all the walks (stocks). The last term corresponds to the coupling to the mean field of the group g to which the walk i belongs and the summation is over all n_g members of the group; for the stock which does not belong to any group, there is no such term. The type of coupling proposed in Eq. (1) is commonly used in coupled maps where the function f_i is usually a nonlinear map [16]. Though the model treats each $\xi_i(t)$ as an uncorrelated noise, while comparing with the actual market data it may be considered as an integrated effect of the random fluctuation between the two discrete times t and $t+1$. In our numerical simulations, ξ's are taken as temporally and mutually uncorrelated Gaussian random values with zero mean and variance σ^2

$$\langle \xi_i(t)\xi_j(t')\rangle = \sigma^2 \delta_{ij}\delta_{tt'} \quad (i,j = 1, \ldots, N), \tag{3}$$

where the averaging $\langle \cdots \rangle$ is over the statistical ensemble.

Equation (1) describes a concise model covering various cases, from the totally random situation ($\epsilon_M = \epsilon_g = 0.0$) to the case where all stocks are fully correlated ($\epsilon_M = 1.0, \epsilon_g = 0.0$). It can be written as an equation of continuity,

$$\mathbf{r}(t+1) = \mathbf{x}(t+1) - \mathbf{x}(t) = \xi(t) + \mathbf{\Delta}(\mathbf{x}(t) + \xi(t)) \tag{4}$$

where $\mathbf{r}$, $\mathbf{x}$, and ξ are the column vectors containing the displacements (the returns), the positions and the noises of the N walks, respectively. $\mathbf{\Delta}$ is a Laplace-type operator describing the flows due to the presence of gradients over an underlying network. The displacements of the walks are, therefore, governed by the mechanism over the underlying network tending to eliminate the ever-changing unbalance arisen from the incoming random pulses. In our mean-field formulation, we assume

$$\Delta = \mathbf{\Delta}_M + \mathbf{\Delta}_G, \tag{5}$$

where the operators $\mathbf{\Delta}_M$ and $\mathbf{\Delta}_G$, defined as the $N \times N$ matrices with, respectively, matrix elements

$$(\mathbf{\Delta}_M)_{ij} = \frac{\epsilon_M}{N} \begin{cases} (1-N) & \text{if } i = j \\ 1 & \text{otherwise} \end{cases}$$

and

$$(\mathbf{\Delta}_G)_{ij} = \begin{cases} \dfrac{\epsilon_g}{n_g}(1-n_g) & \text{if } i = j \in g \text{ and } g \in G \\ \dfrac{\epsilon_g}{n_g} & \text{if } i \neq j; i,j \in g \text{ and } g \in G \\ 0 & \text{otherwise,} \end{cases} \tag{6}$$

which couple to the collections of flows coming to the *ith* site, contributed by the whole system and by the individual groups, respectively. The parameters ϵ_M and ϵ_g are then realized as the kinetic coefficients controlling the magnitudes of such flows.

The information of cross correlation is contained in the matrix $\mathbf{C}$ of cross correlation coefficients C_{ij}, which is defined as the statistical overlap of the fluctuations $\delta r_i = r_i - \langle r_i \rangle_T$ between the two stocks i and j, that is,

$$C_{ij} = \frac{\langle \delta r_i \delta r_j \rangle_T}{\sigma_i \sigma_j}, \tag{7}$$

where $\sigma_i^2 = \langle (\delta r_i)^2 \rangle_T$. The average $\langle \cdot \rangle_T$ is over a time series of T events. Each eigenvector (eigenmode) of $\mathbf{C}$ describes one possible way of activation for the stock fluctuations, with the magnitude of the corresponding eigenvalue measuring its contribution to the fluctuations [17].

II. MARKET MODE

When $\epsilon_M = 1$ and $\epsilon_g = 0$, all the entries of the matrix $\mathbf{C}$ are 1. The matrix possesses one eigenvalue equal to the number of stocks in the market and others are zero. As the correlations decrease, the single large eigenvalue remains isolated from the bulk of dispersed eigenvalues. This roughly explains the origin of the two major components found in the spectrum from the market data, the bulk and the market mode λ_M. The fitting in Fig. 1 suggests the dominant role of the market coupling (ϵ_M close to 1, see the following text).

In the following, we derive the explicit expressions for matrix $\mathbf{C}$ and find analytically the dependence of the eigenvalue λ_M on ϵ_M, for the case with no group couplings. In this case ($\epsilon_g = 0$), the returns in Eq. (4) can be written as [18],

$$\mathbf{r}(t) = \xi(t-1) + \mathbf{\Delta}_M (1-\epsilon_M)^{t-1-t_0} \mathbf{x}(t_0) + \mathbf{\Delta}_M \sum_{s=t_0}^{t-1} (1-\epsilon_M)^{t-1-s} \xi(s). \tag{8}$$

Here we assume the system, with all walks initially at the same position $x_i(0) = 0$, has been statistically steady at time t_0 so that the data can be collected starting from $t = t_0 + 1$, for the evaluation of correlation matrix $\mathbf{C}$. Denoting the $N \times T$ matrices $\Xi \equiv (\xi(t_0), \xi(t_0+1), \ldots, \xi(t_0+T-1))$ and $\mathbf{R} \equiv [\mathbf{r}(t_0+1), \mathbf{r}(t_0+2), \ldots, \mathbf{r}(t_0+T)]$ for the sequences of noises and returns, respectively; and introducing the $T \times T$ matrix Π by

$$\Pi_{vw} = \begin{cases} (1-\epsilon_M)^{w-v} & \text{if } w \geq v \\ 0 & \text{otherwise,} \end{cases} \tag{9}$$

