Minimum spanning tree filtering of correlations for varying time scales and size of fluctuations

Jarosław Kwapień,[1,*] Paweł Oświęcimka,[1] Marcin Forczek,[1] and Stanisław Drożdż[1,2]
[1] Institute of Nuclear Physics, Polish Academy of Sciences, ul. Radzikowskiego 152, 31-342 Kraków, Poland
[2] Faculty of Physics, Mathematics and Computer Science, Cracow University of Technology, ul. Podchorążych 1, 30-084 Kraków, Poland
(Received 19 October 2016; revised manuscript received 31 March 2017; published 17 May 2017)

Based on a recently proposed q-dependent detrended cross-correlation coefficient, ρ_q [J. Kwapień, P. Oświęcimka, and S. Drożdż, Phys. Rev. E **92**, 052815 (2015)], we generalize the concept of the minimum spanning tree (MST) by introducing a family of q-dependent minimum spanning trees (qMSTs) that are selective to cross-correlations between different fluctuation amplitudes and different time scales of multivariate data. They inherit this ability directly from the coefficients ρ_q, which are processed here to construct a distance matrix being the input to the MST-constructing Kruskal's algorithm. The conventional MST with detrending corresponds in this context to $q = 2$. In order to illustrate their performance, we apply the qMSTs to sample empirical data from the American stock market and discuss the results. We show that the qMST graphs can complement ρ_q in disentangling "hidden" correlations that cannot be observed in the MST graphs based on $\rho_{\rm DCCA}$, and therefore, they can be useful in many areas where the multivariate cross-correlations are of interest. As an example, we apply this method to empirical data from the stock market and show that by constructing the qMSTs for a spectrum of q values we obtain more information about the correlation structure of the data than by using $q = 2$ only. More specifically, we show that two sets of signals that differ from each other statistically can give comparable trees for $q = 2$, while only by using the trees for $q \neq 2$ do we become able to distinguish between these sets. We also show that a family of qMSTs for a range of q expresses the diversity of correlations in a manner resembling the multifractal analysis, where one computes a spectrum of the generalized fractal dimensions, the generalized Hurst exponents, or the multifractal singularity spectra: the more diverse the correlations are, the more variable the tree topology is for different q's. As regards the correlation structure of the stock market, our analysis exhibits that the stocks belonging to the same or similar industrial sectors are correlated via the fluctuations of moderate amplitudes, while the largest fluctuations often happen to synchronize in those stocks that do not necessarily belong to the same industry.

DOI: 10.1103/PhysRevE.95.052313

I. INTRODUCTION

A minimum spanning tree (MST) is a subgraph of a weighted network that minimizes the sum of the edge lengths while spanning the network and not containing any cycles. It is often used in optimization and in multivariate data analysis to visualize the key properties of a network representation of the data if the corresponding complete network cannot be shown in a transparent way. Many examples of MST applications can be found in the literature (see, e.g., [1–5]), but they are particularly common in econophysics [3,6–31]. Frequently, a network analysis is used to describe a correlation structure among a number of observables measured simultaneously. Such observables are represented by nodes, and correlations by the weighted edges. These weights can be expressed by any correlation measure, for example, the Pearson coefficient or mutual information. An MST can then be built by using Kruskal's [32] or Prim's [33] algorithm provided the weights have been transformed into Euclidean distances.

One of the key problems that arises while dealing with empirical data is how to overcome nonstationarity. Standard correlation measures, like the ones already mentioned, are highly prone to instability caused by large fluctuations of the data [34] and therefore results obtained with such measures are not fully reliable. Inevitably, the same weakness is inherited by networks built on top of those measures. However, as the nonstationarities are typically associated with trends, simultaneous detrending on different scales can produce stationary data. The detrended fluctuation analysis (DFA) is the most frequently used method in this context [35]. Its outcome, a fluctuation function, plays the role of variance for nonstationary signals. Analogously, the role of covariance is played by the outcome of the detrended cross-correlation analysis (DCCA) [36]. Based on both methods, the so-called detrended cross-correlation coefficient $\rho_{\rm DCCA}$ was introduced, which is a counterpart of the Pearson correlation coefficient [37]. It can be used in the case of nonstationary data to quantify the strength of cross-correlations between detrended signals at a given time scale [37–42].

In statistical analysis, variance and covariance are not able to provide a complete description of the probability distribution function for a given data set and the whole family of statistical moments is necessary to accomplish this. The Pearson correlation coefficient is not sensitive to any nonlinear dependences among data. In the same manner, DFA and DCCA are used to detect power-law auto- and cross-correlations and the related fractal properties of signals, but it is impossible to use them to detect nonlinear structures, which are more complex than monofractals. Therefore, the detrended cross-correlation coefficient $\rho_{\rm DCCA}$ also has a limited power as regards the detection of higher-order statistics than a simple detrended covariance. In order to make it more potent, recently we have proposed its generalization, called the q-dependent detrended cross-correlation coefficient ρ_q ($q \in \mathcal{R}$) [43]. Its main purpose is to identify cross-correlations selectively

*jaroslaw.kwapien@ifj.edu.pl

with respect to the fluctuation amplitudes by amplifying the correlated fluctuations of a particular size, while suppressing those of other sizes. The ρ_q coefficient is based on the multiscale generalizations of DFA and DCCA, i.e., multifractal detrended fluctuation analysis (MFDFA) [44] and multifractal detrended cross-correlation analysis (MFCCA) [45]. Both these methods are principally applied to quantify multifractal structures in one or two signals, but their sensitivity to time scales makes them useful to describe also signals that are not fractal in general.

The analogy between the ρ_{DCCA} coefficient and the Pearson coefficient can be exploited straightforwardly to construct detrended minimum spanning trees [29]. Their main advantage over standard MSTs is that they are more stable with respect to time, being independent of the data nonstationarity. In our present work we want to combine this property of the detrended minimum spanning trees with the ability of ρ_q to focus on fluctuations of a specific size. As a result, we obtain a tool that is capable of quantifying the correlation structure of multivariate nonstationary signals not only with respect to time scale but also with respect to fluctuation size. As in the case of ρ_{DCCA}, the signals under study may be arbitrary, as no fractal properties are required.

Our paper is organized as follows. In Sec. II we provide the necessary definitions, in Sec. III we describe the empirical data and present results of the qMST analysis based on those data, and in Sec. IV we collect the main conclusions.

II. METHODS

Let us start from a brief description of the MFCCA approach [45]. Consider a pair of time series, $x(i)_{i=1,\dots,T}$ and $y(i)_{i=1,\dots,T}$, divided into $2M_s$ boxes of length s (i.e., M_s boxes starting from the opposite ends). In each box ν ($\nu = 0,\dots,2M_s - 1$), we calculate the difference between the integrated signals and the mth-order polynomials $P^{(m)}$ fitted to these signals:

$$X_\nu(s,i) = \sum_{j=1}^{i} x(\nu s + j) - P^{(m)}_{X,s,\nu}(j), \tag{1}$$

$$Y_\nu(s,i) = \sum_{j=1}^{i} y(\nu s + j) - P^{(m)}_{Y,s,\nu}(j), \tag{2}$$

where $m = 2$ typically. The covariance and the variances of X and Y in a box ν are defined as

$$f^2_{XY}(s,\nu) = \frac{1}{s}\sum_{i=1}^{s} X_\nu(s,i)Y_\nu(s,i), \tag{3}$$

$$f^2_{ZZ}(s,\nu) = \frac{1}{s}\sum_{i=1}^{s} Z^2_\nu(s,i). \tag{4}$$

Here Z can be either X or Y. Now we define a family of the fluctuation functions of order q [44,45]:

$$F^q_{XY}(s) = \frac{1}{2M_s}\sum_{\nu=0}^{2M_s-1} \mathrm{sgn}\left[f^2_{XY}(s,\nu)\right]\left|f^2_{XY}(s,\nu)\right|^{q/2}, \tag{5}$$

$$F^q_{ZZ}(s) = \frac{1}{2M_s}\sum_{\nu=0}^{2M_s-1} \left[f^2_{ZZ}(s,\nu)\right]^{q/2}. \tag{6}$$

