Emergence of time-horizon invariant correlation structure in financial returns by subtraction of the market mode

Christian Borghesi
Service de Physique de l'Etat Condensé (CNRS URA 2464), CEA Saclay, 91191 Gif sur Yvette Cedex, France

Matteo Marsili*
The Abdus Salam ICTP, Strada Costiera 11, I-34014 Trieste, Italy

Salvatore Miccichè†
Dipartimento di Fisica e Tecnologie Relative, Università degli Studi di Palermo, Viale delle Scienze, Edificio 18, I-90128 Palermo, Italy

(Received 15 February 2007; published 10 August 2007)

We investigate the emergence of a structure in the correlation matrix of assets' returns as the time horizon over which returns are computed increases from the minutes to the daily scale. We analyze data from different stock markets (New York, Paris, London, Milano) and with different methods. In addition to the usual correlations, we also analyze those obtained by subtracting the dynamics of the "center of mass" (i.e., the market mode). We find that when the center of mass is not removed the structure emerges, as the time horizon increases, from splitting a single large cluster into smaller ones. By contrast, when the market mode is removed the structure of correlations observed at the daily scale is already well defined at very high frequency (5 min in the New York Stock Exchange). Moreover, this structure accounts for 80% of the classification of stocks in economic sectors. Similar results, though less sharp, are found for the other markets. We also find that the structure of correlations in the overnight returns is markedly different from that of intraday activity.

DOI: 10.1103/PhysRevE.76.026104 PACS number(s): 05.90.+m, 89.75.Hc

I. INTRODUCTION

Besides their intrinsic interest, financial markets have also attracted a great deal of attention as a paradigm of complex systems of interacting agents. In this view, the correlation between different assets are one of the signatures of the complexity of the system's interactions and, as such, have been the focus of intense recent research [1,2]. The central object of study is the empirical covariance matrix of a set of N assets, whose elements are the Pearson's correlation coefficients $C_{i,j}^{(\tau)}(T)$ between the log returns of assets i and j over a time horizon τ, measured on historical time series of length T. Early studies have focused mainly on daily returns ($\tau = 1$ day) and have shown that the bulk of the eigenvalue distribution of the correlation matrix is dominated by noise and well described by random matrix theory [3,4]. This band of noisy eigenvalues shrinks as $\sqrt{N/T}$ as the length T of the dataset increases, and it describes a large part of the spectrum in typical cases where N and T are of the order of some hundreds. The few large eigenvalues which leak out of the noise background contain significant information about market's structure. The taxonomy built with different methods [4–7] from financial correlations alone bears remarkable similarity with a classification in economic sectors. This agrees with the expectation that companies engaged in similar economic activities are affected by economic factors in a similar way. With respect to their dynamical properties, it has been found that financial correlations are persistent over time [8] and that they follow recurrent patterns [7].

Furthermore, correlations "build up" as the time horizon τ on which returns are measured increases, and they saturate for returns on the scale of some days [9–11]. This behavior, known as the Epps effect [12], has been analyzed in much detail recently [13]. From the point of view of market efficiency, which entails that prices incorporate information on single asset returns, the Epps effect can be seen as a mechanism by which the mutual information on assets affects the correlation of their returns. Interestingly, it was found that such a process is much faster today than in the past and more pronounced for more capitalized stocks [10]. It has also been remarked [9,14] that the structure of correlations changes as the time horizon τ over which returns are defined increases, i.e., that "pictorially, the market appears as an embryo which progressively forms and differentiates over time" [11].

Here we shall take a closer look on the dependence of the structure of correlations on the time horizon τ and show that the observed evolution of the market structure is due to the dynamics of the market mode. Global correlations play a dominant role at high frequency, thus giving rise to correlation structures which are much more clustered than at the daily scale. However, if global correlations are removed, the structure of correlations at the daily scale is largely preserved across time horizons, down to a scale of 5 min for the most liquid market we have analyzed. Loosely speaking, the network structure, after removing the market mode, appears fully formed and differentiated already at small scales; it only grows in size (of correlations) as the time horizon increases.

The elimination of the market mode from pairwise correlations between stocks is analogous to decomposing the dynamics of a complex interacting system in that of its center of mass and of its internal coordinates. In physics such a

*marsili@ictp.it, also DEMOCRITOS
†micciche@lagash.dft.unipa.it

separation is a consequence of translation invariance in space, which generally implies that the center of mass dynamics is determined by external forces, whereas internal coordinates respond to interparticle interaction forces. A similar translation invariance in log prices exists in principle in financial markets: multiplying all prices $p_i \to p_i' = \lambda p_i$ by the same constant $\lambda > 0$ (including that of the numeraire) amounts to a mere change of units, which should leave the state of the system invariant. This suggests that even in financial markets it makes sense to separate the dynamics of the "center of mass" from that of "internal coordinates," and that the two may follow distinct laws of motion. Our results strongly support this view.

The paper is organized as follows: in the next section we discuss the datasets and how we build correlation matrices. Then we shall discuss the results of data clustering approach first for the New York Stock Exchange and then for the other markets. The following section deals with the minimal spanning trees approach. Finally we shall summarize our results and offer some concluding remarks.

II. DATA

In this paper we empirically investigate the ensemble behavior of price returns for four different markets: the New York Stock Exchange (NYSE), the London Stock Exchange (LSE), the Paris Bourse (PB) and the Borsa Italiana (BI). All data refer to the year 2002. The number of days in the datasets is 251 days for the NYSE and BI data, 250 days for the LSE data, and 254 days for the PB data.

The NYSE data are taken from the Trades and Quotes (TAQ) database maintained by NYSE [15]. In particular, 100 highly capitalized stocks were considered. For each stock and for each trading day we consider the time series of stock prices recorded transaction by transaction. Since transactions for different stocks do not happen simultaneously, we divide each trading day, lasting $6^h30'$, into intervals of length τ. For each trading day, we define N_τ intraday stock price proxies $p_i(t_k)$ of asset i, with $k=1,\cdots,N_\tau$. The proxy is defined as the transaction price detected nearest to the end of the interval (this is one possible way to deal with high-frequency financial data [16]). By using these proxies, we compute the price returns

$$a_i^{(\tau)}(t) = \ln p_i(t) - \ln p_i(t-\tau) \tag{1}$$

at time horizons τ. The time horizons used are $\tau = 5, 15, 30, 65,$ 195 min. For NYSE, values of τ are large enough that all the considered stocks have at least one transaction in each time interval.

The LSE data are taken from the "Rebuild Order Book" database, maintained by LSE [17]. In particular, we consider only the electronic transactions for 92 highly traded stocks belonging to the SET1 segment of the LSE market. The trading activity has been defined in terms of the total number of transactions (electronic and manual) occurring in 2002. However, most of the transactions, a mean value of 75% for the 92 stocks, are of the electronic type. This market is commonly believed to be very active and can be regarded as a realization of a "liquid" market. For each stock i and for each trading day we consider the time series of stock price recorded transaction by transaction and generate N_τ intraday stock price proxies $p_i(t_k)$ according to the procedure explained above. For the LSE data each trading day lasts $8^h30'$ and the time horizons used were 5, 15, 30, 51, 102, 255 min.

The PB data are taken from the "Historical Market Data" database, maintained by EURONEXT [18]. In particular, we consider the electronic transactions of two subsets of stocks traded in the year 2002. For each stock i and for each trading day, lasting $8^h30'$, we consider the time series of stock price recorded transaction by transaction and generate N_τ intraday stock price proxies $p_i(t_k)$ according to the procedure explained above. One first set, which will be analyzed in the section, consists of the 75 most frequently traded stocks at time horizons $\tau_k = 27\times 2^k$ sec, for $k=0,\ldots,10$. An analogous dataset was derived considering tick time: $\tau_k^{(tick)} = 100\times 2^k$. This choice was considered in order to probe the region of very high frequencies and to assess the relevance of time inhomogeneity of trading activity at intraday time scales. In this respect, it is worthwhile to remark that for small τ stocks were not traded in each time interval. A second dataset, that will be considered in Section IV, instead consisted of $N = 39$ stocks which were continuosly traded in the entire year 2002 (i.e., in each time interval) over a time horizon of $\tau = 5, 15, 30, 51, 102,$ 255 min.

The BI data are taken from the "Dati Intraday" database, maintained by Borsa Italiana [19]. In particular, we consider only the electronic transactions occurring for 30 stocks continuosly traded in the entire year 2002. For each stock i and for each trading day we consider the time series of stock price recorded transaction by transaction and generate N_τ intraday stock price proxies $p_i(t_k)$ according to the procedure explained above. For the BI data, the time horizons used were 5, 15, 30, 60, 120, 240 min. Each trading day lasts 8^h.

For all markets, in addition to the intraday time horizons, we have considered returns on the daily time horizon,

$$\begin{aligned} a_i^{(\text{cl-op})}(n) &= \ln p_i^{\text{cl}}(n) - \ln p_i^{\text{op}}(n), \\ a_i^{(\text{cl-cl})}(n) &= \ln p_i^{\text{cl}}(n) - \ln p_i^{\text{cl}}(n-1), \\ a_i^{(\text{op-cl})}(n) &= \ln p_i^{\text{op}}(n) - \ln p_i^{\text{cl}}(n-1), \end{aligned} \tag{2}$$

corresponding to intraday, daily, and overnight returns, respectively. Here $p_i^{\text{op}}(n)$ and $p_i^{\text{cl}}(n)$ are the open and closure prices of stock i in day n.