Eq. (8) for $t = t_0+1, t_0+2, \ldots, t_0+T$ can be summarized as

$$\mathbf{R} = \Xi + \mathbf{\Delta}_M (\mathbf{X}_0 + \Xi)\Pi, \tag{10}$$

where $\mathbf{X}_0$ is the $N \times T$ matrix carrying the information of initial conditions, with vector $\mathbf{x}(t_0)$ in its first column and zeros elsewhere. With properly chosen σ^2 for the random noises [Eq. (3)] to assure the unity variances in Eq. (7), the correlation matrix can be written as

$$\mathbf{C} \approx \frac{1}{T} \mathbf{R}\mathbf{R}^{\mathbf{T}} \tag{11}$$

in the large T limit (see Appendix). From Eqs. (10) and (11), we can see that the cross correlation in our model is due to the statistical overlap managed by *both* the site-wise operator

$\mathbf{\Delta}_M$ and the time-wise operator Π. The presence of such combination between the structural and the temporary effects is true for the general case $\mathbf{\Delta}_G \neq 0$ [19]. As a result, the stocks activated in the eigenmodes deviated from bulk are not restricted to one group for each eigenmode (see the following text for the numerical results). This is in contrast to the model proposed in Ref. [11].

Now, we consider the ensemble average $\langle R_{it}R_{jt}\rangle$, which describes the cross correlation between the two walks i and j at time t and the mean over T time steps of which gives the ensemble averaged correlation coefficient [from Eq. (11)]

$$\langle C_{ij}\rangle = \frac{1}{T}\sum_{t=t_0+1}^{t_0+T}\langle R_{it}R_{jt}\rangle. \tag{12}$$

The quantity can be evaluated analytically by using Eq. (3). For $t \gg 1$, we have (see Appendix)

$$\langle R_{it}R_{jt}\rangle = \sigma^2\left[\delta_{ij}\{1-a(\epsilon_M)\} + \frac{a(\epsilon_M)}{N}\right], \tag{13}$$

where $a(\epsilon_M)=\epsilon_M(3-2\epsilon_M)/(2-\epsilon_M)$ changes monotonically from 0 to 1 as ϵ_M increases from 0 to 1. Note that the correlation between a pair of different walks is $a(\epsilon_M)$, diluted by the size factor $1/N$. From Eqs. (12) and (13), we obtain

$$\langle C_{ij}\rangle = \begin{cases} \dfrac{a(\epsilon_M)}{(1-a(\epsilon_M))N + a(\epsilon_M)} & \text{if } i \neq j \\ 1 \quad \text{if } i = j, \end{cases} \tag{14}$$

for the mean of distribution for each entry over the ensemble of matrices. We can see from Eqs. (13) and (14) that, in the large size limit, the local correlations between a pair of different walks become diminishing. On the other hand, we show in the following that the divergence due to collective activities is present for the global quantity λ_M.

Consider the matrix $\mathbf{C}^* \equiv \langle \mathbf{C}\rangle$ defined by Eq. (14). Its largest eigenvalue λ_M^* is [20]

$$\lambda_M^* = \frac{N}{[\{1-a(\epsilon_M)\}N + a(\epsilon_M)]}, \tag{15}$$

which diverges at $\epsilon_M=1$, as $N\to\infty$. In the large N limit, we can write Eq. (15) as

$$\lambda_M^* \approx [1-a(\epsilon_M)]^{-1} = \frac{2-\epsilon_M}{2(1-\epsilon_M)^2}, \tag{16}$$

which diverges in a power law $\lambda_M^* \propto (1-\epsilon_M)^{-2}$, as $\epsilon_M \to 1$. We found that Eq. (15) describes also the market mode λ_M for the ensemble of matrices, for ϵ_M near 1. Figure 2 verifies in simulation that the presence of such divergence for λ_M, in a power law with exponent 2 as $\epsilon_M \to 1$. The result is similar to critical phenomena in that the divergent fluctuations emerge for a system approaching to the (critical) state with full correlations [20]. This critical state in our model is characterized by a finite-size scaling relationship $\lambda_M^*(\tilde{\epsilon}, N) = \tilde{\epsilon}^{-2}\Lambda(\tilde{\epsilon}N^{1/2})$ in the critical region [21], where $\tilde{\epsilon} \equiv 1-\tilde{\epsilon}_M$ and $\Lambda(x) \equiv x^2/(2x^2+1)$.

III. SPECTRA OF CROSS CORRELATION MATRIX

We now compare the results of our model with the spectra of the actual market data [22]. Figure 1 shows the simulated eigenvalue distribution for the cross correlation matrix for a system of $N=345$ coupled walks (corresponding to 345 stocks) obtained by averaging over $T=3250$ steps. The figure also shows the RMT prediction [23] and the results for 345 US stocks obtained with 3250 values of the 30 min returns in the year 1996 for 250 days [24] and 6.5 trading hours each day. In Fig. 1, the open circles are for one data set starting from 9:30 a.m. while filled circles are ensemble average of five data sets starting at different times from 9:30 to 10:00 a.m. We see from the figure that our model of coupled random walks gives an excellent fit for the actual market data [25] and accounts for the three important deviations from the RMT predictions, namely, a very large eigenvalue (market mode), several eigenvalues larger than RMT upper limit but much less than the market mode, and significant deviations of the bulk from RMT prediction.