The signum function enters the equation for $F^q_{XY}(s)$ in order to preserve the signs of the covariances $f^2_{XY}(s,\nu)$, which would otherwise be lost in the moduli that are necessary to secure real values of $F^q_{XY}(s)$. The real parameter q plays the role of a filter, by amplifying or suppressing the intrabox variances and covariances in such a way that for $q \gg 2$ only the boxes (of size s) with the highest fluctuations contribute substantially to the sums, while for $q \ll 2$ only the boxes with the smallest fluctuations do so. The special case of $q = 2$ allows us to reduce the above formulas to the form

$$F^2_{XY}(s) = \frac{1}{2M_s}\sum_{\nu=0}^{2M_s-1} f^2_{XY}(s,\nu), \tag{7}$$

$$F^2_{ZZ}(s) = \frac{1}{2M_s}\sum_{\nu=0}^{2M_s-1} f^2_{ZZ}(s,\nu), \tag{8}$$

in which all the boxes contribute to the sums with the same weights. The scale dependence of $[F^q_{XY}(s)]^{1/q}$ and $[F^q_{ZZ}(s)]^{1/q}$ can indicate a fractal character of the signals if it is power-law, but here we allow for any form of this dependence.

The q-dependent detrended cross-correlation coefficient $\rho_q(s)$ is defined by means of the qth-order fluctuation functions [43]:

$$\rho_q(s) = \frac{F^q_{XY}(s)}{\sqrt{F^q_{XX}(s)F^q_{YY}(s)}}. \tag{9}$$

For $q = 2$ Eq. (9) reduces to the definition of ρ_{DCCA} [37]. The filtering ability of $\rho_q(s)$ manifests itself in such a way that the more deviated from the value $q = 2$ the exponent q is, the more extreme fluctuations in the corresponding boxes contribute to the coefficient $\rho_q(s)$. For $q \geqslant 0$, values of ρ_q fit within the range

$$-1 \leqslant \rho_q \leqslant 1. \tag{10}$$

As in the case of the Pearson and the ρ_{DCCA} coefficients, $\rho_q = 1$ indicates a perfect correlation, $\rho_q = 0$ indicates independent signals, and $\rho_q = -1$ indicates a perfect anticorrelation. However, for $q < 0$ the coefficient ρ_q is not bound and for independent or weakly correlated signals it may happen that $|\rho_q| > 1$. In this case, an inverted value of ρ_q is considered, which maps the coefficient back into the $[-1,1]$ interval [43].

In a multivariate analysis of N parallel signals, one has to deal with $N(N-1)/2$ different coefficients ρ_q for each considered box size (time scale) s. It is thus convenient to put them in a matrix of size $N \times N$, which can be considered a matrix defining an N-node weighted network. Similar to the Pearson correlation coefficient, ρ_q is not a metric, because it does not fulfill the triangle inequality. In the standard procedure of MST construction, a matrix of the correlation coefficients is transformed into a (metric) distance matrix based on the formula $d_{XY} = \sqrt{2(1 - c_{XY})}$, where c_{XY} is the Pearson coefficient calculated for the signals X and Y [3]. However, an analogous transformation using ρ_q,

$$d^{(q)}_{XY} = \sqrt{2\left(1 - \rho^{(XY)}_q\right)}, \tag{11}$$

might not produce a metric in general.

TABLE I. Number of exceptions from the triangle inequality for 161 700 triples of the $d_{XY}^{(q)}(s)$ distances calculated for the 100 uncorrelated time series of the ARFIMA process.

	q								
s	−4.0	−3.0	−2.0	−1.0	0.0	1.0	2.0	3.0	4.0
20	0	3	2626	0	27	0	0	0	0
200	0	2	1984	80	0	0	0	0	0
2000	0	49	2160	5162	0	0	0	0	0
20 000	52	252	1955	9089	0	0	0	0	0

In order to test numerically whether the triangle inequality $d_{XY}^{(q)}(s) + d_{YZ}^{(q)}(s) \geqslant d_{XZ}^{(q)}(s)$ is valid for any triple of the signals X,Y,Z, we generate a set of $N = 100$ signals of length $T = 10^6$, being the autoregressive fractional moving average (ARFIMA) processes [46]. Next, for a given s and q we calculate ρ_q for each pair of signals (4950 pairs total) and transform its value into the distance $d_{XY}^{(q)}$. Then the inequality is tested for all possible distance triples (161 700 total) and the exceptions are counted. The same sequence of steps is repeated for different values of q ($-10 \leqslant q \leqslant 10$). The results for sample values of s and q are listed in Table I. For $q \geqslant 1$, the distances always fulfill the inequality, which means that $d_{XY}^{(q)}$ behaves like a metric. However, for $q \leqslant 0$ exceptions do occur, indicating that the distance cannot be considered a metric. We also applied the same test to randomized empirical data from a stock market (logarithmic price fluctuations of the 100 largest American companies, the same data set that is studied in Sec. III) and found qualitatively similar results: no violation of the triangle inequality for $q \geqslant 1$ and frequent violations for $q \leqslant 0$. Therefore, our further analysis is restricted to positive q's. This, however, does not limit the robustness of our approach as, in many empirical systems, the small fluctuations that are selected by the negative values of q are associated with noise.

Now we are prepared to define a q-dependent MST by using the q distance given by Eq. (11). qMST can be constructed by applying Kruskal's algorithm to $d_{XY}^{(q)}$. This algorithm consists of two essential steps: (i) sorting the elements of a distance matrix from the smallest to the largest and (ii) going from the smallest distances and connecting not yet connected nodes to the closest ones. After connecting all the nodes, the resulting tree consists of N nodes and $N-1$ weighted edges. It is so optimized that the sum of all the distances $d_{XY}^{(q)}$ is the minimum possible. The qMSTs inherit the properties of ρ_q, so by choosing a particular value of q, the related tree expresses only those correlations that are present among the fluctuations of a selected size: large ones for $q \gg 2$ and small ones for $q < 2$. However, even though for $q = 2$ this procedure of constructing the detrended MSTs gives the same weight to all fluctuation amplitudes, one has to remember that even for this value of q it differs substantially from the standard procedure, which does not involve detrending.

III. qMSTs FOR EMPIRICAL DATA

In order to present an example of a qMST analysis, we consider a set of time series representing logarithmic stock-price fluctuations (returns),

$$r_X(t,\Delta t)) = \ln p_X(t+\Delta t) - \ln p_X(t), \qquad (12)$$

for the $N = 100$ largest American companies traded on the New York Stock Exchange over the years 1998–1999 [47]. We choose a sampling interval of $\Delta t = 1$ min to obtain sufficiently long time series ($T = 203,190$). We calculate $\rho_q(s)$ for all possible pairs of stocks; a few time scales s—20, 60, and 390 min (a trading day), 1950 min (a trading week), and 7800 min (approx. a trading month); and a few values of q from $q = 1.0$ to $q = 6.0$ (although we use only integer values, arbitrary positive real values may also be used). Next, we derive the metric distances $d_{XY}^{(q)}(s)$ and apply Kruskal's algorithm to obtain a minimum spanning tree for each value of s and q (36 trees total).