Each stock can be associated to an economic sector of activity (see Ref. [20]). The relevant economic sectors are reported in Table I.

Given the price return at a selected time horizon τ, we built the correlation matrices in the usual way,

$$A_{i,j}^{(\tau)} = \frac{\langle a_i^{(\tau)} a_j^{(\tau)}\rangle - \langle a_i^{(\tau)}\rangle\langle a_j^{(\tau)}\rangle}{\sqrt{\langle [a_i^{(\tau)} - \langle a_i^{(\tau)}\rangle]^2\rangle\langle [a_j^{(\tau)} - \langle a_j^{(\tau)}\rangle]^2\rangle}}. \tag{3}$$

Here and in what follows, $\langle \ldots \rangle = (1/N_\tau)\Sigma_{t=1}^{N_\tau} \ldots$ denotes time average.

In order to disentangle different components of the dynamics and to understand their effect, we considered also series of datasets derived from $a_i^{(\tau)}(t)$. In all derived datasets

TABLE I. Color codes for the economic sectors of activity for the stocks.

	Sector	Color
1	Technology	red
2	Financial	green
3	Energy	blue
4	Consumer noncyclical	yellow
5	Consumer cyclical	brown
6	Healthcare	grey
7	Basic materials	violet
8	Services	cyan
9	Utilities	magenta
10	Capital goods	light green
11	Transportation	maroon
12	Conglomerates	orange

we subtract a particular component of market dynamics from the rest. When the structure of the derived dataset differs substantially from that of the matrix $\hat{A}$ we can conclude that the decomposition is meaningful and informative.

First we removed the "center of mass" dynamics:

$$b_i^{(\tau)}(t) = a_i^{(\tau)}(t) - \frac{1}{N}\sum_{j=1}^{N} a_j^{(\tau)}(t). \tag{4}$$

From this, a covariance matrix $B_{i,j}^{(\tau)}$ was computed in the same way as in Eq. (3).

In a further dataset we removed the effect of the market index from $a_i^{(\tau)}(t)$. This was done by first considering the time series $I^{(\tau)}(t)$ of the corresponding market index at the same time horizon τ and then by estimating the coefficients of a one factor model,

$$a_i^{(\tau)}(t) = \alpha_i + \beta_i I^{(\tau)}(t) + c_i^{(\tau)}(t). \tag{5}$$

The residues $c_i^{(\tau)}(t)$ were used to build the covariance matrix $C_{i,j}^{(\tau)}$. We could build the time series $I^{(\tau)}(t)$ only in the case of NYSE data, for which we had access to intraday data of the SP500 composite index.

In all datasets we computed an "endogenous" market index using the market average return

$$\bar{a}^{(\tau)}(t) = \frac{1}{N}\sum_{j=1}^{N} a_j^{(\tau)}(t).$$

By using $\bar{a}^{(\tau)}(t)$ instead of the market index $I^{(\tau)}(t)$ in Eq. (5) and by considering the appropriate residues $d_i^{(\tau)}(t)$, we computed a further covariance matrix $D_{i,j}^{(\tau)}$.

Finally, we produced a dataset by removing the contribution of the largest eigenvector of the matrix $A_{i,j}^{(\tau)}$. This can be done by zeroing the largest eigenvalue of $\hat{A}$, as discussed in Ref. [4]. An alternative method, which we prefer, is that of removing the "optimal" factor, $G^{(\tau)}(t)$ which is obtained by minimizing

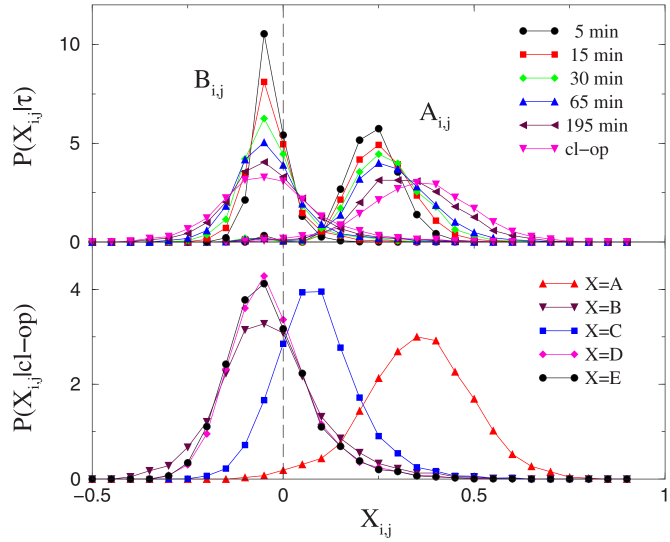

FIG. 1. (Color online) Distribution of correlation coefficients $A_{i,j}$ and $B_{i,j}$ for different time horizons τ (top) and at the intraday time horizon for different datasets (NYSE data).

$$\chi^2 = \sum_{i=1}^{N}\sum_{t=1}^{N_\tau}[a_i^{(\tau)}(t) - \alpha_i - \beta_i G^{(\tau)}(t)]^2 \tag{6}$$

on α_i, β_i, and $G^{(\tau)}(t)$. The residues $e_i^{(\tau)}(t)$ resulting from this operation coincide with the time-series obtained from $a_i^{(\tau)}(t)$ by subtracting the leading contribution of its singular value decomposition. We call $E_{i,j}^{(\tau)}$ the correlation matrix of the residues $e_i^{(\tau)}(t)$.

In summary, we consider the original time series (set A), the one obtained subtracting the average market return (set B), and those obtained from the residues of a one factor model with the market index (set C), the average market return (set D), and the optimal factor (set E). Set C represents a case where the market mode is exogenously determined whereas in sets D and E it is determined by the data itself. This allows us to understand how much an index, such as SP500 which is a weighted average, accounts for the collective dynamics of the market.

The distribution of matrix elements is shown in Fig. 1 as a function of time horizon (top, for the sets A and B) and for different datasets at the intraday time horizon. We observe that the distribution spreads out as the time horizon increases, as a manifestation of the Epps effect. However, while the distribution of $A_{i,j}$ is centered around a positive value, that of correlations of derived datasets is peaked on values close to zero and is narrower. For set B (D and E) the peak is at slightly negative values, whereas for set C it occurs at positive values. This suggests that the removal of correlations is more efficient when the single factor is computed from the data. In particular, the dynamics of the mean $\bar{a}(t)$ explains the correlations better than the market index.

We also find that intraday and overnight returns have distinctly different distribution of correlation coefficients. This difference is particularly pronounced in dataset C which again suggests that the market index is even less explicative

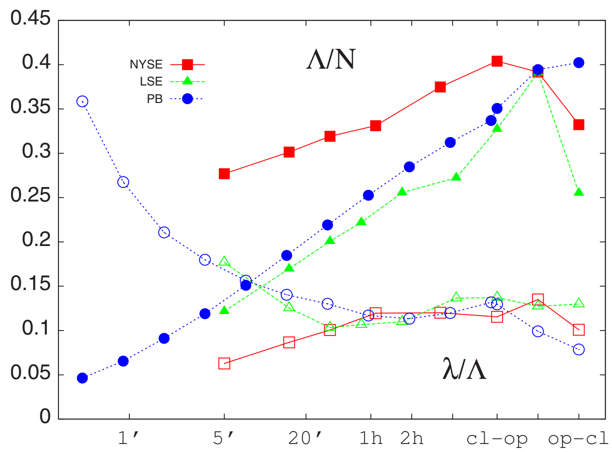

FIG. 2. (Color online) Largest eigenvalue Λ/N, divided by the number of assets, of the matrix $\hat{A}_\tau$ as a function of τ for NYSE, LSE and PB (full symbols). Ratio λ/Λ of the second largest to the largest eigenvalue of $\hat{A}_\tau$, as a function of τ (open symbols). The labels cl-op and op-cl refer to the intraday and overnight returns, respectively, defined in Eq. (2).

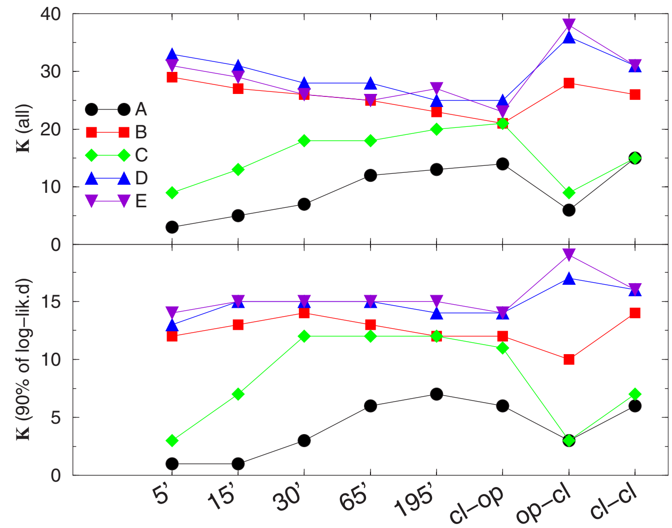

FIG. 3. (Color online) Top: Number of clusters for datasets A, B, C, D, and E. Bottom: Number of clusters accounting for 90% of the likelihood (NYSE data). The labels cl-op, cl-cl, and op-cl refer to the time horizons defined in Eq. (2).

of the market's collective behavior at these scales.

