Correlation structure and dynamics in volatile markets

New Journal of Physics
The open-access journal for physics

Correlation structure and dynamics in volatile markets

T Aste[1,2,3,4], **W Shaw**[2] **and T Di Matteo**[2,3]
[1] School of Physical Sciences, University of Kent, Canterbury,
Kent CT2 7NH, UK
[2] Department of Mathematics, King's College London, The Strand, London
WC2R 2LS, UK
[3] Department of Applied Mathematics, School of Physical Sciences,
The Australian National University, Canberra, ACT 0200, Australia
E-mail: t.aste@kent.ac.uk

New Journal of Physics **12** (2010) 085009 (21pp)
Received 9 February 2010
Published 19 August 2010
Online at http://www.njp.org/
doi:10.1088/1367-2630/12/8/085009

Abstract. The statistical signatures of the 'credit crunch' financial crisis that unfolded between 2008 and 2009 are investigated by combining tools from statistical physics and network theory. We devise measures for the collective behavior of stock prices based on the construction of topologically constrained graphs from cross-correlation matrices. We test the stability, statistical significance and economic meaningfulness of these graphs. The results show an intriguing trend that highlights a consistently decreasing centrality of the financial sector over the last 10 years.

[4] Author to whom any correspondence should be addressed.

Contents

1. Introduction

Financial systems are archetypal of complexity. In markets, several individuals, groups, humans and machines operate at different frequencies with different strategies interacting at different levels, both individually and collectively, within an intricate network of complicated relations. The emerging dynamics continuously evolve through bubbles and crashes with erratic trends. Although intrinsically ‘complex’, financial systems are very well suited to statistical studies. Indeed, the framework of governing rules is rather stable and the time evolution of the system is continuously monitored, providing a very large amount of data for scientists to investigate [1].

One typical concern for investors is to diversify risk by investing in assets that are differently affected by external or internal events [2]. This, in principle, requires the study and understanding of the collective dynamics of all the assets in a portfolio and the identification of sets of stocks that behave similarly or dissimilarly, or independently.

In this paper, we look at the collective dynamics of 395 stocks traded in the US equity market in the time period between 1 January 1996 and 30 April 2009 (data from Reuters [3]). We look at the relative variations in their prices and we study their interdependence by building from the cross-correlation matrix a network of significant links among the stocks. We study the time evolution of such networks and highlight some important changes that occurred during the unfolding of the 2008–2009 crisis.

The paper is organized as follows. In section 2, we introduce the cross-correlations among stocks as a measure of similarity and dependency. We discuss the effect of the finite size of the time series and we introduce a measure of significance against the null hypothesis. The significance of the correlation is also discussed by measuring the spectrum of the eigenvalues and comparing it with the spectrum from randomized data. In section 3, we introduce a method to extract only a subset of meaningful correlations from the whole cross-correlation matrix by means of a network approach. In section 4, we discuss the economic and financial interpretation of such networks of relevant links between stocks and we look at the dynamical effects associated with the recent 2008–2009 financial crisis. In section 5, we look at the position within these networks of a set of stocks belonging to the financial sector by monitoring their relative centrality and peripherality across a time period of 13 years. Section 6 summarizes the main results.

2. Correlations and dependency

In markets, each stock price is not evolving in isolation. Indeed, individual price changes affect the global variation of the market trend as much as the collective dynamics are reflected in the individual behavior. Therefore, the understanding of the properties of such a system of interactions, dependencies and co-variations is crucial to the understanding of the overall system.

A commonly used measure of dependency between two data series $x_i(t)$ and $x_j(t)$ (with $t = 1, 2, \ldots, T$) is the correlation coefficient. The Pearson's estimator of the correlation coefficient between two data series is defined as

$$\rho_{i,j} = \frac{\frac{1}{T}\sum_{t=1}^{T}(x_i(t)-\mu_i)(x_j(t)-\mu_j)}{\sigma_i\sigma_j}, \tag{1}$$

where $\mu_{i(j)}$ and $\sigma_{i(j)}$ are, respectively, the sample mean and the sample standard deviation of the data series $x_{i(j)}(t)$. In this paper, we calculate the correlation coefficients for the log-returns of the daily prices that, for a given stock 'i', are given by [4, 5]

$$x_i(t) = \log(\text{price}_i(t+1)) - \log(\text{price}_i(t)). \tag{2}$$

The Pearson's correlation coefficient between these time series of log-returns of the stock prices is a measure of the dependency between the dynamical evolution of the prices of two stocks. Such a measure is common and widespread in the literature and it turns out to be rather efficient at catching similarities between the evolution of stock prices. However, one should be aware that this measure can sometimes be problematic, especially for nonlinear dependencies or for data series characterized by large fluctuations with power-law tails.