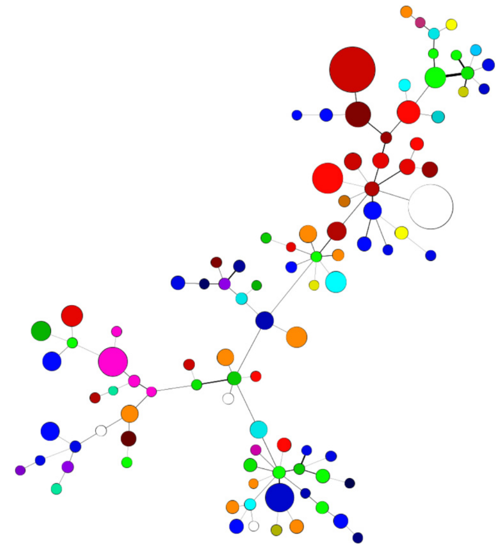

FIG. 1. A q-dependent minimum spanning tree example ($s = 20$, $q = 5.0$). The symbol size is proportional to the market capitalization of a stock on December 31, 1999, while different sectors are denoted by different colors and different subsectors within the same sector are denoted by shades of a particular color: technology (from light red to dark brown), services (from light blue to dark blue), energy (from light magenta to dark magenta), financial (from light green to dark green and olive), consumer noncyclical (from cyan to turquoise), basic materials (from yellow to greenish yellow), health care (orange and light brown), consumer cyclical (violet), capital goods (seagreen), utilities (violetred), and conglomerates (white). The edge thickness is proportional to ρ_q (the thicker it is, the more correlated the stocks are). For node captions see file Figure1.png in the Supplemental Material [48].

Figure 1 displays the qMST for the shortest time scale $s = 20$ and for the sample exponent $q = 5.0$. Different industrial sectors and subsectors are distinguished by different colors and color shades, respectively, and the node symbols' size is proportional to the market capitalizaton of the corresponding company. The higher is the calculated $\rho_q^{(XY)}$, the thicker is the edge connecting the nodes X and Y. One can see that the tree topology is somewhere between a centralized and a distributed network. There are hubs with a degree up to $k = 9$ and a number of peripheral nodes with $k = 1$, but the shortest path length is rather high: $L \approx 7.9$. The edge weights are also

TABLE II. Thresholds $\tau_\rho = \bar{\rho}_q + 2\sigma_{\rho_q}$ representing the $\rho_q(s)$ values above which we reject the null hypothesis of random correlations.

	q					
s	1.0	2.0	3.0	4.0	5.0	6.0
20	0.0155	0.0114	0.0179	0.0235	0.0248	0.0255
60	0.0261	0.0184	0.0223	0.0296	0.0316	0.0309
390	0.0733	0.0507	0.0486	0.0607	0.0703	0.0740
1950	0.1493	0.1032	0.0841	0.0837	0.0905	0.0971
7800	0.3129	0.2183	0.1840	0.1842	0.1948	0.2029

rather low, with a number of edges denoted by lines as thin as possible. Several industrial sectors are visible as clusters, but a number of nodes are connected to nodes from different sectors. This suggests that there is a significant randomness in the distribution of the nodes across this tree. The question follows immediately whether these effects are genuine or, rather, statistically insignificant and pertinent to noise.

We therefore have to estimate the statistical significance of the edge weights, at least approximately. Since our data are strongly leptokurtic [49–52], we define the null hypothesis stating that the $\rho_q(s)$ values can fully be explained by the cross-correlations between random processes with the same probability density function (p.d.f.) as the empirical data under study. To test the results against this hypothesis, we shuffle all the time series 50 times and create 50 independent sets of surrogate data. Then for each set we derive all possible values of $\rho_q^{(XY)}(s)$ for the same choices of s and q as we did before and identify the maximum values of ρ_q in each case. Next, for a given s and q, we collect all the maximum coefficients and calculate their mean $\bar{\rho}_q$ and standard deviation σ_ρ. Finally, we assume Gaussianity of the p.d.f. and set a threshold at $\tau_\rho = \bar{\rho}_q + 2\sigma_\rho$. If for some X, Y we obtain $\rho_q^{(XY)}(s) \geqslant \tau_\rho$, we consider the null hypothesis to be rejected in this case and not rejected otherwise. The specific thresholds for different values of s and q arelisted in Table II.

The so-filtered qMST graphs for $s = 20$ are shown in Fig. 2 ordered by increasing q. Three observations can be drawn from these pictures. First, the tree transforms itself from a highly centralized structure ($L \approx 2.2$) with a dominant hub (General Electric) of degree $k = 87$ and a secondary hub (Cisco) with $k = 10$ for $q = 1.0$ to a distributed, randomlike structure ($L \approx 8.2$) with a few regional hubs with $k > 4$: Yahoo! ($k = 5$), Bank of New York ($k = 6$), Morgan Stanley ($k = 5$), and Chase Manhattan ($k = 8$) for $q = 6.0$. This topologic transition occurs gradually with increasing q. Filtering out the insignificant edges has no effect at $q = 1.0$ as the tree consists of all the 100 nodes, but the effect is striking at $q = 6.0$, where the structure is disconnected, with only 50 nodes forming the main component (the main tree), while the remaining nodes either form three smaller components, one with $n = 6$ nodes and two with $n = 2$ nodes, or do not form any connections (40 nodes). From the market perspective, a centralized structure with high edge weights corresponds to a global coupling among the stocks that causes the stock prices to move collectively. Their evolution can thus be described by a one-factor model. In a correlation matrix analysis this

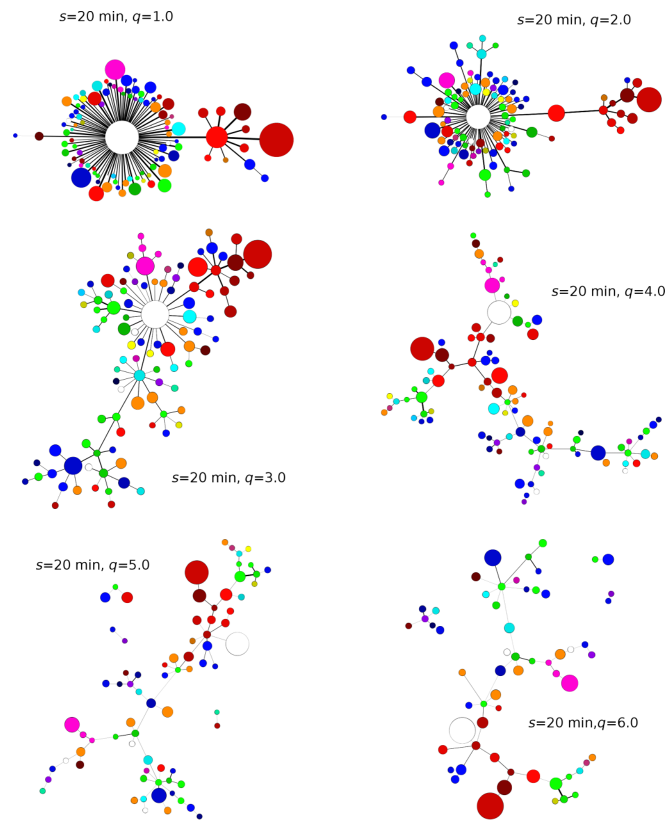

FIG. 2. The q-dependent minimum spanning trees formed for $s = 20$ and for a few values of q from 1.0 to 6.0 (from left to right and from top to bottom). Trees have been filtered to remove statistically insignificant edges ($\rho_q < \tau_\rho$) and nodes with degree $k = 0$. (Differences in the sizes of the node symbols between the trees are artifacts.) For node captions see file Figure2.png in the Supplemental Material [48].

factor manifests itself as the distant largest eigenvalue and it is often called the market factor. On the other hand, a distributed topology with a few hubs of a small degree and many disconnected nodes indicates that the market can be decomposed into clusters often related to the industrial sectors and that many stocks have independent dynamics. The eigenvalue spectrum of a correlation matrix would show a few nonrandom eigenvalues and a bulk of random ones in this case [26,34,53–55].