Correlation $D_{i,j}$ and $E_{i,j}$ were found to have a distribution which is similar to that of $B_{i,j}$. This anticipates a generic conclusion: the subtraction of a global component from the dynamics is most meaningful when it eliminates (either implicitly as in $\hat{B}$ or explicitly as in $\hat{E}$) the market mode by setting the corresponding eigenvalue to zero.

Before analyzing the structure of correlations, it is of interest to provide some estimate of the relative strength of global correlations and of noise in the correlation matrices $\hat{A}$. Figure 2 plots the share of correlation carried by the largest eigenvalue Λ (which is Λ/N, by normalization) for set $\hat{A}$ of NYSE, LSE, and PB, as a function of time horizon τ. As a manifestation of the Epps effect [12], Λ/N increases with τ in a way which is reasonably well approximated by a logarithmic growth. The ratio of the second largest eigenvalue λ to the largest, which could be taken as a measure of the relative strength of interasset correlations against global correlations, has a declining trend with τ for small time horizons and then saturates at around 0.1.

III. DATA CLUSTERING

We performed data clustering analysis following the method of Ref. [6]. Here we only sketch the basic idea of the method and we refer the interested reader to Ref. [6] for details. In brief, assume we wish to cluster N standardized [26] time series $x_i(t)$ in K groups having a similar dynamics. This task can be formalized in the problem of finding the labels $s_i=1,\ldots,K$ of the group to which the ith time series belongs. Reference [6] assumes that $x_i(t)$ is generated according to the model

$$x_i(t) = g_{s_i}\eta_{s_i}(t) + \sqrt{1-g_{s_i}^2}\,\epsilon_i(t), \quad t=1,\ldots,T, \tag{7}$$

where $\eta_s(t)$ and $\epsilon_i(t)$ are independent gaussian variables with zero mean and unitary variance. Here $\eta_s(t)$ describes the component of the dynamics which is common to all time series $x_i(t)$ with $s_i=s$ whereas $\epsilon_i(t)$ describes idiosyncratic fluctuations. Equation (7) is consistent with a correlation matrix $X_{i,j}=\langle x_i x_j\rangle$ which has a block diagonal structure for $T\to\infty$: $X_{i,j}=g_{s_i}^2$ if $s_i=s$, $i\neq j$ $(X_{i,i}=1)$, and $X_{i,j}=0$ otherwise. The parameters g_s entering Eq. (7) as well as the cluster structure $\{s_i\}$ can be determined by maximum likelihood estimation. Approximate maximization of the log likelihood can be done following a hierarchical clustering procedure [27]: start with N clusters, each composed of a single asset $(s_i^{(N)}=i)$. From the configuration $\{s_i^{(K+1)}\}$ with $K+1$ clusters, compute the log likelihood of all configurations obtained by merging two clusters. The configuration $\{s_i^{(K)}\}$ with K clusters is the one corresponding to the maximal log likelihood $\mathcal{L}_K$. This operation can be iterated with K going from $N-1$ to 1, and the optimal configuration can be chosen as that for which $\mathcal{L}_K$ is maximal. This also predicts the optimal number K^* of clusters which describes our dataset. This method has already been used to analyze stock market data: in Refs. [6] the emergent clusters were found to be highly correlated with economic activity. Furthermore, the method was extended to perform noise undressing. In Ref. [7] the method has been applied to investigate market dynamics, showing that well defined recurrent states of marketwide activity can be defined.

Here we apply this method to investigate how the structure of market's correlations evolves as the time lag τ increases from the high-frequency range to the daily scale. We shall first focus on NYSE and then discuss the differences found in other markets.

A. NYSE

Figure 3 shows the evolution of the number of clusters with the time horizon τ for the different datasets in the NYSE. For $\hat{A}$ we find fewer clusters than with other methods and the number of clusters increases with τ. This is consistent with results of Refs. [9] which observes an evolution of the structure of correlations, where more and more details are added as the time horizon increases. The other datasets, however, reveal that this is due to the fact that $\hat{A}$ includes the correlations induced by the common factor. When this is removed, as for $\hat{B}$, $\hat{D}$, and $\hat{E}$, we find that the number of clusters which accounts for most of the log-likelihood is remarkably stable from the 5 min to the intraday scale. When the S&P500 index is removed from the data ($\hat{C}$), we find a fast evolution of the structure between 5 and 30 min and then the number of clusters saturates to a constant level. Again, in all cases, a significant variation takes place in the overnight and hence at the daily (cl-cl) scale.

A closer view on the evolution of the cluster structure is presented in Fig. 4. This plots the cluster label $s_i^{(\tau)}$ as a function of τ, for each asset belonging to clusters accounting for 90% of the log likelihood [28]. Hence assets i and j belonging to the same cluster for all τ follow parallel trajectories in the figure. In this representation, cluster splitting and merging can be clearly read off. In dataset $\hat{A}$ and $\hat{C}$ we see considerable splitting of clusters as we move from τ=5 min to the daily time horizons. A substantial reshuffling and merging takes place when going to overnight returns. On the contrary, in datasets $\hat{B}$, $\hat{E}$, and $\hat{D}$ (not shown), cluster membership exhibits a remarkable stability at intraday scales: the vast majority of assets within a cluster at 5 min follows the same "trajectory" across time-horizons. Some reshuffling takes place in the order of clusters, suggesting that the structure of correlations among sectors evolves with time horizons. Again, the structure of overnight returns is considerably different.

In order to make the comparison of different cluster structure quantitative, we have introduced an information distance $\mathfrak{I}(s^{(1)},s^{(2)})$ between any two structures $\{s_i^{(1)}\}$ and $\{s_i^{(2)}\}$. In words, this tells us how much the knowledge of the cluster label $s_i^{(1)}$ of a randomly chosen stock i yields information on the value of $s_i^{(2)}$. Information is quantified by entropy reduction, in the following manner: let $p^{(\ell)}(s)$ be the fraction of stocks with $s_i^{(\ell)}=s$ for $\ell=1,2$ and $p^{(1|2)}(s|s')$ be the fraction of stocks with $s_i^{(1)}=s$, among those which have $s_i^{(2)}=s'$. From these, we can compute the entropies $S^{(\ell)}$ in the usual way and the conditional entropy

$$S^{(1|2)} = -\sum_{s'} p^{(2)}(s')\sum_{s} p^{(1|2)}(s|s')\ln p^{(1|2)}(s|s').$$

The information gain is then given by

$$\mathfrak{I} = \frac{S^{(1)} - S^{(1|2)}}{S^{(1)}}. \tag{8}$$

Because of the normalization, a value of $\mathfrak{I}\approx 1$ implies that $s^{(2)}$ yields a rather precise information on $s^{(1)}$, so if $\mathfrak{I}=0.8$ we shall say that $s^{(2)}$ accounts for 80% of the information contained in $s^{(1)}$. Table II shows the values of $\mathfrak{I}$ between different cluster structures and that obtained from set A at τ =1 day time horizon. This shows that at this time horizon,

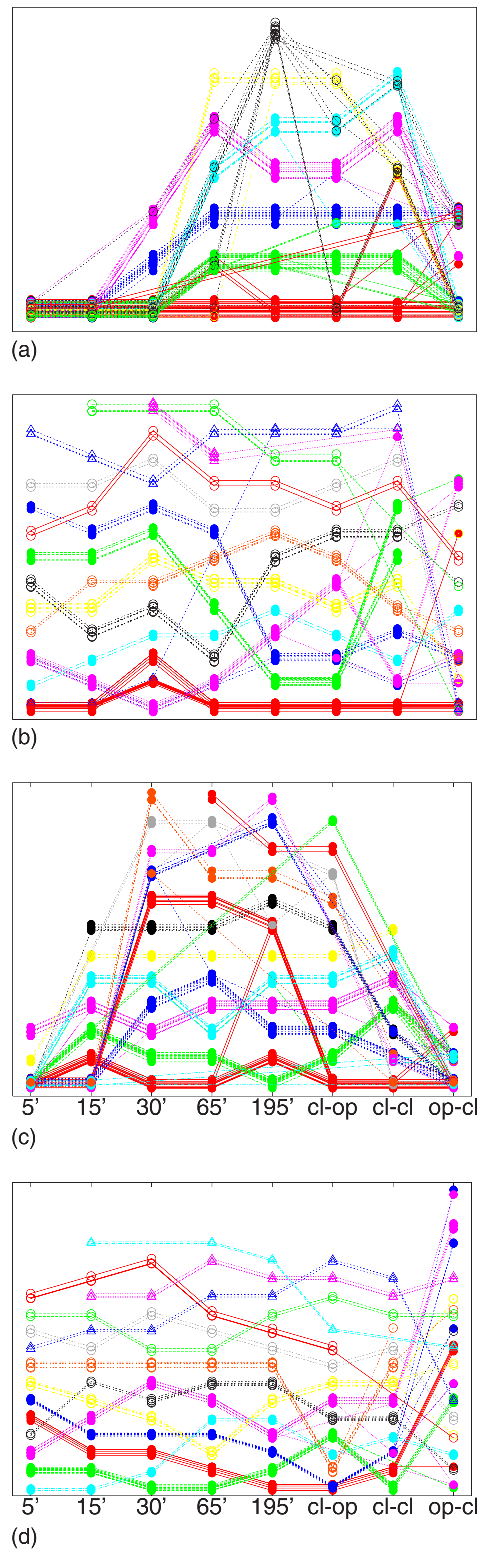

FIG. 4. (Color online) Evolution of the cluster structure with time horizon for the sets A, B, C, and E, from top to bottom of NYSE. The cluster label $s_i^{(\tau)}$ of each asset belonging to the most relevant clusters is shown as a function of τ. In this way, assets who always belong to the same cluster follow the same "trajectory" (indeed, trajectories of different assets i are shifted by a small random variable ϵ_i to distinguish them). The color is relative to the cluster structure at the intraday scale.