The fitting of the market data to the model includes the dividing of the stocks into groups and the estimation of the coupling parameters. Without exploring the sophisticated algorithms for partitioning [26] and fitting [15], we investigate the feasibility of a procedure of grouping based on the integrated information carried by the eigenmodes of the correlation matrix and the fluctuation properties of the time sequences of each group. The grouping obtained in this approach may not be necessary identical with the division by the market sectors or by the maximum-likelihood fitting [15,27]. After all, the grouping for the cooperative activities signaled by eigenmodes (see following paragraph) are affected by many dynamic factors, such as the trading behavior of the stock owners, and may be different from those based on static interpretations [14] of the correlation data.

The parameters of the model used for the fit in Fig. 1 are determined as follows. The coupling ϵ_M is estimated by Eq. (15) using the market mode λ_M obtained from the correlation matrix $\mathbf{C}$ and is amended with the determination of the rest parameters. The number of groups and the number of stocks n_g in each group g are determined by finding the comovements of the stocks in the eigenvectors corresponding to the K eigenvalues in between the bulk component and the market mode. The procedure is as follows. The stocks are considered as in the same group in the mean-field scenario if they are activated correspondingly in all K eigenvectors. We define a base noise level corresponding to the fluctuations in the components of the eigenvector of the market mode. In the components of the K eigenvectors a component (stock) is treated as active if it has a magnitude larger than this base noise level otherwise its contribution is neglected. To determine whether two stocks are in the same group, we consider two K-dimensional vectors composed of the corresponding components in the K eigenmodes and take the absolute value of the cosine of the angle between these two vectors. If this absolute value is larger than some critical value β then the two stocks are coupled. After we get all the couples we classify them into groups. Two groups are merged if they have one or more overlapping elements and the direction of the