The second observation is that, for the 20-min time scale, the periods (the MFCCA boxes) characterized by fluctuations of moderate amplitude are universally correlated among the stocks, while the periods with the largest fluctuations are correlated only within small groups of stocks and typically they are uncorrelated. This can be understood if one realizes that it needs time to develop large movements of the whole market, and 20 min is too short to accomplish this.

Finally, the third observation from Fig. 2 is that some node clusters notably overlap with the industrial sectors, especially for $3.0 \leqslant q \leqslant 5.0$. It is the most evident for the IT cluster (red; this cluster is seen even for $q = 1.0$), the financial cluster (green), and the basic materials cluster (magenta). For the remaining sectors, their stocks are distributed across the qMST for $q \geqslant 3.0$. It is also noteworthy that since the

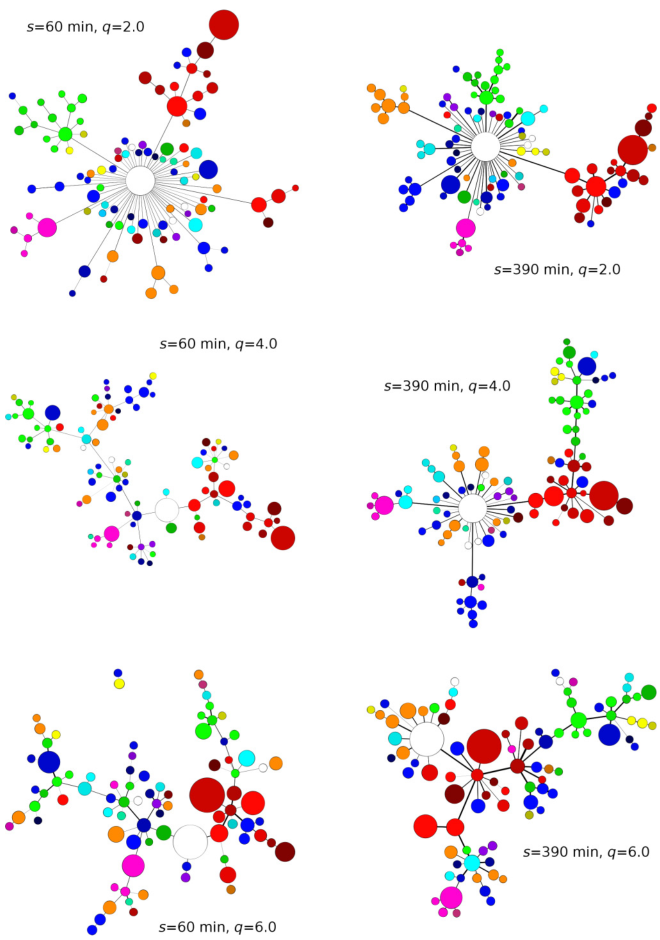

FIG. 3. The q-dependent minimum spanning trees created for the hourly ($s = 60$ min; left) and daily ($s = 390$ min; right) time scales and for $q = 2.0$ (top), $q = 4.0$ (middle), and $q = 6.0$ (bottom). Trees were filtered to remove statistically insignificant edges ($\rho_q < \tau_\rho$) and nodes with degree $k = 0$. For node captions see file Figure3.png in the Supplemental Material [48].

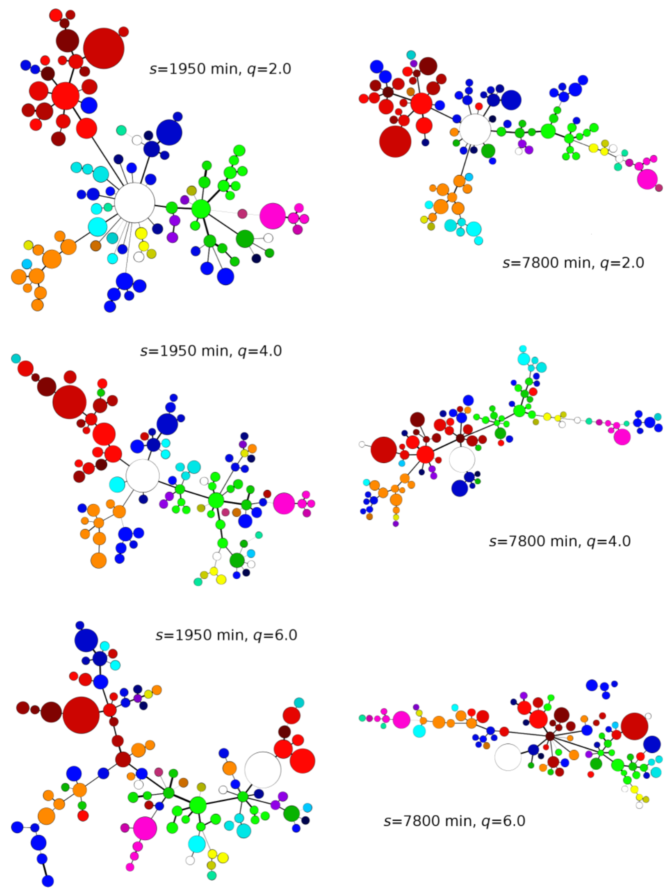

FIG. 4. The q-dependent minimum spanning trees created for the weekly ($s = 1950$ min; left) and monthly ($s = 7800$ min; right) time scales and for $q = 2.0$ (top), $q = 4.0$ (middle), and $q = 6.0$ (bottom). Trees were filtered to remove statistically insignificant edges ($\rho_q < \tau_\rho$) and nodes with degree $k = 0$. For node captions see file Figure4.png in the Supplemental Material [48].

edges shown on the filtered qMSTs are statistically significant, any restructuring of the connections while changing q can be related to uncovering some previously "hidden" correlations.

These results can be compared with those obtained for other time scales: $s = 60$ min and $s = 390$ min (Fig. 3) as well as for $s = 1950$ min. and $s = 7800$ min. (Fig. 4). For any given s, if we augment q, the main hub (the General Electric stock) shown at $q = 2.0$ loses its centrality and typically becomes a peripheral one. The other, less important hubs can either preserve their degree or lose it as well, which depends on s or individually on a node. An additional general rule is that the shorter the time scale is, the more dramatic the changes in topology observed with increasing q. The edge weights gradually become higher with increasing s (the edges become thicker) for all the exponents q, which means that the average correlation strength among the stock pairs also increases. This remains in perfect agreement with the results obtained earlier with other means, suggesting the existence of the so-called Epps effect [56–58].

It is interesting to observe the changes in tree topology for a fixed q and variable s. For example, for $q = 2.0$ there is a clear, gradual transition from a centralized one-factor structure for $s = 20$ min (Fig. 2, top right) to an almost-perfect sectorial structure for the monthly scale (Fig. 4, top right), indicating that cross-correlations between the stocks build up progressively, from the most general marketwide ones at short time scales to more specific intraindustry ones at long time scales.