TABLE II. Overlaps $\Im$ between cluster structures at different time horizons and different sets and the structure of set A at $\tau=1$ day (op-cl).

$\Im(\%)$	5 min	15 min	30 min	65 min	195 min	cl-op	cl-cl	op-cl
A	0.5	0.11	0.42	0.77	0.86	1.00	0.89	0.24
B	0.91	0.90	0.91	0.90	0.92	0.90	0.90	0.72
C	0.33	0.66	0.84	0.86	0.87	0.92	0.89	0.30
D	0.91	0.90	0.92	0.91	0.89	0.92	0.90	0.78
E	0.91	0.87	0.90	0.87	0.90	0.90	0.90	0.80

the cluster structure is essentially the same in the five datasets, with an overlap larger than 90%. An overlap of the same order of magnitude attains for all intraday scales in sets B, D, and E. Even though the overlap drops down as one moves to overnight returns, the difference is much smaller in sets B, D, and E than in sets A and C. This suggests that, even though overnight returns have a structure which is markedly different from that of intraday returns, still removing the market mode allows one to reveal more invariant features.

Such invariant features, we claim, are related to economic sectors. In order to support this, we compare the cluster structures with the classification of assets in the sectors of economic activity given in Table I. The latter yields a sector label $e_i \in \{1,\ldots,12\}$ for each stock i, for which we can compute an information gain $\Im$, as above, setting $s_i^{(1)}=e_i$. Figure 5 shows the behavior of $\Im$ for different datasets across time horizons. This suggests that the most informative sets are those where the market mode is removed and these account for 80% of the information contained in e_i. For these, the information content is remarkably constant across time horizons. On the contrary, for set A the information gain $\Im$ increases with τ in the intraday range, as if information on the economic activity of assets were "released" gradually, as the time horizon increases. It is worthwhile to remark that, for all datasets, overnight returns (specially for sets A and C) carry much less information on the economic structure of the market than intraday returns.

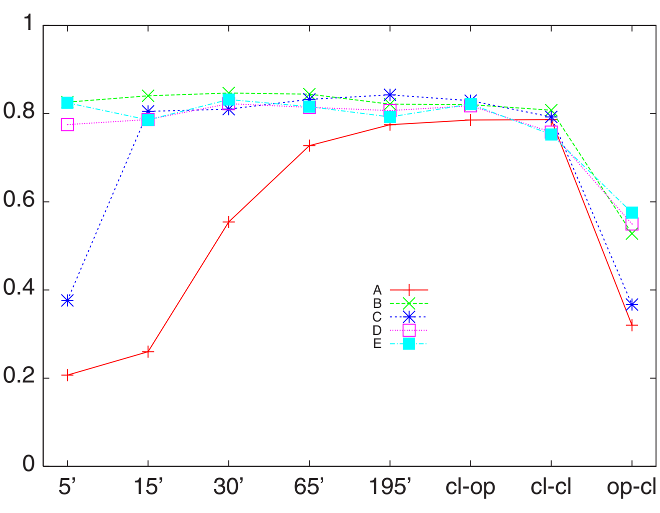

FIG. 5. (Color online) Information gain on the classification in economic sectors given by the knowledge of cluster structures $s_i^{(\tau)}$ at different time horizons τ, for different datasets $a,\ldots,e$. In order to avoid effects due to differences in the number of clusters, we considered maximum likelihood structures with 20 clusters for all datasets. Notice that by normalization $0 \leqslant \Im \leqslant 1$.

Hence we conclude that in datasets A and C the evolution in the cluster structure is due to the interplay between the "center of mass" motion (i.e., the market mode) and the internal dynamics. Indeed when the latter contribution is subtracted from the data, as in datasets B, D, and E, we find that the structure of correlations is remarkably stable with the time horizon. This is consistent with a notion of market's informational efficiency by which information is incorporated very quickly in market's returns. From the above analysis, we infer that the information on the relations between assets is efficiently incorporated in returns over time horizons shorter than 5 min in NYSE.

B. Other markets

We have performed data clustering analysis also on LSE and PB data. Again we found that removing the market mode allows one to reveal the structure of correlations much more clearly. Indeed, while set A is characterized by one or two clusters at intraday time scales, set $B,\ldots,E$ are characterized by a richer structure, as shown in Figs. 6 and 7. In both cases, we see that a significant part of the structure forms at intermediate time horizons of 15–30 min. In PB data, we pushed our analysis to ultrahigh frequency, probing very

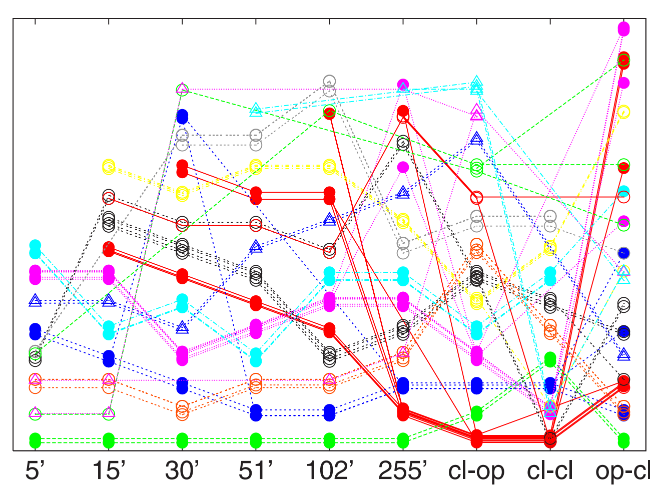

FIG. 6. (Color online) Evolution of the cluster structure for set E of LSE.

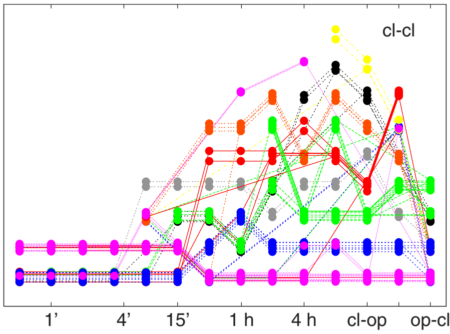

FIG. 7. (Color online) Evolution of the cluster structure for set B of 75 stocks in PB.

short time scales. We found that for $\tau < 5$ min barely any structure can be seen in the correlation matrix.

As for NYSE, we found that the cluster structure of set $\hat{A}$ is poorly correlated with the classification of assets in economic sectors, whereas datasets B and E cluster in a way which reflects up to 70% of the (entropy of a) classification in economic sectors for LSE, and that this information content is roughly constant across (intraday) time scales. As for the NYSE, we found that overnight returns have a cluster structure which is markedly different from that of intraday returns. Different markets, however, exhibit different patterns in this respect. While the LSE has a fragmented cluster structure of overnight returns similar to NYSE, PB shows a more compact structure.

In contrast with our findings on NYSE data, the cluster structure of set A is now markedly different from that of other sets even at the daily scale. This suggests that the role of global correlation is much stronger in LSE and PB.

In order to compare different markets, we performed the Kolmogorov-Smirnov (KS) test [21] on the distribution of cluster sizes. This provides a p value for the hypothesis that two different samples $\{s_i^a\}$ and $\{s_i^b\}$ of cluster sizes can be considered as different populations drawn from the same unknown parent distribution. If this is not the case (i.e., p is small), we can conclude that the two samples have a different structure, whereas if p is close to 1, we cannot reject the hypothesis that the two samples have the same structure. We found that LSE and PB have a cluster size distribution which is different from that of NYSE ($p \simeq 0.1$), but which are remarkably similar one to the other ($p \simeq 1$).

The similarity between LSE and PB, and their difference with NYSE, is also visible in the dependence of the largest eigenvalues on τ shown in Fig. 2. Remarkably, the market mode seems stronger in NYSE than in LSE and PB, whereas data clustering suggests the opposite.