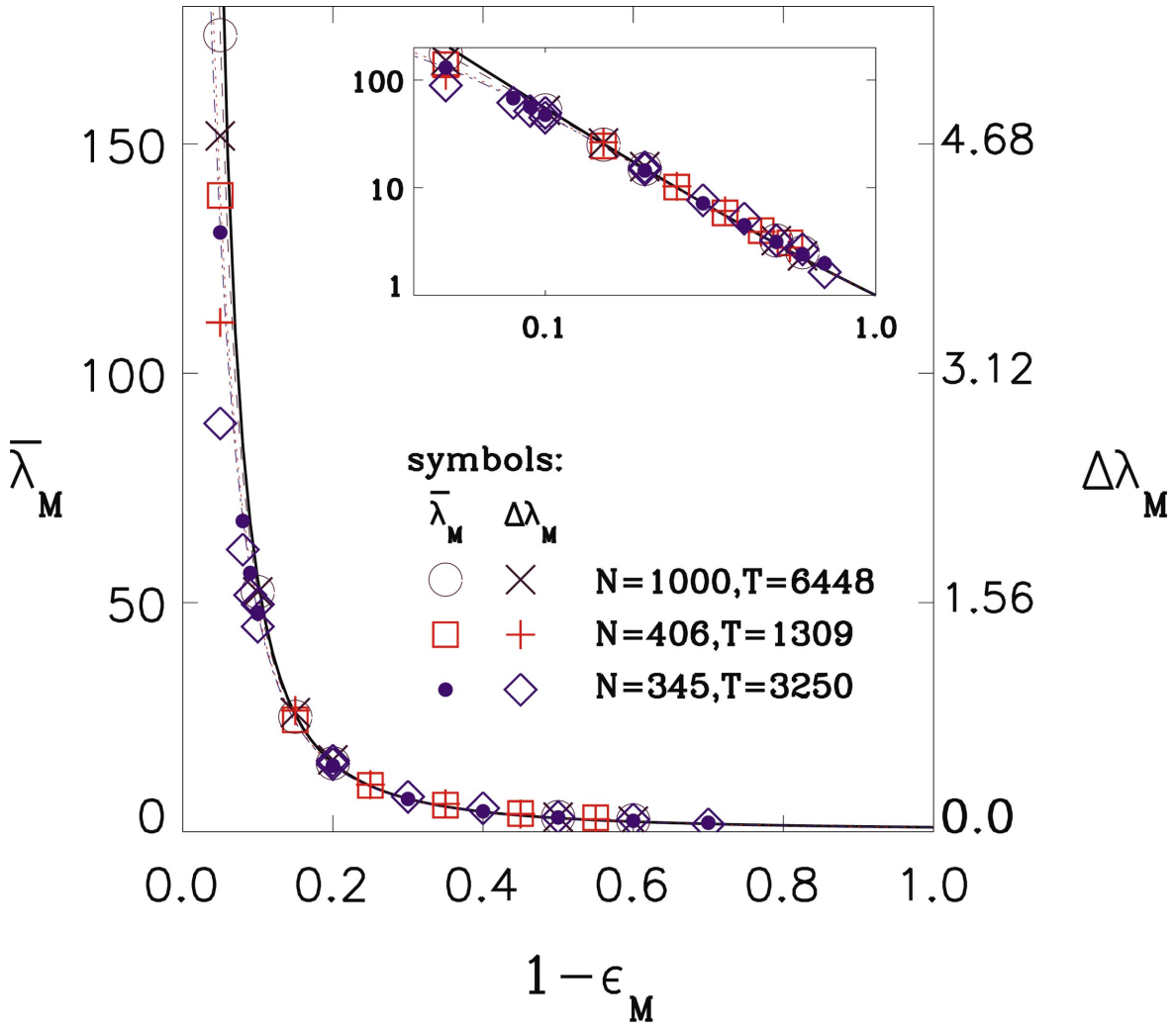

FIG. 2. (Color) The plot of the ensemble averaged value $\overline{\lambda}_M$ and the width $\Delta\lambda_M$ of the distribution of λ_M in the ensemble corresponding to the *market mode* for various values of $\epsilon_M(\epsilon_g=0.0)$ for three different systems with $N=(345,406,1000)$ walks, lengths $T=(3250,1309,6448)$, and over ensembles of $(106,400,50)$ cross correlation matrices, respectively. [These are the same N and T values as the pools of stocks analyzed in this study (Fig. 1), Refs. [3] and [4], respectively.] The corresponding curves for the analytical expression, Eq. (15), for finite N (thin broken, dotted, and dashed lines are for $N=345$, 406, and 1000, respectively) are plotted for comparison. The series of curve approaches in increasing N to that of Eq. (16) (solid line), which is a power law with exponent 2 for $\epsilon_M \approx 1$. The inset shows the same data in log-log plot.

overlapped part is within the deviation of each group in all the K directions. The critical value $\beta \approx 0.835$ is determined such that the final number of major groups after merging is K (or close to K). The coupling ϵ_g for each group g can be estimated by matching the ϵ_g-dependent properties between the market and the model [28]. In Table I, we list the parameters $\{n_g\}$ and $\{\epsilon_g\}$ for the groups obtained in fitting to our model.

TABLE I. Fitted grouping parameters.

Group index g	A^*	B^*	C^*	D^*	E^*	F^*	G^*	H^*	I	J	K	L	M
n_g	70	62	30	29	26	17	14	12	8	7	6	5	4
$\epsilon_g \times 10^3$	104	139	111	44.6	9.25	46.9	13.1	98.7	61.4	135	95.2	41.3	0.06
Group index g	N	O	P	Q	R	S	T	U	V	W	X	Y	
n_g	4	3	3	2	2	2	2	2	2	2	2	2	
$\epsilon_g \times 10^3$	0.11	123	7.09	2.14	78.6	0.91	17.7	11.4	94.4	1.60	1.76	62.0	

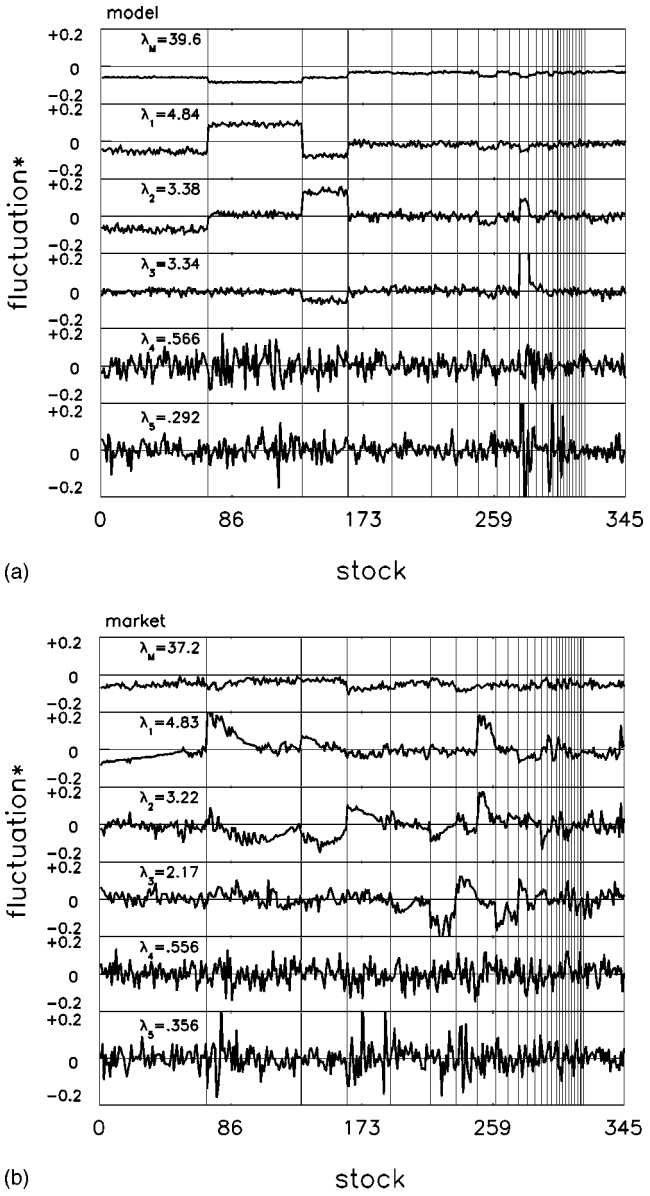

FIG. 3. The components (scaled fluctuations of returns) for the eigenvectors of a few typical eigenvalues corresponding to our model (a) and to the market data (b) as described in Fig. 1. They are the eigenvectors of (top down) the market mode λ_M, the next three largest eigenvalues (λ_1,λ_2,λ_3) and two typical modes with their eigenvalues falling inside the bulk (λ_4) and at the lower edge of the bulk component (λ_5), respectively. The stocks are numbered so that the stocks belonging to the same group are consecutively placed and the groups are arranged in the plots, from left to right, according to their sizes in the descending order. There are 27 stocks placed at the end that do not belong to any group. The vertical lines indicate the boundaries of the groups and the horizontal lines mark the zero levels.