From the perspective of practical application of the qMST graphs, it is important to note that, since their topology depends on q, the correlation structure of the market is different if one considers different fluctuation amplitudes. That is, with the help of such graphs one can quickly realize what this structure looks like in volatile periods and what it looks like in more typical or quiet periods. They can allow one to fine-tune the investment portfolios, for instance, in order to avoid a composition of the stocks that tend to be more strongly correlated during volatile times, which increases the portfolio's risk and which can remain unnoticed if one uses standard measures. For example, in the considered interval of time (1998–1999) for $q = 2.0$ the Corning shares were the most closely related to the General Electric ones, while for larger q's it occurs that they were even more strongly related to the AIG shares, so that a simultaneous investment in

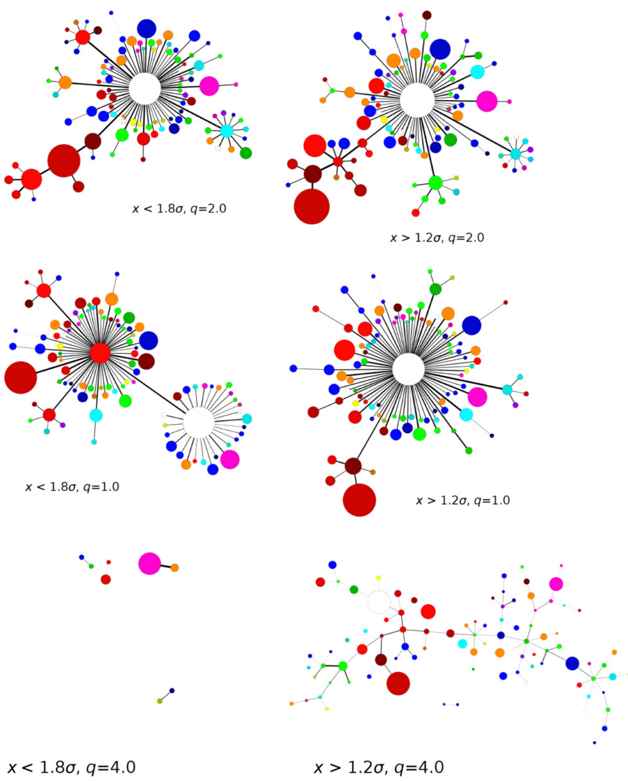

FIG. 5. The q-dependent minimum spanning trees ($\Delta t = 1$ min, $s = 20$ min) created for two data sets with different statistical properties: set 1 (large fluctuations randomized, small and medium ones preserved), on the left; and set 2 (small and medium fluctuations randomized, large fluctuations preserved), on the right. The respective qMSTs for different values of q are compared: $q = 2.0$ (top; $c_e = 51$ edges common to both sets), $q = 1.0$ (middle; $c_e = 21$), and $q = 4.0$ (bottom; $c_e = 0$). Trees were filtered to remove statistically insignificant edges ($\rho_q < \tau_\rho$) and nodes with degree $k = 0$. For node captions see file Figure5.png in the Supplemental Material [48].

both these stocks was in fact more hazardous than might have been expected from an analysis of the MST for $q = 2.0$ or an analysis using the standard MST approach.

Another situation, in which generalized qMST graphs are preferred over $q = 2$ MSTs, is illustrated in Fig. 5. We consider two data sets that differ from each other in their statistical properties. Both data sets consist of signals that were obtained from the original time series of returns ($\Delta t = 1$ min, $s = 20$ min) by partial randomization based on amplitude. Each signal was randomized independently. Set 1 comprises signals in which large fluctuations $[r(t) > 1.8\sigma]$ were shuffled among themselves, while small and medium fluctuations $[r(t) \leqslant 1.8\sigma]$ were preserved as they occurred in the signals (σ stands for standard deviation). Set 2 was produced in a similar manner, but here the small and medium fluctuations ($r(t) < 1.2\sigma$) were shuffled among themselves, while large ones $[r(t) \geqslant 1.2\sigma]$ were preserved. Such randomization destroys all the correlations carried by the fluctuations of the respective amplitudes. Thus, only correlations within the range $(1.2\sigma, 1.8\sigma)$ remained common to both sets, while correlations outside that range were different in each set.

Despite that such differences exist, the qMSTs for $q = 2$ reveal comparable topologies for both data sets as indicated in the top panels in Fig. 5. To express this quantitatively, we calculated how many edges are common to both trees (by a common edge we mean an edge connecting the same pair of nodes in both cases): the respective number is $c_e = 51$ for $q = 2$. This indicates that the $q = 2$ MSTs are not sufficiently sensitive to differentiate between the sets. In contrast, the differences appear more significant for both $q = 1$ (middle panels; $c_e = 21$ common edges) and $q = 4$ (bottom panels; no common edges). In the latter case, the discrepancy is particularly striking as the tree for set 1 is essentially random and virtually disappears when tested against a null hypothesis of no genuine correlations, while almost the whole tree structure for set 2 is preserved as being statistically significant. In fact, we expected this result, because for $q = 4$ only those parts of the signals are taken into consideration that comprise many high-amplitude fluctuations in any pair of the signals, and after reshuffling, the large fluctuations become decorrelated in set 1, while they are still correlated in set 2. Therefore, in this situation, applying the generalized qMSTs proved advantageous since it allowed us to differentiate between the analyzed data sets.

Topological variability of the qMST graphs for a range of q may also be considered a property of the data, expressing stability of the correlation structure at different fluctuation amplitudes. The more static the topology of qMSTs for different q's is, the broader range of fluctuations is correlated in the same manner. Figure 6 presents such an application of the qMST graphs. Trees calculated for the original set of time series of returns (top panels) are compared with trees calculated for time series of the Gaussian returns (middle panels) and time series of the return signs (bottom panels). While the original returns have a p.d.f. with the power-law tails with a scaling index $\alpha \approx -3$, the Gaussian returns were obtained from the original ones by suppressing their amplitudes in such a way that their p.d.f. became Gaussian, while the amplitude ranks of the fluctuations remained preserved [59]. The sign series were obtained by replacing the returns with their signs, i.e., $r'(t) = 1$ if $r(t) > 0$, $r'(t) = -1$ if $r(t) < 0$, and $r'(t) = 0$ otherwise. For a larger effect, in Fig. 6 we compare qMSTs at $q = 1$ (left) and $q = 6$ (right). As the most variable (i.e., different for different stock pairs) correlations are observed in the original data for the largest fluctuations (only $c_e = 2$ edges are common to $q = 1$ and $q = 6$), suppressing these fluctuations either by Gaussianizing their p.d.f. or by considering only their signs leads to a more uniform strength of the correlations ($c_e = 71$ and $c_e = 65$ common edges, respectively). Thus, the qMSTs presented in the top panels in Fig. 6 have much more heterogeneous edge weights than their counterparts shown in the middle and bottom panels.

In this context, one can view the parameter q in qMST in the same way as it is commonly viewed in studies using the Rényi entropy or in multifractal analysis, where one computes a spectrum of the generalized fractal dimensions D_q, the generalized Hurst exponents H_q, or the multifractal spectrum $\tau(q)$ [60]. By applying different values of q, one selects certain portions of the analyzed data and acquires some insight into their properties. Here, by varying q, we obtain a family of trees describing correlations between different parts