In the case of PB data, we also performed several tests in order to assess the sensitivity of our results on the inhomogeneity of trading activity. One may indeed think that particular times of the day, such as the opening or the closure of the market, peak, or lunch break hours, might be characterized by different statistical properties. In order to test for these effect, we removed the first and the last 20 min of trading from the data in each day and considered the resulting correlation matrices $\hat{A}', \hat{B}', \ldots$. We computed the relative

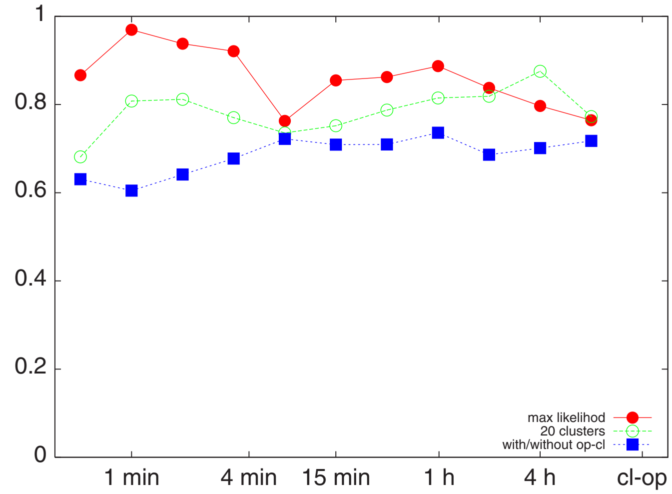

FIG. 8. (Color online) Relative entropy $\mathfrak{I}$ of the cluster structures of set B of PB obtained for (i) tick and real time (circles) on maximum likelihood structures (filled) or structures with 20 clusters (open) and (ii) with and without the opening and closure period of roughly 30′ (filled squares).

information $\mathfrak{I}$ between the maximum likelihood structures obtained in this way and the original ones, at different time scales τ. The result is that, for set B of PB, at all τ roughly $\mathfrak{I} \simeq 70\%$ of the structure found in the whole dataset coincides with that obtained eliminating the opening and the closing period (see Fig. 8). An even stronger similarity ($\mathfrak{I} = 0.83$) was found in NYSE between the structure of intraday correlations and those obtained from returns measured roughly 30 min after opening and before closing. We conclude that a significant part of the structure is not affected by the activity at the market opening or at closure.

As a further test to check the effects of time inhomogeneity of trading activity, we compute correlation matrices in tick time for PB, over intervals of $\tau_k^{(\mathrm{tick})} = 100 \times 2^k$ ticks, which correspond on average to the time scales τ_k used in real time (here a tick is defined as a transaction on any of the stocks considered). The results, shown in Fig. 8, suggest that the structure of market correlation is largely independent of the definition of time, as indeed roughly 80% of the information found with real time is recovered using tick time.

IV. SINGLE LINKAGE CLUSTERING ANALYSIS

In this section we review the results obtained by applying the single linkage clustering algorithm (SLCA) to the data considered in Section II. For each time-horizon considered, the SLCA allows us to obtain a hierarchical tree (HT) and a minimum spanning tree (MST), which give complementary information about the network structure of the considered set of stocks. Indeed, the HT gives a description of the hierarchical organization of the stocks, while the MST gives an indication about their topological organization. For a review of SLCA in the context of multivariate financial time series we refer to [14,22,23].

As much as in the previous section, here we apply the SLCA to the different datasets in order to investigate how the structure of market's correlations evolves as the time horizon τ increases from intraday scales to the daily scale. We shall first focus on NYSE and then discuss the differences found in other markets. The colors used in the representation of both the HTs and the MSTs refer to the classification sectors of economic activity given in Table I.

A. NYSE

The investigation of NYSE data by using the SLCA reveals that the role of the market mode in the structure of the correlation is twofold. On one side, the level of clustering in all the HTs in the sets where the market mode is removed is at a higher distance than the corresponding HTs of set A. Such effect is expected since, by removing the market mode, the mean correlation becomes approximately zero, as shown in Fig. 1. On the other side, the cluster structure seems now to be more evident than in the case of the original data.

In Fig. 9 we present the data for set A (top) and set E (bottom) at the two extreme time horizons of 5 min [panels (a) and (c)] and 1 day [panels (b) and (d)]. Contrary to what we find in set A [panel (a)], the HT of set E at the 5-min time horizon [panel (c)] shows a significant level of structure that, additionally, is similar to the one found at the 1-day (op-cl) time horizon [panel (d)].

This is also confirmed by comparing the structure of the MST in sets A and set E. The 5-min MST of set A shows a typical structure with a few hubs characterized by a high degree (Fig. 10). The 1-day (op-cl) MST of set A indicates that the number of hubs has increased, reflecting the progressive organization of stocks according to their sectors of activity as the time horizon increases (Fig. 11). The MSTs shown in Figs. 12 and 13 for set E are markedly different from the corresponding ones for set A. No pre-eminent hub is traceable in the two MSTs. In addition, they have a structure which is remarkably similar to one another, to the extent that one could not say which is which, on the basis of their statistical structure alone.

In order to quantify the difference between the structure of the MSTs of different datasets at different time horizons, we performed the Kolmogorov-Smirnov (KS) test [21] on the degree distributions of MSTs. This provides a p value for the probability that the two samples are drawn from the same distribution. The results for different sets are collected in Table III and they largely confirm the conclusions based on visual inspection of Figs. 10–13. First, we see that the structure of the MSTs at the extremes of the intraday scale range are markedly different in set A and become increasingly similar as we move to set E. Second, Table III shows that the structure of set A is similar to that of other sets at the same time horizon at $\tau=1$ day (op-cl), but this is not true at smaller time horizons.

We also compared the MSTs with random MST (r-MST) generated by uncorrelated random walks of the same length. This reveals that, apart from set A, we are not able to detect any statistical feature in the degree distribution which differentiates the MSTs of sets B, C, D, and E at $\tau=1$ day (op-cl) from those generated by pure noise. Even the diameter of the MSTs is not able to discriminate them from those generated

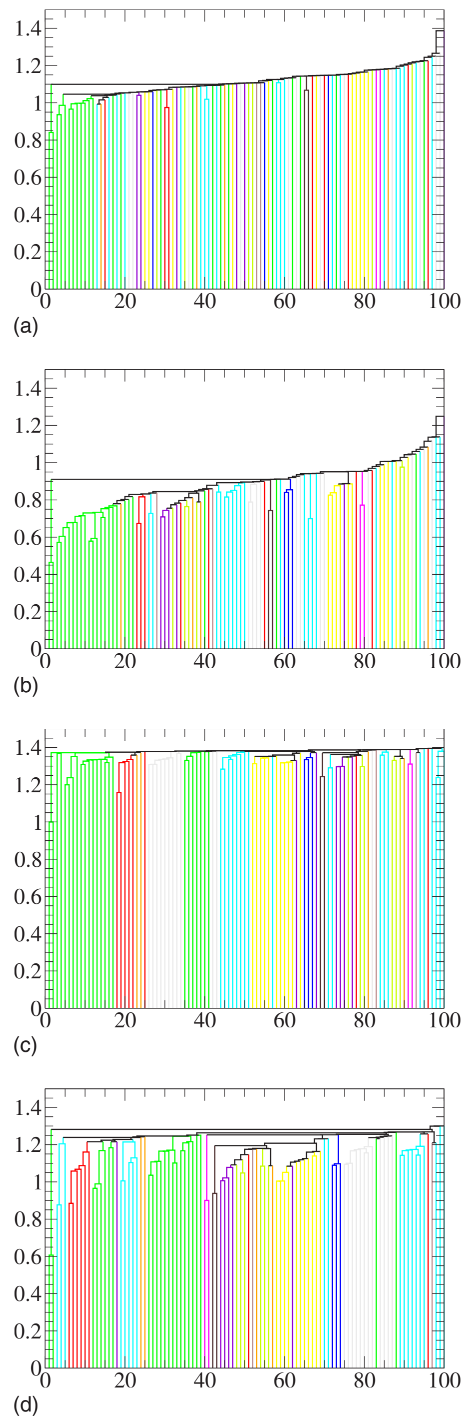

FIG. 9. (Color online) HT for set A [panels (a) and (e)] and E [panels (c) and (d)] of NYSE at $\tau=5$ min [panels (a) and (c)] and at daily (op-cl) time horizon [panels (b) and (d)]. The vertical lines represent different stocks. For each stock, colors refer to its economic sector of activity, see Table I. Economic sectors of activity are defined according to the classification scheme used in the web site http://finance.yahoo.com/.

by pure noise. However, the similarity of MSTs with r-MST disappears for larger datasets of $N=500$ or $N=2000$ stocks of NYSE, for which KS yields values of $p\simeq 0$ for all sets, at both $\tau=5$ min and 1 day (op-cl). Furthermore, MSTs turn out to be considerably more compact than r-MSTs. For example,