Here, we are interested in the variation in the correlation structure with time and therefore we look at correlations computed over a moving window of size Δ. This means that equation (1) is applied over a subset of the time series within the time window $[t-\Delta+1, t]$. A practical example is given in figure 1, where cross-correlations between the log-returns of *Ford Motor* (FORD) and *General Electric* (GE) are calculated over moving windows of $\Delta = 62$, 125 and 250 working days (corresponding roughly to 3 months, 6 months and 1 year, respectively). From this figure, it is clear that the correlations computed over smaller windows have larger fluctuations, which are, however, around the corresponding values computed over larger windows.

By looking at the whole cross-correlation matrix for all the $n = 395$ stocks calculated over the whole time period, we observe that the vast majority of correlation coefficients are positive and that they are distributed around an average correlation of $\langle\rho\rangle = \frac{2}{n(n-1)}\sum_{i<j}\rho_{i,j} \sim 0.25$ with only a few coefficients above 0.5. This is shown in figure 2(a), where the histogram of the correlation coefficients is reported. As one can see from figures 2(b)–(d), broader distributions are measured when smaller sub-periods are analyzed. Interestingly, the average correlation

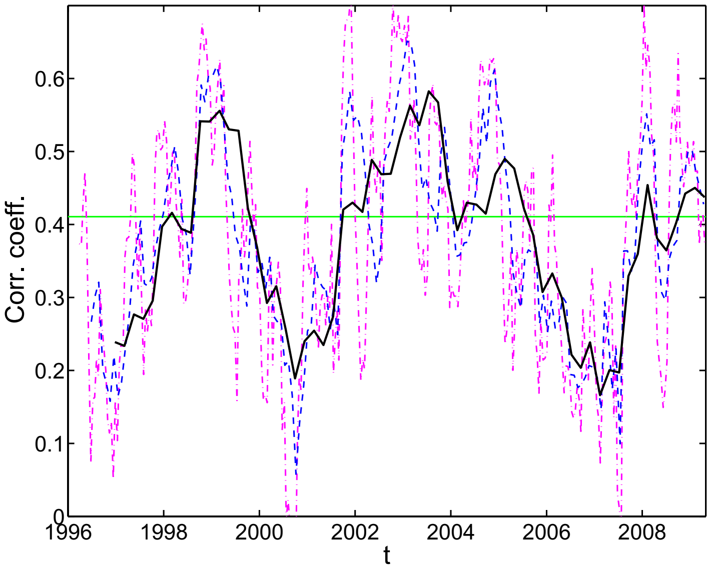

Figure 1. Correlation coefficients ($\rho_{i,j}$) as a function of time (t) between the log-returns of the two time series of *Ford Motor* and *General Electric* over moving windows of, respectively, $\Delta \sim 62$ (dash-dot magenta line), $\Delta \sim 125$ (dashed blue line) and $\Delta \sim 250$ (solid black line) working days. The horizontal line is the correlation coefficient between the log-returns time series of the two companies calculated over the whole period.

varies significantly by changing the observation period, as clearly visible by comparing figures 2(b)–(d). This can be better seen from figure 3, where the average correlation over a moving window of 250 working days ($\sim$1 year) is reported. One can note that, indeed, in the year 2000 the correlations had average values below 0.1, whereas in 2009 the average values were above 0.25. Indeed, periods of larger market instabilities trigger collective reactions that result in larger correlations. In figure 3, one can clearly see larger average correlations following the 9/11 turmoil and then, after an easing in 2004–2007, a substantial and fast increase in the average correlation during the credit crunch crisis (2008–2009). Interestingly, such an increase preceded the crisis and it seems to be a signature of the growing price bubble during the bull market in 2007–2008.