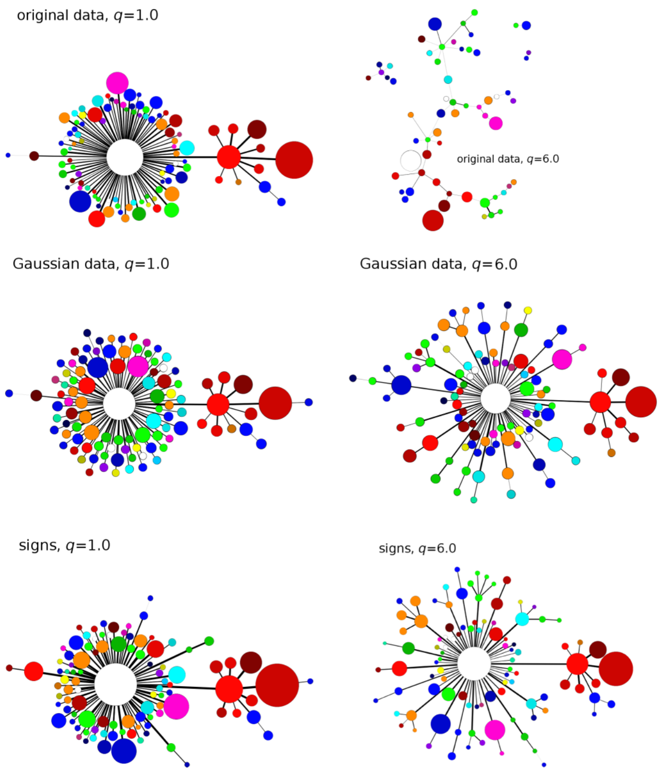

FIG. 6. The q-dependent minimum spanning trees created for three data sets ($N = 100$): the original time series of stock returns ($\Delta t = 1$ min, $s = 20$ min; top), the time series of Gaussianized returns (middle), and the time series of return signs (bottom). Two values of q are used—$q = 1.0$ (left) and $q = 6.0$ (right)—and the number of edges common to the respective qMSTs is computed: $c_e = 2$ (original returns), $c_e = 71$ (Gaussian returns), and $c_e = 65$ (return signs). Trees were filtered to remove statistically insignificant edges ($\rho_q < \tau_\rho$) and nodes with degree $k = 0$. For node captions see file Figure6.png in the Supplemental Material [48].

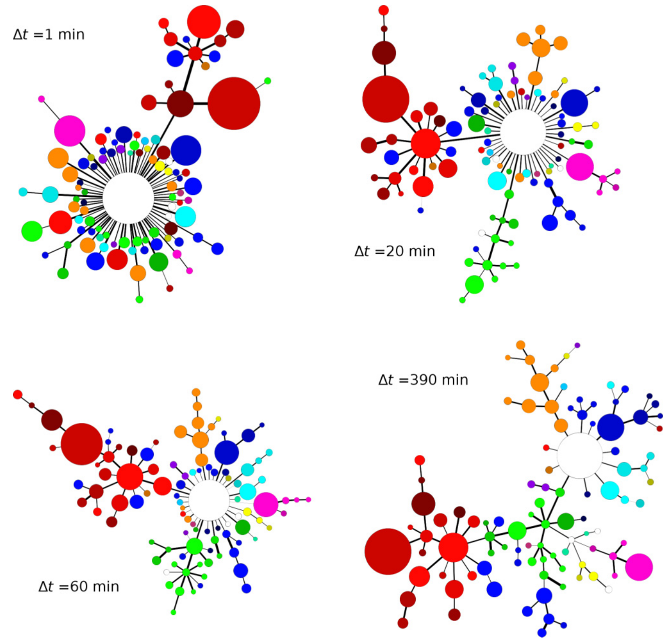

FIG. 7. Minimum spanning trees constructed in the standard way by using the Pearson correlation coefficient for time series of returns with $\Delta t = 1$ min (top left), 20 min (top right), 60 min (bottom left), and 390 min (bottom right). For node captions see file Figure7.png in the Supplemental Material [48].

of signals based on standard deviations of their fluctuations. This makes the qMST graphs a possible method of quantifying the heterogeneity of correlations in a way resembling that in which the singularity spectra $f(\alpha)$ quantify the richness of multifractal structures.

For the sake of comparison, we also present the MSTs calculated in accordance with their standard definition, i.e., by using the Pearson correlation coefficients [3] for nondetrended time series of returns. A straightforward choice of the data sampling interval could be the same $\Delta t = 1$ min as before, but one has to realize that there is not any direct counterpart of the temporal scale s in this case. The difference is that in DFA-related approaches, the temporal scale is determined by a window size, in which detrending is performed and in which the variance of the residual signal is calculated, while in the standard approach the time scale is associated directly with the sampling interval Δt. Therefore, here we have to check different values of Δt—1, 20, 60, and 390 min—and to construct one MST for each of them. Figure 7 shows the resulting trees. As usual in such analyses, by increasing the time scale, we obtain more and more visible clusters that may be identified with the market sectors. By comparing the present trees with the ones in Figs. 2, 3, and 4, it is evident that they resemble more the centralized trees at $1.0 \leqslant q \leqslant 3.0$ than the trees at $q \geqslant 4.0$, which were significantly less centralized.

In order to compare the results of both methods in a quantitative manner, we prefer to consider the matrices consisting of all the respective coefficients for all $N(N-1)/2 = 4950$ pairs of stocks instead of considering only the minimal spanning trees. This is because by taking all possible coefficients into consideration, we will increase the statistical significance of our comparison. Therefore, here we take the values of ρ_q and the Pearson correlation coefficients c for all pairs of signals and investigate how similar the sets are for different choices of the parameters Δt, s, and q. For each choice, we order both coefficients into vectors in such a way that, first, we take $N-1$ coefficients (of the same type) involving a stock X_1, then $N-2$ coefficients involving a stock X_2, then $X-3$ coefficients involving a stock X_3, and so on, until, finally, the coefficient for stocks X_{N-1} and X_N.

By proceeding in this way, we eventually obtain two vectors, $\mathbf{c}(\Delta t)$ and $\boldsymbol{\rho}_q(s)$, for each Δt, q, and s. [Actually, the coefficient $\rho_q(s)$ also depends on some sampling time $\Delta t'$, but as we do not change the latter and consequently use the data sampled at $\Delta t' = 1$ min, we neglect $\rho_q(s)$ dependence on $\Delta t'$.] Then we calculate the normalized scalar product of these vectors,

$$P(\Delta t,s,q) = \frac{\sum_{m=1}^{4950} c_m(\Delta t)\rho_q^{(m)}(s)}{|\mathbf{c}(\Delta t)||\boldsymbol{\rho}_q(s)|}, \qquad (13)$$
$$-1 \leqslant P(\Delta t,s,q) \leqslant 1,$$

where the vector components labeled with m correspond to unique stock pairs ($m = 1,\ldots,4950$). We prefer the

TABLE III. Values of the normalized scalar product for 4 Pearson-coefficient-based ($\Delta t = 1, 20, 60, 390$ min) and 24 ρ_q-based ($s = 20, 60, 390, 7800$ min, $q = 1.0, 2.0, 3.0, 4.0, 5.0, 6.0$) weighted networks with nodes representing the 100 largest companies. Values greater than 0.950 are distinguished in boldface.

		q					
Δt	s	1.0	2.0	3.0	4.0	5.0	6.0
1	20	**0.97**	**0.98**	0.92	0.61	0.33	0.20
	60	**0.96**	**0.96**	0.95	0.86	0.70	0.53
	390	0.95	0.95	0.94	0.90	0.84	0.77
	7800	0.89	0.87	0.83	0.77	0.71	0.66
20	20	**0.96**	**0.95**	0.86	0.53	0.28	0.17
	60	**0.97**	**0.97**	0.95	0.84	0.68	0.52
	390	**0.98**	**0.98**	**0.96**	0.92	0.86	0.79
	7800	0.94	0.92	0.88	0.82	0.77	0.71
60	20	**0.95**	0.95	0.86	0.55	0.29	0.19
	60	**0.97**	**0.97**	0.95	0.85	0.69	0.53
	390	**0.98**	**0.99**	**0.98**	0.94	0.88	0.81
	7800	**0.97**	**0.96**	0.93	0.89	0.83	0.78
390	20	0.91	0.90	0.83	0.53	0.28	0.19
	60	0.93	0.93	0.91	0.83	0.70	0.53
	390	**0.95**	**0.95**	0.95	0.92	0.86	0.79
	7800	**0.98**	**0.98**	**0.96**	0.91	0.86	0.81

normalized scalar product to the Pearson correlation coefficient here, because we do not want the matrix elements $c_{X_i X_j}$ and $\rho_q^{(X_i X_j)}$ to be altered by the normalization.