In Fig. 3, we plot the eigenvectors of a few typical modes obtained for the model and for the 345 US stocks described in Fig. 1. It shows the presence of several simultaneously activated groups in each of the eigenmodes deviated from bulk [eigenvectors corresponding to λ_1, λ_2, and λ_3 in Figs. 3(a) and 3(b)], for both the market data [Fig. 3(b)] and the simulation [Fig. 3(a)]. Such situation, however, cannot be directly interpreted in a model based on the blocked-structured correlation matrix **C** [11,15,19]. In our model, the groupings cause not only the deviation for the eigenvalues in between market mode and the continuous bulk component, but also modify the bulk leading to a shift and an extension of the distribution at the lower edge (Fig. 1). The eigenvectors corresponding to the eigenvalues at lower edge possess very different patterns compared with those (discretely) distributed above the upper edge of the bulk. The differences are easily demonstrated in our model. [Compare the eigenvectors corresponding to λ_5 and λ_3 in Fig. 3(a).] The stocks in the latter modes are coherently activated in groups, while only individual stocks are activated with large amplitudes in the former modes [29]. Such differences can also be recognized in the market data. [Compare those for λ_5 and λ_3 in Fig. 3(b).]

It is worth to note that, in our model, the grouped activities are realized as the perturbing parts to the market activities, as a result of the large magnitude in λ_M as compared with the other eigenvalues. The ingredient of cooperative activities carried by the market mode (the eigenvector corresponding to λ_M) has been excluded in determining the grouping parameters (see Table I). Such division can only be feasible via the decomposing of the fluctuations into collective modes. It is supported by the observation that, in market mode [see Fig. 3(b)], the stocks prices of the whole market change coherently in the same direction (sign). The grouping obtained in our approach is quite different from that via a clustering procedure based merely on the static information of cross correlation [14]. In Fig. 4, we compare the grouping obtained by our approach with the information of S&P 500 sectors, and with the minimum-spanning tree based on the analysis of correlation coefficients [14]. It is apparent that the structure of grouping for the lower level collective activities obtained in our analysis is different from that based on static information. In the latter case, the dominant market cooperative activities are not separated.

IV. FINAL COMMENTS

To conclude, we have proposed a dynamical model of coupled random walks for the evolution of stock prices, which properly accounts for the important deviations from the RMT predictions. The correlations are realized on the level of time sequences rather than on the statistics of correlation matrix, which deserves further investigation for wider applications. In applying to market data, the coupling parameters suggest the usage of the similar concept of (the inverse of) impedance which effectively describes systems of dense fluid. To our knowledge, such kind of analysis has never been carried out from the standpoint of the whole market [8]. The kinetic contents of these parameters and the implication for the partitioning of the stocks and for the structure of underlying network are the relevant issues to explore. Though we have presented results only for the economic data, we feel that our model and analysis will be useful in other problems as well where significant deviations from the RMT predictions are found [5,6].