2.1. Significance of the correlation coefficients against the null hypothesis

A first important question to answer is whether these coefficients are significant or are instead just the product of some random coincidences. A measure of significance can be inferred by comparing the values of the correlation coefficients computed from the measured time series with the coefficients obtained from the same time series but with the time entries randomly shuffled, eliminating in this way any real temporal correlation. Random, uncorrelated data should result in zero correlation coefficients. However, for data series of finite window sizes Δ, there will be some residual finite non-zero fluctuations in the coefficients. We have produced several cross-correlation matrices from different randomly shuffled series. The residual correlation coefficients between the shuffled series provide a threshold value for the

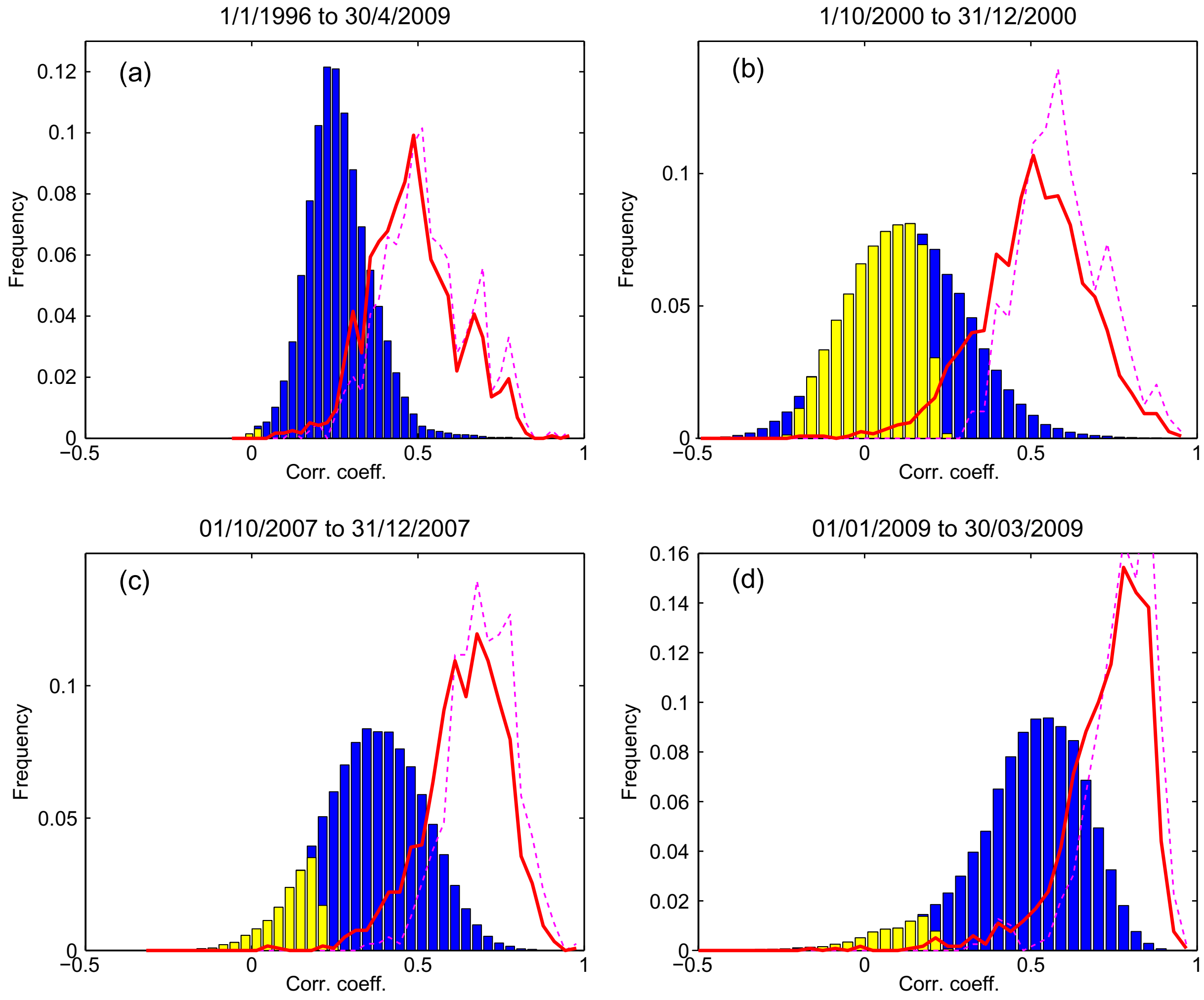

Figure 2. (**a, blue bars**) Frequency distributions of the cross-correlation coefficients calculated over the whole period from 1/1/1996 to 30/4/2009 (3353 working days in total) for the 395 companies analyzed. (**b, c, d, blue bars**) Frequency distributions of the cross-correlation coefficients calculated for the same set of 395 companies over shorter time periods of 3 months: (**b**) from 1/10/2000 to 31/12/2000 (63 working days); (**c**) from 1/10/2007 to 31/12/2007 (64 working days); (**d**) from 1/01/2009 to 30/03/2009 (61 working days). (**Yellow bars**) In all figures the superimposed bars in 'yellow' are the frequencies of the non-significant correlations (which score less than 90% significance against the null hypothesis, see text). (**Red lines**) Frequency distributions of the subset of correlation coefficients contained in the connected links of the PMFG. (**Magenta dashed lines**) Frequency distributions of the subset of correlation coefficients contained in the connected links of the MST.

'null hypothesis'. We assess a 90% confidence factor against the null hypothesis by calculating for 100 times the correlation coefficients $\rho_{i,j}^{\text{shaff}}$ from differently shuffled data series and then considering significant the correlation coefficients $\rho_{i,j}$ from the real series that, at least 90 times out of 100, are larger (in absolute value) than the corresponding correlation coefficients from the