The results are listed in Table III. They show that the two methods coincide most for $1 \leqslant q \leqslant 2$ (or sometimes up to $q = 3$) and for $s \geqslant \Delta t$ (but not for $s \gg \Delta t$), exactly as one would expect on formal grounds. In this case qMSTs offer similar information as the standard MST. In contrast, if s is of the order of an hour or less, at $q \geqslant 4$ the presently introduced method provides us with information that cannot be found with the standard approach. This means that in the latter case, by using detrended MSTs of order q, we obtain significantly more extensive information about the correlation structure of the data under study. This constitutes a strong quantitative argument in favor of the new method.

IV. CONCLUSIONS

In this work we have introduced a family of q-dependent minimum spanning trees that are able to visualize the cross-correlations that are restricted to a specific range of fluctuation amplitudes. These trees are a generalization of the detrended MSTs proposed in [29]. They are defined on the basis of the q-dependent detrended cross-correlation coefficient ρ_q by transforming it to a metric distance. We applied the technique of qMST graphs to visualize the correlation structure of the American stock market and found that by changing the fluctuation amplitudes and time scales focused on, one observes substantial changes in the topology of the graphs. We identified a few previously known features of the financial data like the Epps effect, the division of the market into industrial sectors, the presence of the so-called market factor, and the privileged role of some specific stocks (like General Electric and Cisco) in the market structure. We showed that the trees for small values of q (e.g., $q \leqslant 3$) exhibit clusters that are strongly related to the industrial sectors of the market, especially for longer temporal scales s, while for larger values of this parameter ($q \geqslant 4$) this industrial structure becomes less and less evident. This thus indicates that what binds these industrial clusters are the typical fluctuations of moderate amplitude, while the largest fluctuations are more likely to synchronize among stocks that do not necessarily belong to the same industries.

The detrended qMSTs can complement the coefficient ρ_q in disentangling "hidden" cross-correlations, which cannot be observed either in analyses based on the measures defined for $q = 2.0$, like $\rho_{\rm DCCA}$ and the related detrended MSTs, or in analyses using the standard minimal spanning trees. In principle, the role of the MST is to decrease significantly the number of edges in a network by focusing on the most important ones. By considering a family of such trees with varying q's, the number of edges remains the same in each tree but the routes of connections among nodes may change with different choices of q. Changes indicate heterogeneity of cross-correlations, like, for instance, their varying strength for different fluctuation amplitudes in the time series, while independence of q provides evidence of homogeneity of cross-correlations. Our generalization of the MST concept to qMST indicates the direction of studying and quantifying such effects.

In order to illustrate how advantageous the qMST approach is, we applied it to two data sets with different, known statistical properties and documented that the trees representing $q = 2$ performed poorly in distinguishing between these sets, unlike the generalized trees with $q \neq 2$, whose performance was satisfactory. We thus showed that qMSTs, if constructed for different choices of q, are capable, as expected, of detecting diversity of correlations. The topological variability of the trees for different q's is related to how diverse the correlations are. Therefore, we envisage that both ρ_q and qMSTs will prove helpful in portfolio analysis and, of course, also in many other areas where cross-correlations are to be extracted from nonstationary time series.

[1] J. D. Barrow, S. P. Bhavsar, D. H. Sonoda, Minimal spanning trees, filaments and galaxy clustering, Mon. Not. R. Astron. Soc. **216**, 17 (1985).
[2] C. Adami and A. Mazure, The use of minimal spanning tree to characterize the 2d cluster galaxy distribution, Astron. Astrophys. Suppl. Ser. **134**, 393 (1999).
[3] R. N. Mantegna, Hierarchical structure in financial markets, Eur. Phys. J. B **11**, 193 (1999).
[4] J. Živković, B. Tadić, N. Wick, and S. Thurner, Statistical indicators of collective behavior and functional clusters in gene networks of yeast, Eur. Phys. J. B **50**, 255 (2006).

[5] E. Hernandez-García, E. A. Herrada, A. F. Rozenfeld, C. J. Tessone, V. M. Eguíluz, C. M. Duarte, S. Arnaud-Haond, and E. Serrão, Evolutionary and ecological trees and networks, in *Nonequilibrium Statistical Mechanics and Nonlinear Physics. AIP Conference Proceedings, Vol. 913*, edited by O. Descalzi, O. A. Rosso, and H. A. Larrondo (AIP, Melville, NY, 2007), p. 78.

[6] N. Vandewalle, F. Brisbois, and X. Tordoir, Non-random topology of stock markets, Quant. Fin. **1**, 372 (2001).

[7] J.-P. Onnela, A. Chakraborti, K. Kaski, and J. Kertesz, Dynamic asset trees and Black Monday, Physica A **324**, 247 (2003).

[8] S. Micciche, G. Bonanno, F. Lillo, and R. N. Mantegna, Degree stability of a minimum spanning tree of price return and volatility, Physica A **324**, 66 (2003).

[9] G. Bonanno, G. Caldarelli, F. Lillo, C. Micciché, N. Vandewalle, and R. N. Mantegna, Networks of equities in financial markets, Eur. Phys. J. B **38**, 363 (2004).

[10] M. McDonald, O. Suleman, S. Williams, S. Howison, and N. F. Johnson, Detecting a currency's dominance or dependence using foreign exchange network trees, Phys. Rev. E **72**, 046106 (2005).

[11] T. Mizuno, H. Takayasu, and M. Takayasu, Correlation networks among currencies, Physica A **364**, 336 (2006).

[12] C. Eom, G. Oh, and S. Kim, Topological properties of a minimal spanning tree in the Korean and the American stock markets, J. Korean Phys. Soc. **51**, 1432 (2007).

[13] M. J. Naylor, L. C. Rose, and B. J. Moyle, Topology of foreign exchange markets using hierarchical structure methods, Physica A **382**, 199 (2007).

[14] R. Coelho, C. G. Gilmore, B. Lucey, P. Richmond, and S. Hutzler, The evolution of interdependence in world equity markets—Evidence from minimum spanning trees, Physica A **376**, 455 (2007).

[15] A. Z. Górski, S. Drożdż, and J. Kwapień, Scale free effects in world currency exchange network, Eur. Phys. J. B **66**, 91 (2008).

[16] A. Garas, P. Argyrakis, and S. Havlin, The structural role of weak and strong links in a financial market network, Eur. Phys. J. B **63**, 265 (2008).

[17] P. Sieczka and J. A. Hołyst, Correlations in commodity markets, Physica A **388**, 1621 (2009).

[18] J. Kwapień, S. Gworek, S. Drożdż, and A. Z. Górski, Analysis of a network structure of the foreign currency exchange market, J. Econ. Interact. Coord. **4**, 55 (2009).

[19] R. D. Smith, The spread of the credit crisis: View from a stock correlation network, J. Korean Phys. Soc. **54**, 2460 (2009).

[20] M. Eryigit and R. Eryigit, Network structure of cross-correlations among the world market indices, Physica A **388**, 3551 (2009).

[21] C. Eom, G. Oh, H. Jeong, and S. Kim, Topological properties of stock networks based on minimal spanning tree and random matrix theory in financial time series, Physica A **388**, 900 (2009).

[22] T. Aste, W. Shaw, and T. Di Matteo, Correlation structure and dynamics in volatile markets, New J. Phys. **12**, 085009 (2010).

[23] M. Keskin, B. Deviren, and Y. Kocakaplan, Topology of the correlation networks among major currencies using hierarchical structure methods, Physica A **390**, 719 (2011).