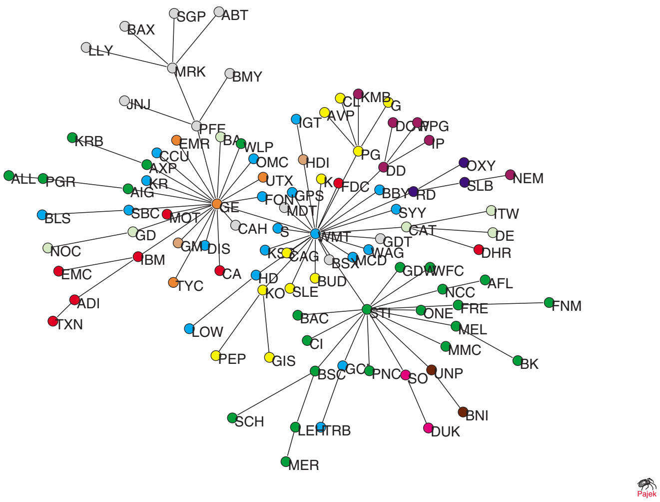

FIG. 10. (Color online) MST for set A of NYSE at τ=5 min. The vertices represent different stocks. For each stock, colors refer to its economic sector of activity; see Table I. Economic sectors of activity are defined according to the classification scheme used in the web site http://finance.yahoo.com/.

we find that with N=500 the r-MST has a diameter of 53 whereas at τ=1 day (op-cl) the largest value of the diameter is 37 for set E. Finally, in the case of N=500 stocks, for set B, set C, set D, and set E we have also performed the KS test in order to compare the degree distribution of the MSTs at 5 min and 1 day (op-cl). Such tests confirm the result of Table III, valid for N=100 stocks, that the degree distributions are essentially indistinguishable, with p values which are close to 1. Hence the removal of the "market mode" generates residues whose MSTs still contain nontrivial statistical features, although these are not clearly observable in the case of N=100 assets. When considering a larger set, say N=500, the noise threshold lowers enough to reveal a topological organization which is different from the one associated to uncorrelated random walks.

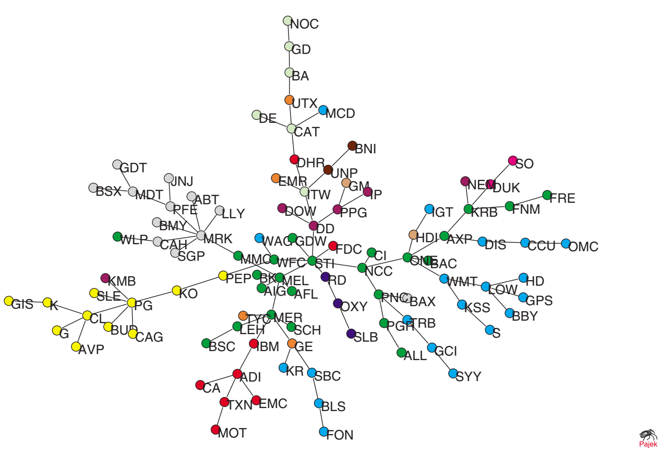

FIG. 11. (Color online) MST for set A of NYSE at τ=1 day (op-cl). The vertices represent different stocks. For each stock, colors refer to its economic sector of activity; see Table I. Economic sectors of activity are defined according to the classification scheme used in the web site http://finance.yahoo.com/.

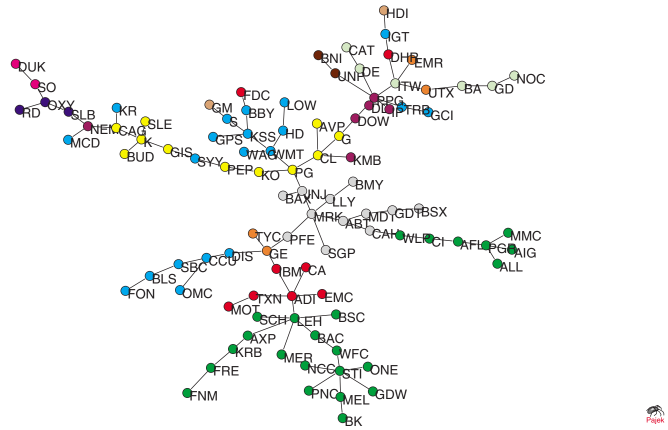

FIG. 12. (Color online) MST for set E of NYSE at τ=5 min. The vertices represent different stocks. For each stock, colors refer to its economic sector of activity; see Table I. Economic sectors of activity are defined according to the classification scheme used in the web site http://finance.yahoo.com/.

In Fig. 14 we show the HTs relative to set A [panel (a)] and set E [panel (b)] in the case when the overnight time horizon is considered. The structure of such trees is different form the ones at intraday time-horizons. In particular, for set A, the HT of Fig. 14 shows that some stocks are highly correlated with each other. However, the organization in economic sectors of activity is in general less marked than in the corresponding HT at the daily time horizon, see Fig. 9. Such effect is also observable when considering set E, i.e. panel (b) of Fig. 14. Here the average level of correlation increases, as expected. It is therefore evident that at the overnight time horizon the organization of stocks in clusters is different than at intraday time horizons, i.e., when the market is open.

Even though the topology of the MSTs has no specific statistical features, the distribution of stocks on them is definitely not random. Indeed, the cluster structure seen in the HTs of Fig. 9 corresponds to the fact that companies belong-

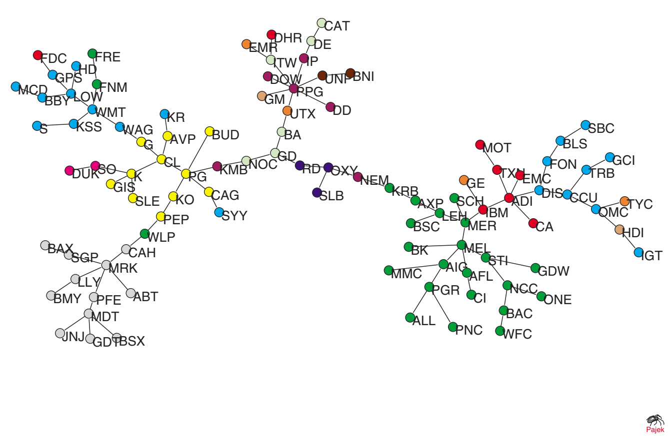

FIG. 13. (Color online) MST for set E of NYSE at τ=1 day (op-cl). The vertices represent different stocks. For each stock, colors refer to its economic sector of activity; see Table I. Economic sectors of activity are defined according to the classification scheme used in the web site http://finance.yahoo.com/.

TABLE III. p values of the Kolmogorov-Smirnov test on the degree distribution of MSTs for different datasets and time horizons in NYSE. The first row compares the MSTs at $\tau'=5$ min and $\tau=1$ day (op-cl) in different datasets $X=A,\ldots,E$. The second (third) row compares the structure of the MST of set A at $\tau=5$ min (1 day) with the MSTs in different datasets at the same horizon τ. The fourth (fifth) row compares the MSTs of sets $A,\ldots,E$ at $\tau=5$ min (1 day) with one generated by a random sample of $N=100$ random walks of the same length. The last two rows report the diameters of the MSTs at $\tau=5$ min and 1 day (op-cl). These should be compared with the diameter $d=22\pm3$ of a r-MST generated by uncorrelated random walks.

$X=A,B,C,D,E$	A_τ	B_τ	C_τ	D_τ	E_τ
$X_{\tau=5\text{ min}}, X_{\tau=1\text{ day}}$	0.031	0.677	0.961	0.992	1.000
$X_{\tau=5\text{ min}}, A_{\tau=5\text{ min}}$	1.000	0.047	0.021	0.001	0.000
$X_{\tau=1\text{ day}}, A_{\tau=1\text{ day}}$	1.000	0.794	0.894	0.677	0.794
$X_{\tau=5\text{ min}}, R_{\tau=5\text{ min}}$	0.000	0.443	0.677	0.794	0.992
$X_{\tau=1\text{ day}}, R_{\tau=1\text{ day}}$	0.556	1.000	1.000	1.000	1.000
d at $\tau=5$ min	8	17	16	23	23
d at $\tau=1$ day	15	26	22	22	25

ing to the same economic sector appear clustered in the same region of the MST. Such organization, when the market mode is removed, is revealed already at the 5-min time horizons. It is worth remarking, though, that the location of sectors themselves along the tree is different at 5 min and at the intraday scale. In other words, the intrasector structure evolves in time, while the sector composition remains stable.

In order to give a quantitative description of this effect, for each set and at each time horizon we have measured the fraction of the MST links that are conserved with respect to the open-to-close case. The results are reported in Fig. 15. The top panel refers to the case when all links in the MST are considered. The other two panels refer to the case when we also use the information about the economic sectors of activity; see Table I. In particular, we consider only intrasector links (middle panel) or only intersector links (bottom panel).

When we consider all links (top) or only those between stocks in the same sector (middle), the subtraction of the center of mass yields, apart from one exception, a higher fraction of conserved links with respect to set A. The middle panel of Fig. 15 shows that 70–80% of the MST links between stocks belonging to the same economic sector are conserved with respect to the open-to-close case, whereas a much smaller fraction is conserved between stocks belonging to different economic sectors. This is consistent with our observation that while sector composition remains stable, intrasector correlations evolve with the time horizon. Moreover, such results are also consistent with the ones shown in Fig. 5 that the amount of economic information contained in the clusters is constant.