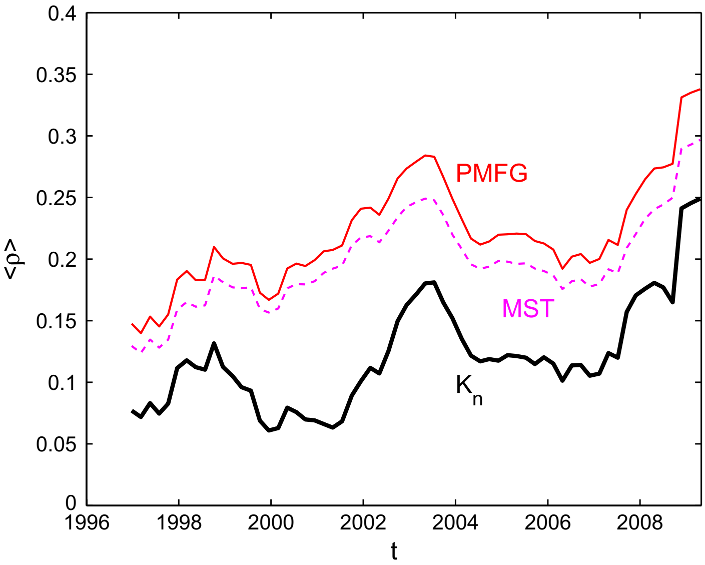

Figure 3. Average correlation coefficient calculated, respectively, for (from bottom to top) the whole cross-correlation matrix (complete graph K_n, black line), the MST (magenta line) and the PMFG (red line) calculated over a moving window of 250 working days.

shuffled series (i.e. $|\rho_{i,j}| > |\rho_{i,j}^{\mathrm{shaff}}|$). In figure 2, the histograms of the non-significant correlation coefficients are superimposed on the histograms of all correlation coefficients. We can first notice that longer time series produce a larger portion of significant coefficients. For instance, when the correlation coefficients are calculated over the whole time period (therefore computed from time series containing over 3300 data points), they contain only 0.5% non-significant coefficients (see figure 2). On the other hand, the correlation coefficients calculated over 3 month periods (computed therefore from time series containing only about 60 data points) contain much larger portions of non-significant coefficients with the cases reported in figures 2(b)–(d) having, respectively, 65%, 16% and 7% non-significant coefficients. We can note that there are relevant changes over different observation periods with a much larger portion of significant correlations observed in the first 3 months of the year 2009 (figure 2(d)) with respect to the last 3 months of the year 2000 (figure 2(b)). It should also be noted that most of these changes are due to the increase in the average correlation, that is over three times larger in 2009 than in 2000 (see figure 3). Indeed, signals with weaker correlations are more affected by noise and finite window size than strongly correlated signals. The fact that in some cases over 60% of the correlation coefficients are not significant might appear alarming. However, we must consider that we are going to filter this correlation, extracting only a subset of meaningful links. We will see shortly that such a subset of filtered correlations is almost entirely made up of significant coefficients.