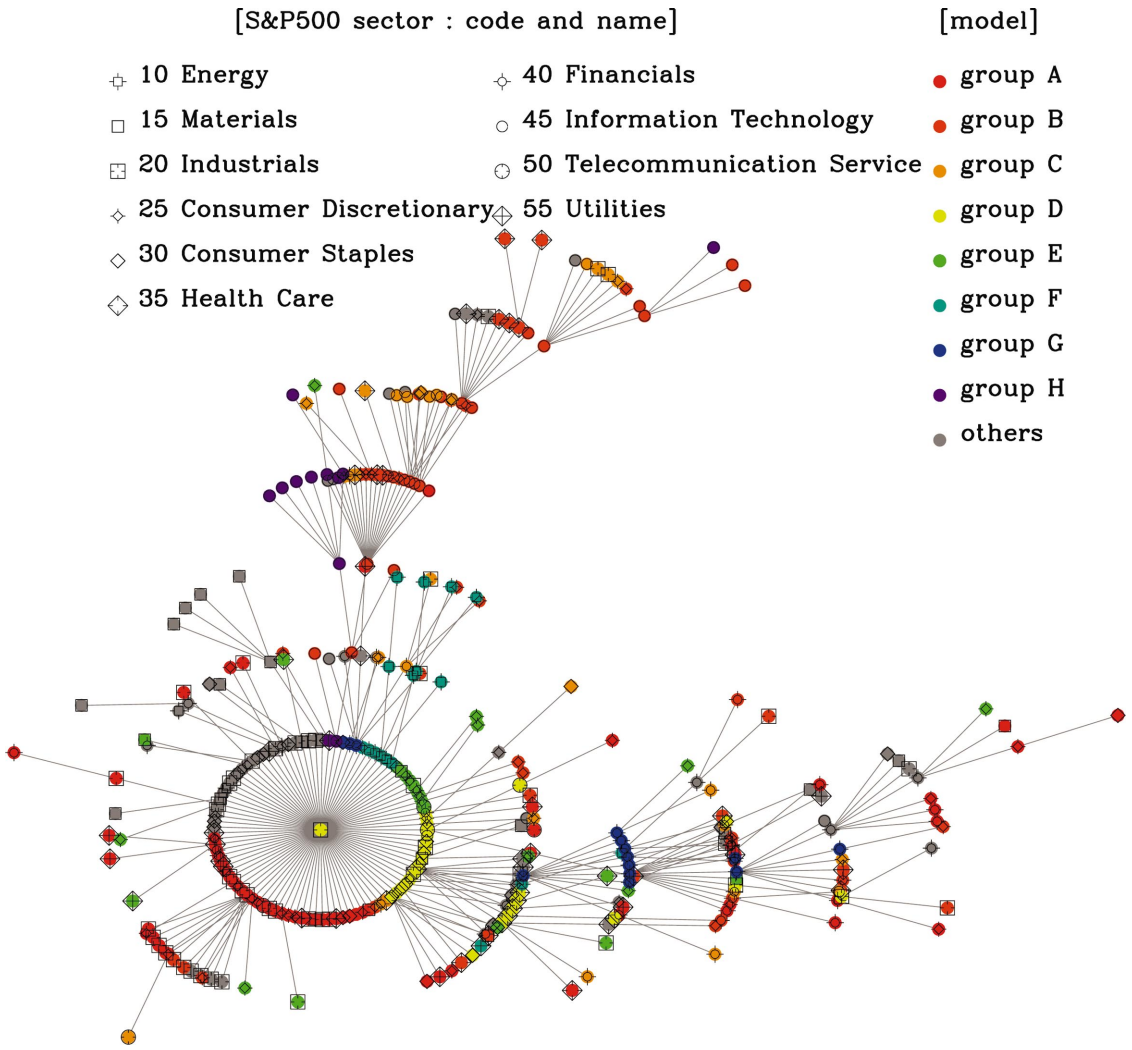

FIG. 4. (Color) The 345 US stocks classified by our method and by the minimum spanning tree analysis on the correlation coefficients in Ref. [14]. We label the stocks in the eight largest groups obtained by our method (see text and Table I) in different colors. The stocks are also marked by different symbols according to their sectors in S&P 500 classification. In our approach, the grouping is realized as the perturbing part of the cooperative activities in the fluctuations to the dominant market activities (see text). In the minimum spanning tree analysis, there is no such division.

ACKNOWLEDGMENTS

We thank J. Voit and T. Lux for comments. W.J.M. acknowledges the helpful conversation with C. N. Chen, K. T. Leung, D. B. Saakian, T. Shimada, J. Skřivánek, and T. M. Wu. This work was supported in part by National Science Council of the Republic of China (Taiwan) under Grant No. NSC 91-2112-M-001-056.

APPENDIX: CORRELATION MATRIX

According to Eq. (7), the correlation matrix is

$$\mathbf{C} = \mathbf{\Gamma}^{-1/2}(\mathbf{R} - \mathbf{R\Omega})(\mathbf{R} - \mathbf{R\Omega})^{\mathbf{T}}\Gamma^{-1/2}, \tag{A1}$$

where the $N \times N$ matrix $\mathbf{\Gamma}$ and the $T \times T$ matrix $\mathbf{\Omega}$ are defined by

$$\mathbf{\Gamma}_{ij} = \begin{cases} ((\mathbf{R} - \mathbf{R\Omega})(\mathbf{R} - \mathbf{R\Omega})^{\mathbf{T}})_{ii} & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \text{ and } \mathbf{\Omega}_{ij} = \frac{1}{T} \text{ for } i,j = 1, \ldots, T, \tag{A2}$$

respectively.

To find the distributions for the entries of the correlation matrix **C**, we evaluate the ensemble averages of $\mathbf{RR^T}$, $\mathbf{R\Omega R^T}$, and $\mathbf{R\Omega\Omega R^T}$. It is useful to evaluate $\langle R_{is}R_{jt}\rangle$ first. We have [from Eq. (3)]

$$\langle R_{is}R_{jt}\rangle = \sigma^2\left[\delta_{ij}\delta_{st}(1-\epsilon_M) + \delta_{ij}A_{st} + \delta_{st}\frac{\epsilon_M}{N} - \frac{A_{st}}{N}\right], \tag{A3}$$

where

$$A_{st} = \frac{\epsilon_M}{\epsilon_M - 2}[(1-\epsilon_M)^{s+t} + (1-\epsilon_M)^{|s-t|+1}]. \tag{A4}$$