In this respect, the botton panel of Fig. 15 shows that set D and set E reveal better than the others the topogical organization within different economic sectors at all time horizons. Finally, it is worth remarking that set C, where the market mode is *exogenously* given by the SP500 index, gives results which are comparable with those of set A.

Summarizing, the SLCA analysis of sets A, B, C, D, and E shows that (i) the removal of the center of mass reveals the organization of the sectors within different economic sectors even at small time horizons and (ii) this is better achieved in set D and set E, where the center of mass is *endogeneously* obtained either by miminizing the χ^2 function of Eq. (6) or by using a mere return market average. Finally, we find that the degree distributions of the MST at different time horizons are statistically the same, especially in set E, according the the Kolmogorov-Smirnov test, but they cannot be distinguished from those of a set of N independent random walks, for such a small market ($N=100$). The distribution of stocks on the MST reflects the organization of stocks in economic

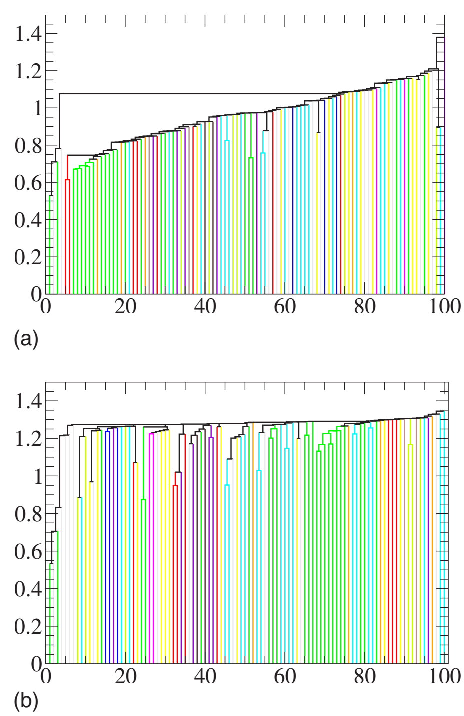

FIG. 14. (Color online) HT for set A (a) and E (b) of NYSE at overnight time horizon. The vertical lines represent different stocks. For each stock, colors refer to its economic sector of activity; see Table I. Economic sectors of activity are defined according to the classification scheme used in the web site http://finance.yahoo.com/.

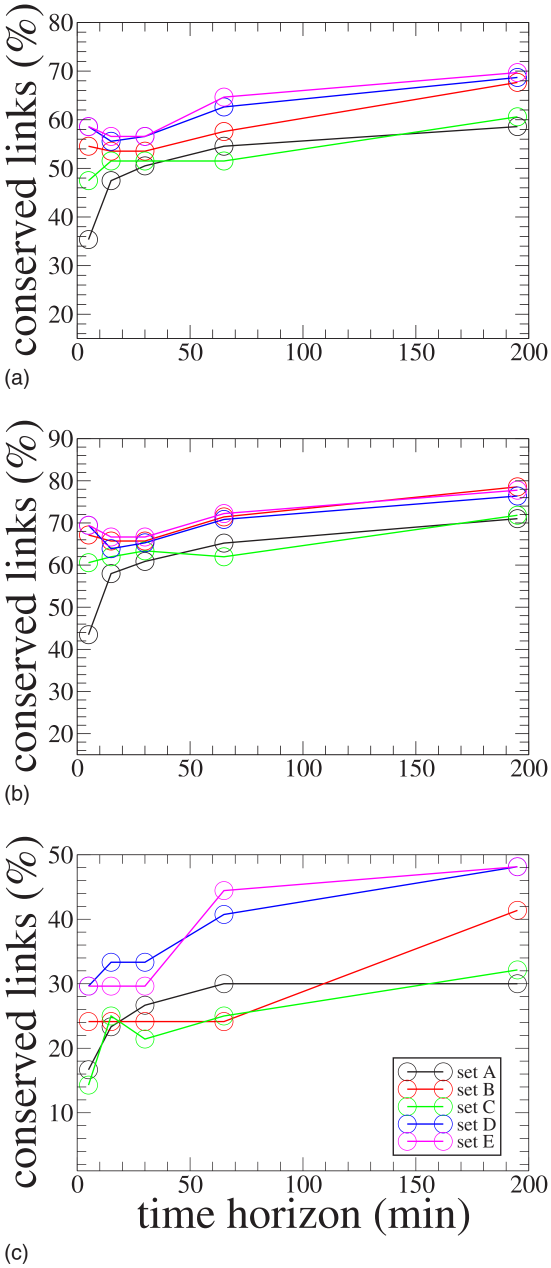

FIG. 15. (Color online) Fraction of the intraday MST links that are conserved with respect to the open-to-close case in the NYSE data. We report the cases where we consider all the links (a), only intrasector links (b) or only intersector links (c). Economic sectors of activity are defined according to the classification scheme used in the web site http://finance.yahoo.com/.

sectors, and indeed links between companies in the same sector are "conserved" across time scales.

B. Other markets

The question arises whether the above results have some degree of universality or whether they are peculiar to the NYSE market. We have therefore repeated the above investigations for different markets, i.e., for LSE, PB, and BI. Generally we confirmed the main conclusions: we find that HT of sets B, C, D, and E reveal the organization of stocks in economic sectors better than set A, and that the structure of HTs for the former is less dependent on the time horizon τ than for the latter. The structure of MSTs has a clear evolution in set A as the time horizon increases (e.g., KS test yields $p_{LSE}=0.051$ for the degree distributions of MSTs of set A between $\tau=5$ min and 1 day), whereas it has a remarkably stable structure in the other sets (particularly for set E, for which $p_{LSE}=0.999$ between $\tau=5$ min and 1 day). A comparison of the MST for set E for NYSE and LSE yields a KS test value of $p>0.9$ for all time horizons τ and similar results were found comparing NYSE and PB or BI MSTs. This apparent universality is rather a consequence of the fact that the topology of MSTs for $N\approx 100$ stocks or less is dominated by noise. Indeed, as for NYSE, the topology of MSTs is indistinguishable from that of r-MSTs generated from uncorrelated random walks.

When the market is open, the location of stocks on the MSTs, as in NYSE, is consistent with economic classification, across time horizons. In Fig. 16 we report, for different sets and at each time horizon, the fraction of the MST links that are conserved with respect to the open-to-close case for LSE (left), PB (middle), and BI (right). Again, we consider all the links (top), only intrasector links (middle), or only intersector links (bottom). As much as in the NYSE case, the sectors considered here are the economic sectors of activity mentioned above. In the case of LSE data the results are less sharp than in the NYSE case. Set B, set D, and set E give results which are more similar to each other with respect to the NYSE case. One possible exception is given by set B at 5-min time horizon. In all cases, it is confirmed that the removal of the market mode reveals the organization of stocks in economic sector already at small time horizons. As an example, the fraction of conserved intrasector links in set E is always ranging between 50% and 60%, while in set A such percentage drops to 30% at the smallest time horizon. At larger time scales. however, the fraction of conserved links for set A has roughly the same value as for other sets. This is different from what we found for NYSE, where the fractions of conserved links were systematically smaller for set A than for other sets.

When considering the overnight time horizon, we confirm that the organization of stocks in economic sectors of activity is less evident than in the case when the market is open. However, such differences are less marked than in the NYSE case.

V. CONCLUSIONS

We found that removing the dynamics of the center of mass (i) decreases the level of correlations and (ii) makes the cluster structure more evident. Naïvely one would expect that reducing the level of correlations reduces the "signal"