[24] Y. Zhang, G. H. T. Lee, J. C. Wong, J. L. Kok, M. Prusty, and S. A. Cheong, Will the US economy recover in 2010? A minimal spanning tree study, Physica A **390**, 2020 (2011).

[25] L. Sandoval Jr., Pruning a minimum spanning tree, Physica A **391**, 2678 (2012).

[26] J. Kwapień and S. Drożdż, Physical approach to complex systems, Phys. Rep. **515**, 115 (2012).

[27] Z. Zheng, K. Yamasaki, J. N. Tenenbaum, and H. E. Stanley, Carbon-dioxide emissions trading and hierarchical structure in worldwide finance and commodities markets, Phys. Rev. E **87**, 012814 (2013).

[28] M. Wiliński, A. Sienkiewicz, T. Gubiec, R. Kutner, and Z. R. Struzik, Structural and topological phase transitions on the German Stock Exchange, Physica A **392**, 5963 (2013).

[29] G.-J. Wang, C. Xie, Y.-J. Chen, and S. Chen, Statistical properties of the foreign exchange network at different time scales: Evidence from detrended cross-correlation coefficient and minimum spanning tree, Entropy **15**, 1643 (2013).

[30] E. Kantar, B. Deviren, and M. Keskin, Hierarchical structure of the European countries based on debts as a percentage of GDP during the 2000–2011 period, Physica A **414**, 95 (2014).

[31] P. Skowron, M. Karpiarz, A. Fronczak, and P. Fronczak, Spanning trees of the world trade web: Real-world data and the gravity model of trade, Acta Phys. Pol. A **127**, A-123 (2015).

[32] J. Kruskal, On the shortest spanning subtree of a graph and the traveling salesman problem, Proc. Am. Math. Soc. **7**, 48 (1956).

[33] R. C. Prim, Shortest connection networks and some generalizations, Bell Syst. Tech. J. **36**, 1389 (1957).

[34] S. Drożdż, F. Grümmer, F. Ruf, and J. Speth, Towards identifying the world stock market cross-correlations: DAX versus Dow Jones, Physica A **294**, 226 (2001).

[35] C.-K. Peng, S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley, and A. L. Goldberger, Mosaic organization of DNA nucleotides, Phys. Rev. E **49**, 1685 (1994).

[36] B. Podobnik and H. E. Stanley, Detrended Cross-Correlation Analysis: A New Method for Analyzing Two Nonstationary Time Series, Phys. Rev. Lett. **100**, 084102 (2008).

[37] G. F. Zebende, DCCA cross-correlation coefficient: Quantifying level of cross-correlation, Physica A **390**, 614 (2011).

[38] R. T. Vassoler and G. F. Zebende, DCCA cross-correlation coefficient apply in time series of air temperature and air relative humidity, Physica A **391**, 2438 (2012).

[39] G. F. Zebende, M. F. da Silva, and A. Machado Filho, DCCA cross-correlation coefficient differentiation: Theoretical and practical approaches, Physica A **392**, 1756 (2013).

[40] J. C. Reboredo, M. A. Rivera-Castro, and G. F. Zebende, Oil and US dollar exchange rate dependence: A detrended cross-correlation approach, Energy Econ. **42**, 132 (2014).

[41] M. F. da Silva, E. J. A. L. Pereira, A. M. da Silva Filho, A. P. N. de Castro, J. G. V. Miranda, and G. F. Zebende, Quantifying cross-correlation between Ibovespa and Brazilian blue-chips: The DCCA approach, Physica A **424**, 124 (2015).

[42] M. Hussain, G. F. Zebende, U. Bashir, and D. Donghong, Oil price and exchange rate co-movements in Asian countries: Detrended cross-correlation approach, Physica A **465**, 338 (2017).

[43] J. Kwapień, P. Oświęcimka, and S. Drożdż, Detrended fluctuation analysis made flexible to detect range of cross-correlated fluctuations, Phys. Rev. E **92**, 052815 (2015).

[44] J. W. Kantelhardt, S. A. Zschiegner, A. Bunde, S. Havlin, E. Koscielny-Bunde, and H. E. Stanley, Multifractal detrended fluctuation analysis of nonstationary time series, Physica A **316**, 87 (2002).

[45] P. Oświęcimka, S. Drożdż, M. Forczek, S. Jadach, and J. Kwapień, Detrended cross-correlation analysis consistently extended to multifractality, Phys. Rev. E **89**, 023305 (2014).

[46] J. R. M. Hosking, Fractional differencing, Biometrika **68**, 165 (1981).

[47] Stock tickers (1999): AA, ABT, AHP, AIG, ALD, AMGN, AOL, ARC, AT, AUD, AXP, BA, BAC, BEL, BK, BLS, BMY, BUD, C, CA, CBS, CCL, CCU, CHV, CL, CMB, CMCS, COX, CPQ, CSCO, DD, DELL, DH, DIS, DOW, EDS, EK, EMC, EMR, ENE, F, FBF, FNM, FRE, FTU, G, GCI, GE, GLW, GM, GTE, GTW, HD, HWP, IBM, INTC, JNJ, JPM, KMB, KO, LLY, LOW, LU, MCD, MDT, MER, MMC, MMM, MO, MOT, MRK, MSFT, MTC, MWD, ONE, ORCL, PEP, PFE, PG, PNU, QCOM, QWST, SBC, SCH, SGP, SLB, SUNW, T, TWX, TX, TXN, UMG, USW, UTX, VIAB, WCOM, WFC, WMT, XON, YHOO (TAQ database; http://www.taq.com).

[48] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevE.95.052313 for Figs. 1–7 with node captions.

[49] V. Plerou, P. Gopikrishnan, L. A. N. Amaral, M. Meyer, and H. E. Stanley, Scaling of the distribution of price fluctuations of individual companies, Phys. Rev. E **60**, 6519 (1999).

[50] S. Drożdż, J. Kwapień, F. Grümmer, F. Ruf, and J. Speth, Are the contemporary financial fluctuations sooner converging to normal? Acta Phys. Pol. B **34**, 4293 (2003).

[51] S. Drożdż, M. Forczek, J. Kwapień, P. Oświęcimka, and R. Rak, Stock market return distributions: from past to present, Physica A **383**, 59 (2007).

[52] R. Rak, S. Drożdż, J. Kwapień, and P. Oświęcimka, Stock returns versus trading volume: Is the correspondence more general? Acta Phys. Pol. B **44**, 2035 (2013).

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

[54] J. Kwapień, S. Drożdż, F. Grümmer, F. Ruf, and J. Septh, Decomposing the stock market intraday dynamics, Physica A **309**, 171 (2002).

[55] G.-J. Wang, C. Xie, S. Chen, J.-J. Yang, and M.-Y. Yang, Random matrix theory analysis of cross-correlations in the US stock market: Evidence from Pearson's correlation coefficient and detrended cross-correlation coefficient, Physica A **392**, 3715 (2013).

[56] T. W. Epps, Comovements in stock prices in the very short run, J. Am. Stat. Assoc. **74**, 291 (1979).

[57] J. Kwapień, S. Drożdż, and J. Speth, Time scales involved in emergent market coherence, Physica A **337**, 231 (2004).

[58] B. Tóth and J. Kertész, Increasing market efficiency: Evolution of cross-correlations of stock returns, Physica A **360**, 505 (2005).

[59] J. Theiler, S. Eubank, A. Longtin, B. Galdrikian, and J. D. Farmer, Testing for non-linearity in time series: The method of surrogate data, Physica D **58**, 77 (1992).

[60] H. G. Schuster and W. Just, *Deterministic Chaos: An Introduction*, 4th ed. (Wiley, New York, 2006).