From Eq. (A3), we obtain Eq. (13) in the stationary time regime and, accordingly,

$$\langle(\mathbf{RR^T})_{ij}\rangle = \sigma^2 T\left[\left(\delta_{ij}(1-a(\epsilon_M)) + \frac{a(\epsilon_M)}{N}\right)\right] \tag{A5}$$

and

$$\begin{aligned}\langle(\mathbf{R\Omega R^T})_{ij}\rangle &= \langle(\mathbf{R\Omega\Omega R^T})_{ij}\rangle \\ &= \sigma^2\left[\delta_{ij}(1-\epsilon_M) + \delta_{ij}\frac{1}{T}\sum_{t=t_0+1}^{t_0+T}\sum_{s=t_0+1}^{t_0+T}A_{st} + \frac{\epsilon_M}{N}\right. \\ &\quad \left. - \frac{1}{T}\sum_{t=t_0+1}^{t_0+T}\sum_{s=t_0+1}^{t_0+T}\frac{A_{st}}{N}\right] \\ &\approx \frac{\sigma^2}{N} + o\left(\frac{1}{T}\right),\end{aligned} \tag{A6}$$

where we have used the equality

$$\sum_{t=1}^{u}\sum_{s=1}^{u}A_{st} = \left[\frac{(1-\epsilon_M)^{2u+2} - (1-\epsilon_M)^2}{\epsilon_M(\epsilon_M-2)}\right] - u(1-\epsilon_M).$$

Comparing Eqs. (A5) and (A6), we conclude that, with properly chosen σ^2 for the random noises [Eq. (3)] to assure the unity variances in Eq. (7), Eq. (11) is a good approximation in the large T limit.

[1] M. L. Mehta, *Random Matrices*, 2nd ed. (Academic Press, New York, 1991).

[2] See, for example, V. V. Flambaum *et al.*, Phys. Rev. A **50**, 267 (1994); A. D. Mirlin, Phys. Rep. **326**, 259 (2000); W.-J. Ma *et al.*, Physica A **321**, 364 (2003); S. N. Taraskin and S. R. Elliott, Phys. Rev. B **65**, 052201 (2002).

[3] L. Laloux, P. Cizeau, J.-P. Bouchaud, and M. Potters, Phys. Rev. Lett. **83**, 1467 (1999).

[4] V. Plerou *et al.*, Phys. Rev. Lett. **83**, 1471 (1999); V. Plerou *et al.*, Phys. Rev. E **65**, 66126 (2002).

[5] M. Barthélemy, B. Gondran, and E. Guichard, Phys. Rev. E **66**, 056110 (2002).

[6] M. S. Santhanam and P. K. Patra, Phys. Rev. E **64**, 016102 (2001).

[7] J. Voit, *The Statistical Mechanics of Financial Markets* (Springer, New York, 2001); R. N. Mantegna and H. E. Stanley, *An Introduction to Econophysics: Correlations and Complexity in Finance* (Cambridge University Press, Cambridge, 2000); J.-P. Bouchaud and M. Potters, *Theory of Financial Risk: From Statistical Physics to Risk Management* (Cambridge University Press, Cambridge, 2000).

[8] E. J. Elton and M. J. Gruber, *Modern Portfolio Theory and Investment Analysis*, 3rd ed. (Wiley, New York, 1995); J. Y. Campbell, A. W. Lo, and A. C. MacKinlay, *The Econometrics of Financial Markets* (Princeton University, Princeton, NJ, 1997).

[9] L. Bachelier, Ann. Sci. Ec. Normale Super. **17**, 21 (1900); English translation of *The Random Character of Stock Market Prices*, edited by P. H. Cootner (MIT, Cambridge, MA, 1964).

[10] T. Lux and M. Marchest, Nature (London) **397**, 498 (1999).

[11] J. D. Noh, Phys. Rev. E **61**, 5981 (2000).

[12] S. Drozdz *et al.*, Physica A **287**, 440 (2000).

[13] A. M. Sengupta, and P. P. Mitra, Phys. Rev. E **60**, 3389 (1999).

[14] N. Vandewalle, F. Brisbois, and X. Tordoir, Quant. Finance **1**, 372 (2001); G. Bonanno, F. Lillo, and R. N. Mantegna, *ibid.* **1**, 96 (2001); R. N. Mantegna, Eur. Phys. J. B **11**, 193 (1999).

[15] L. Giada and M. Marsili, Phys. Rev. E **63**, 061101 (2001).

[16] K. Kaneko, Physica D **124**, 322 (1998); **34**, 1 (1989); S. Jalan and R. E. Amritkar, Phys. Rev. Lett. **90**, 014101 (2003); P. M. Gade and C.-K. Hu, Phys. Rev. E **60**, 4966 (1999); **62**, 6409 (2000).

[17] The fluctuation δr_i can be decomposed as the summation of products of the site-dependent and time-dependent parts, $\delta r_i(t) = \Sigma_s \lambda_s^{1/2} u_s(i) v_s(t)$, where $u_s(i)$ is the ith component of the eigenvector corresponding to the eigenvalue λ_s.

[18] The concise algebraic properties of $\mathbf{\Delta}_M$ and $\mathbf{\Delta}_G$, $\mathbf{\Delta}_M\mathbf{\Delta}_M = (-\epsilon_M)\mathbf{\Delta}_M$ and $\mathbf{\Delta}_M\mathbf{\Delta}_G = \mathbf{\Delta}_G\mathbf{\Delta}_M = (-\epsilon_M)\mathbf{\Delta}_G$, allow for the analytic evaluation of various quantities in the approximation to the first order of the ϵ_g's for $\epsilon_g \ll \epsilon_M$.