2.2. *Eigenvalue distribution*

We have discussed in the previous section the significance of the cross-correlation between each couple of variables. Another measure of significance concerns the degree of independence between linear combinations of variables with respect to the correlation structure resulting from

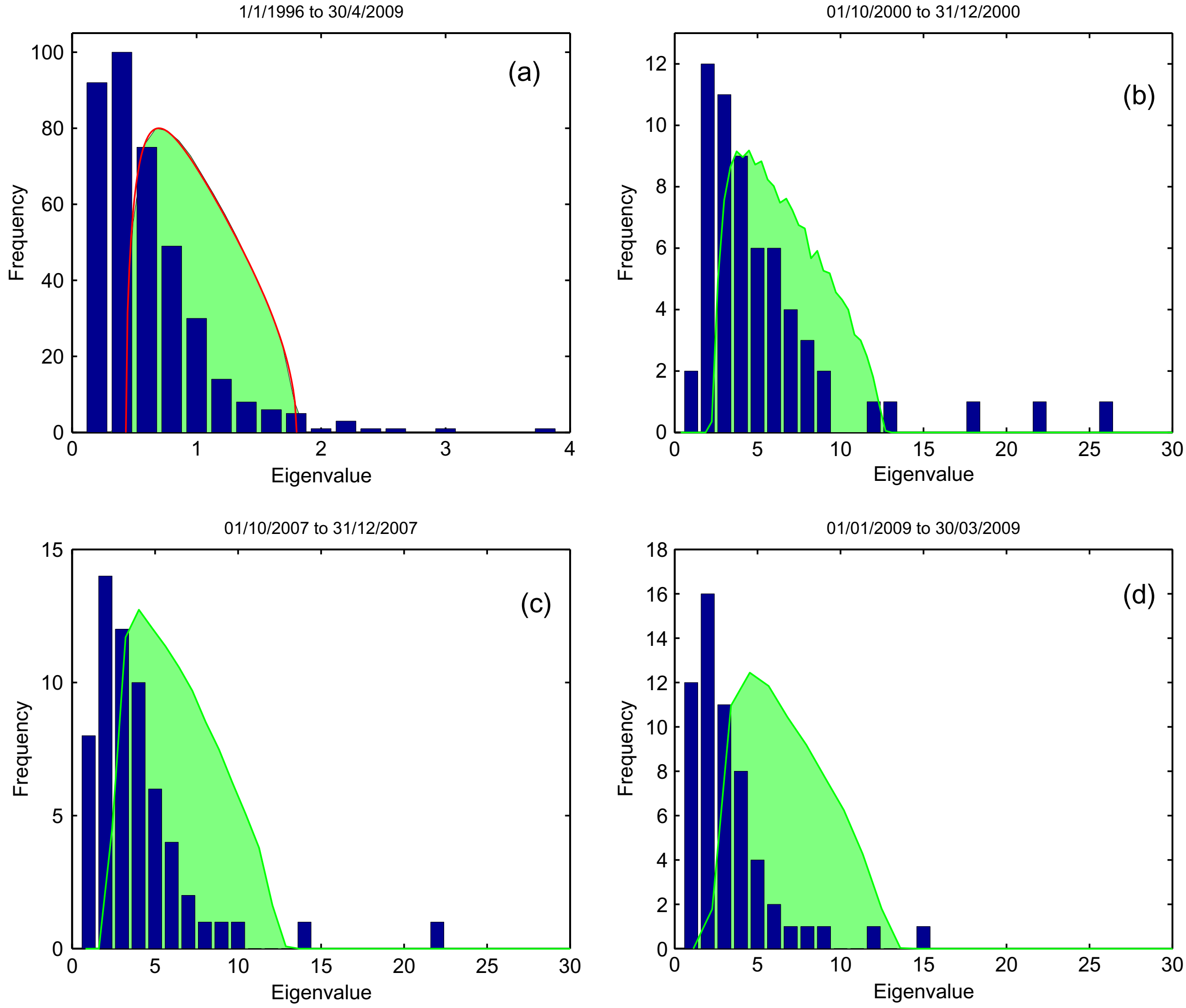

Figure 4. (a) Frequency distribution of the eigenvalues of the correlation coefficient matrix calculated over the whole time period. There are other six larger eigenvalues (at: 5.6, 6.3, 9.3, 10.8, 14.1 and 109.7), which are not reported. The red line is the theoretical prediction for such a distribution. (b) The same distribution over the sub-period from 1/10/2000 to 31/12/2000. There are also 332 zero eigenvalues and one larger eigenvalue at 69.5, which are not reported. (c) The same distribution over the sub-period from 1/10/2007 to 31/12/2007. There are also 331 zero eigenvalues and one larger eigenvalue at 152.1, which are not reported. (d) The same distribution over the sub-period from 1/01/2009 to 30/03/2009. There are also 334 zero eigenvalues and one larger eigenvalue at 204.1, which are not reported. In all these figures, the green area plots are the distributions of the eigenvalues for the cross-correlation matrices generated from randomly shuffled time series.

noisy uncorrelated data. This can be done by comparing the eigenvalue spectrum from the real cross-correlation matrices with the ones from the correlation matrices from the shuffled data series. In figure 4(a), we report (blue bars) the spectrum of eigenvalues from the correlation matrix between the $n = 395$ firms quoted on the NYSE calculated over the whole period from 01/01/1996 to 30/04/2009 ($T = 3358$). In this figure, the green area plot is the (average) spectrum from the shuffled data series and the red line is the theoretical behavior of the eigenvalue spectrum for n random uncorrelated data series calculated over a window of finite length $T > n$, which is known analytically [6],

$$p(\lambda) = \frac{Q}{2\pi}\frac{\sqrt{(\lambda_{\max} - \lambda)(\lambda - \lambda_{\min})}}{\lambda},$$

where $Q = T/n$ and $\lambda_{\min} = 1 + 1/Q - 2\sqrt{1/Q}$ and $\lambda_{\max} = 1 + 1/Q + 2\sqrt{1/Q}$ are, respectively, the maximum and minimum eigenvalues. As one can see, the analytical expression matches almost perfectly the distribution from correlations of shuffled data series and it is instead markedly different from the spectra of the correlations from the real-time series. We have measured that the spectrum of eigenvalues counts 16 eigenvalues with values above $\lambda_{\max}$. This is an indication that, in the cross-correlation structure, there are at least $16 \times n$ significant and independent variables.