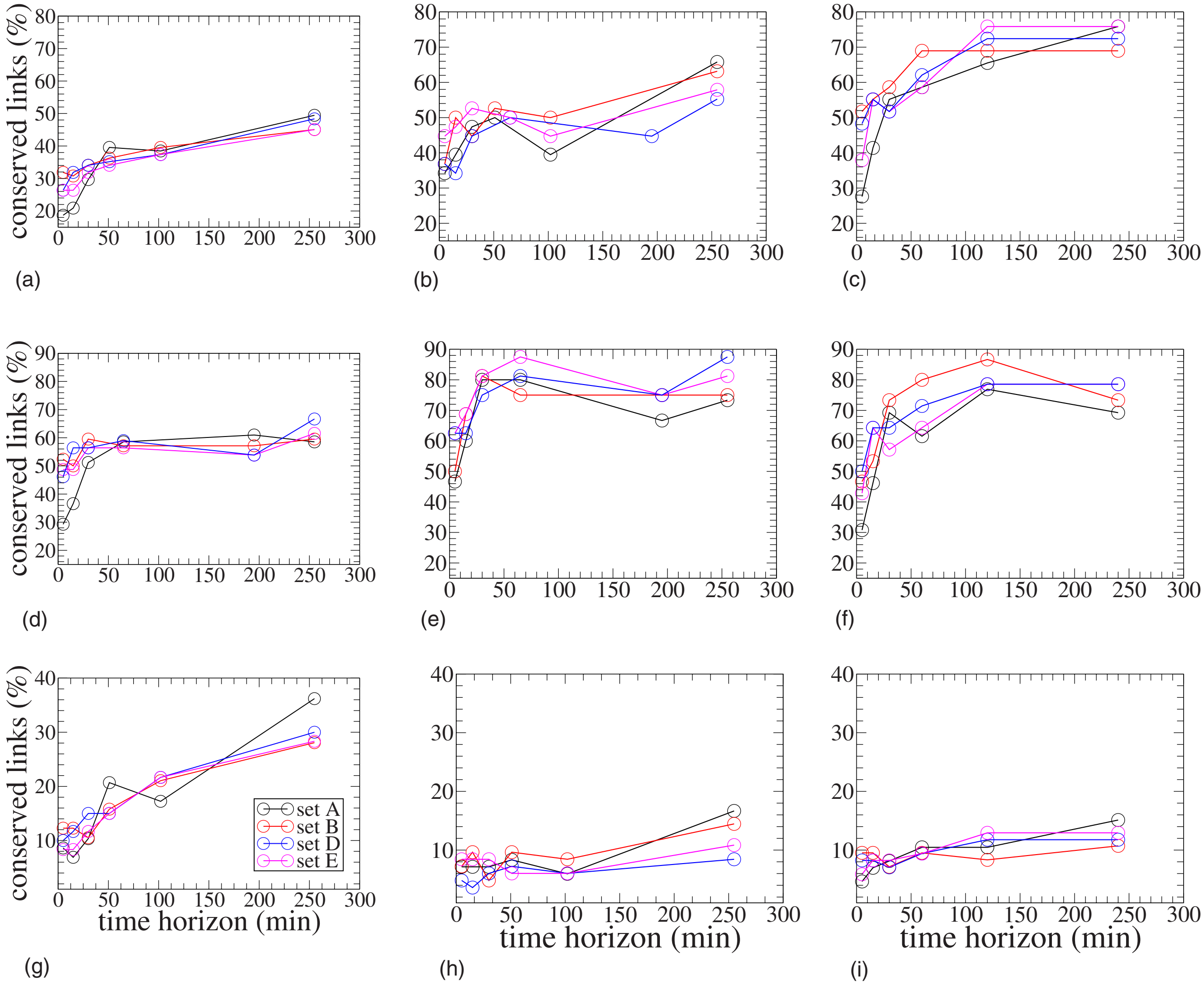

FIG. 16. (Color online) Fraction of the intraday MST links that are conserved with respect to the open-to-close case in the LSE (left), PB (middle), and BI (right) data. We report the cases where we consider all the links (top), only intrasector links (middle), or only intersector links (bottom). Economic sectors of activity are defined according to the classification scheme used in the web site http://www.euroland.com/.

and hence enhances the role of noise in the dataset. On this ground, one might expect a less sharply defined structure, i.e., the opposite of (ii). The fact that we observe (i) and (ii) implies that the market mode dynamics bears little or no information on the market structure. It also suggests that the market mode dynamics and the dynamics of "internal coordinates" are to a large extent separable, in much the same manner as in particle systems of classical mechanics, where the center of mass dynamics accounts for the effect of external forces, whereas relative coordinates respond to internal forces arising from interparticle potentials. In that context, the separation of the dynamics of the center of mass is ultimately a consequence of translation invariance in space. A similar symmetry exists in financial markets and in economics in general and it is related to the undeterminacy of the value of money, see, e.g., [24]: rescaling money and all prices by the same constant λ—which is equivalent to translation invariance in log prices—is expected to leave all economic forces invariant. We believe our results can be interpreted as one of the consequences of such invariance principle.

It is not difficult to imagine components of trading activity which might contribute to the dynamics of the center of mass or to relative coordinates. It is worthwhile to remark, in this respect, that a simple phenomenological model for the dynamics of the market mode, taking into account the impact of trading in risk minimization strategies, has been recently proposed [25]. Besides reproducing the main statistical properties of the dynamics of the largest eigenvalue of the covariance matrix, this model also shows that the behavior of the market mode is largely insensitive to a finer structure of correlations. The invariance of the structure of internal correlations down to time scales of minutes suggests that the dynamical origin of financial correlations has to be found at small time scales. However, its similarity with the classification of assets in economic sectors also suggests that trading activity is also driven by the dynamics of the economy, whose evolution takes place on much longer time scales. In this respect, the finding of a scale-invariant correlation structure is highly nontrivial.

The scale invariance of correlation structure might have important implications for risk management, because it suggests that correlations on short time scales might be used as a proxy for correlations on longer time horizons. If the structure of correlations at short time scales can be computed using shorter time series, this might allow us to detect structural changes more efficiently.

Finally, uncovering the dynamical origin of such a complex phenomenology poses exciting challenges to theoretical modeling of multiasset markets.

ACKNOWLEDGMENTS

Authors acknowledge support from research project MIUR 449/97 "High frequency dynamics in financial markets," M.M. acknowledges support from EU-STREP project no. 516446 COMPLEXMARKETS. S.M. acknowledges support from MIUR-FIRB RBNE01CW3M "Cellular Self-Organizing nets and chaotic nonlinear dynamics to model and control complex systems," from the EU-STREP projects no. 012911 "Human behavior through dynamics of complex social networks: An interdisciplinary approach." We wish to thank Dr. Claudia Coronnello for assistance in the preparation of data.

[1] J.-P. Bouchaud and M. Potters, *Theory of Financial Risk and Derivative Pricing: From Statistical Physics to Risk Management* (Cambridge University Press, Cambridge, England, 2003).

[2] R. N. Mantegna and E. Stanley, *Introduction to Econophysics* (Cambridge University Press, Cambridge, England, 1999).

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

[4] V. Plerou, P. Gopikrishnan, B. Rosenow, L. A. Nunes Amaral, and H. E. Stanley, Phys. Rev. Lett. **83**, 1471 (1999).

[5] G. Bonanno, G. Caldarelli, F. Lillo, and R. N. Mantegna, Phys. Rev. E **68**, 046130 (2003).

[6] L. Giada and M. Marsili, Phys. Rev. E **63**, 061101 (2001); Physica A **315**, 57 (2002).

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

[8] J.-P. Onnela, A. Chakraborti, K. Kaski, J. Kertesz, and A. Kanto, Phys. Rev. E **68**, 056110 (2003).

[9] G. Bonanno, F. Lillo, and R. N. Mantegna, Quant. Finance **1**, 96 (2001).

[10] J. Kwapien, S. Drodz, and J. Speth, Physica A **337**, 231 (2004).

[11] M. Potters, J.-P. Bouchaud, and L. Laloux, J. Stat. Mech.: Theory Exp. 2005, P08010.

[12] T. W. Epps, J. Am. Stat. Assoc. **74**, 291 (1974).

[13] B. Toth and J. Kertesz, e-print arXiv:0704.1099.

[14] C. Coronnello, M. Tumminello, F. Lillo, S. Miccichè, and R. N. Mantegna, Acta Phys. Pol. B **36**, 2653 (2005); M. Tumminello, T. Di Matteo, T. Aste, and R. N. Mantegna, Eur. Phys. J. B **55**, 209 (2007); M. Tumminello, C. Coronnello, F. Lillo, S. Miccichè, and R. N. Mantegna, Int. J. Bifurcation Chaos Appl. Sci. Eng. (to be published).

[15] http://www.nysedata.com

[16] M. M. Dacorogna, R. Gencay, U. A. Müller, R. B. Olsen, and O. V. Pictet, *An Introduction to High-Frequency Finance* (Academic Press, New York, 2001).

[17] http://www.rebuildorderbook.com

[18] http://www.euronext.com

[19] http://www.borsaitaliana.it

[20] For the NYSE data we considered the classification scheme given in the web site http://finance.yahoo.com/, for the LSE and BI data we considered the classification scheme used in the web site http://www.euroland.com, for the PB data we considered the classification scheme used in the web site http://www.euronext.com/.

[21] W. H. Press, S. A. Teukolsky, W. T. Veterling, and B. P. Flannery, *Numerical Recipes in Fortran: The Art of Scientific Computing*, 2nd Ed. (Cambridge University Press, Cambridge, England, 1992).

[22] R. N. Mantegna, Eur. Phys. J. B **11**, 193 (1999).

[23] G. Bonanno, G. Caldarelli, F. Lillo, S. Miccichè, N. Vandewalle, and R. N. Mantegna, Eur. Phys. J. B **38**, 363 (2004).

[24] P. Bak, S. F. Norrelykke, and M. Shubik, Phys. Rev. E **60**, 2528 (1999).

[25] G. Raffaelli, M. Marsili, JSTAT L08001 (2006).

[26] The time series $x_i(t)$ are derived from $a_i^{(\tau)}(t),\ldots,e_i^{(\tau)}(t)$ by normalization to zero average and unit variance, e.g., $x_i(t)=[a_i^{(\tau)}(t)-\langle a_i^{(\tau)}\rangle]/\sqrt{\langle[a_i^{(\tau)}-\langle a_i^{(\tau)}\rangle]^2\rangle}$.

[27] We choose this simple option, rather than more elaborate maximization procedures based, e.g., on simulated annealing, because for the data sets used the optimal configuration we find depend only marginally on the algorithm used.

[28] Labels are sorted with respect to their contribution to the log likelihood, e.g., $s_i^{(\tau)}=3$ means that asset i belongs to the third most relevant cluster at time horizon τ.