[19] The formulation Eq. (10) is valid for the general case $\mathbf{\Delta}_G \neq 0$ (with $\mathbf{\Delta}_M$ and Π replaced, respectively, by $\mathbf{\Delta} = \mathbf{\Delta}_M + \mathbf{\Delta}_G$ and a new Π_G which possesses grouping dependence). Using the same notations, the returns in the model of Ref. [11] are driven by Ξ adding a set of noises Ξ_G which has its N rows divided into groups, $\mathbf{R} = \Xi_G + \Xi$. For each column of Ξ_G, the entries of the rows of the same group are the same. Ξ_G provides the statistical overlap in Eq. (11), leading to the blocked structure in the correlation matrix which accounts for the presence of those deviated eigenvalues from the bulk of the distribution.

[20] We are looking for the maximum of $\mathbf{f^T C^* f}$ over any column vector $\mathbf{f}$ with unity norm

$$\lambda_M^* = \max_{\mathbf{f^T f}=1}\{\mathbf{f^T C^* f}\}$$

since the maximum occurs for the column vector $\mathbf{f}$ equal to the eigenvector corresponding to the largest eigenvalue. For matrix $\mathbf{C}^*$, $\mathbf{f}$ happens to have all its components in the same value, which reflects the general feature of the cooperative activities of the market mode. The above equation is also valid between

any **C** and its largest eigenvalue λ_M. If we consider the time sequences of the column vectors of the fluctuations of returns $\delta\mathbf{r}(t)$ which generate the correlation matrix **C**, we can write

$$\lambda_M = \max_{\mathbf{f}^T\mathbf{f}=1}\left\{\frac{1}{T}\sum_{t=1}^{T}[\mathbf{f}^T\delta\mathbf{r}(t)]^2\right\}.$$

λ_M describes then the maximum possible mean square fluctuation over time, and the components of the vector **f** giving that maximum describe the corresponding portfolio.

[21] Using $\tilde{\epsilon}\equiv 1-\epsilon_M$, Eq. (15) can be written as

$$\lambda_M^* = \frac{1+\tilde{\epsilon}}{\tilde{\epsilon}^2}\frac{\tilde{\epsilon}^2 N}{2\tilde{\epsilon}^2 N + (1+\tilde{\epsilon}-\tilde{\epsilon}^2)}.$$

We are interested in the region that $\tilde{\epsilon}$ is small and N is large. It is apparent that the equation can be reduced as an equality for two new quantities $\tilde{\lambda}_M^* \equiv \tilde{\epsilon}^2\lambda_M^*$ and $x\equiv\tilde{\epsilon}N^{1/2}$,

$$\tilde{\lambda}_M^* = \frac{x^2}{2x^2+1}\left(1+o\left(\tilde{\epsilon},\frac{1}{\tilde{\epsilon}N}\right)\right),$$

for $1/N \ll \tilde{\epsilon} \ll 1$.

[22] For market data, we analyze the correlation matrix of the log returns, that is, the changes of log price $\ln(P_i(t+1)) - \ln(P_i(t))$ for all stock i's. The log return can be approximated by the price return $\delta P_i/P_i$ normalized by the price, which justifies the use of the stochastically identical walks in our model.

[23] According to the random matrix theory [13], the spectrum of the eigenvalues for totally random case ($\epsilon_M=0$ and $\epsilon_g=0$) is given by $\rho(\lambda)=Q/2\pi\lambda\sqrt{(\lambda_{\max}-\lambda)(\lambda-\lambda_{\min})}$ with $Q=T/N$ and $\lambda_{\max,\min}=1+1/Q\pm 2\sqrt{1/Q}$. For Fig. 1, we have $Q=9.420$ and $\lambda_{\max,\min}=1.7578\ldots,0.4545\ldots$.

[24] We survey the time sequences of trading prices from the Trades and Quotes (TAQ) database for one pool of $N=345$ US stocks selected from S&P 500, which have complete records over the four years 1996–1999.

[25] We also use our model to fit the spectral data of Ref. [4] with $N=1000$ and $T=6448$ (not shown) and find good agreement. For this fit the groups were not chosen from eigenvector analysis.

[26] D. Korenblum and D. Shalloway, Phys. Rev. E **67**, 056704 (2003), and references therein.

[27] M. Marsili, Quant. Finance **2**, 297 (2002).

[28] We can express the deviation $\langle r_i^2\rangle$ of the return for walk i analytically as a linear function of the deviations $\langle\xi_k^2\rangle$ of all the noises ξ_k's $k=1,\ldots,N$, with the coefficients containing the parameters, ϵ_M, n_g's, and ϵ_g's. The couplings ϵ_M and ϵ_g's can be estimated by filling $\langle r^2\rangle$'s and $\langle\xi^2\rangle$'s with the market data. (In practice, it is necessary to carry out certain coarse graining for this mean-field model by neglecting certain stock-by-stock differences in the real market data.) The estimation is then further finely adjusted to match the spectra. In the procedure, there is always some compromise between reproducing the grouped activities in the eigenvectors and matching the distribution spectra.

[29] Normally, the inverse participation ratio [4,5] is used to estimate the fraction of components which significantly contribute to that eigenvector. However, it cannot tell us about the elements (walks) which are common to different groups, an information crucial to determine the different groups in our case. Hence, we use a different procedure to determine the groups as described in the text.