In figures 4(b)–(d), we also report the spectra of correlations computed over shorter periods of time of 3 months: (b) 1/10/2000–31/12/2000; (c) 1/10/2000–31/12/2007 and (d) 1/1/2009–31/3/2009. In this case, a large number of eigenvalues are zero because the matrix rank cannot be larger than T, which in this case is $\sim$63, leaving therefore $n - T \sim 332$ eigenvalues equal to zero. Figures 4(b)–(d) report only the nonzero eigenvalues for both the correlation matrix from the real-time series (blue bars) and the shuffled ones (green area plot). In this case, the analytical prediction cannot be applied because $T < n$. For these shorter time series, we can see that the spectrum of the eigenvalues from shuffled data reaches larger values around $\lambda_{\max} \sim 13$, and we also have a smaller number of eigenvalues above $\lambda_{\max}$ (namely 5, 3 and 2 for the time periods (b), (c) and (d), respectively). Let us note that this is an indication that an information filtering procedure that aims to keep both significant and non-redundant correlations should not exceed $3 \times n$ filtered entries. This is discussed in the next section.

Interestingly, we see again also for these shorter time series that there are very significant differences between the spectra from real data and the spectra from the shuffled series. This is somehow different from what is reported in [7, 8] where, on the contrary, the authors found an overall similarity between the eigenvalue spectrum from real data and the prediction for random data series. Given that the authors of [7, 8] have studied a rather similar set of data, but over the period 1991–1996, we might conclude that something has changed in the market's correlation structure since 1996.

3. Complex networked systems

One of the most significant recent advances in complex systems studies has been the use of networks to represent and model the complex structure of interactions between the system's elements [9]–[13]. The general idea is that, in a complex system consisting of many interacting elements, one can associate a node with each element and an edge with each interaction/relation. In the case of financial markets, in principle, each stock price is related to the price of all other stocks, and we have already seen that a measure of such interrelation is the cross-correlation matrix. Straightforwardly, one could associate with such a system the complete graph K_n (where every node is connected to all other nodes) and eventually assign a weight to the edges between stocks with the corresponding correlation coefficient $\rho_{i,j}$ (or other related measures such as the distance $d_{i,j} = \sqrt{2(1-\rho_{ij})}$). For a system of n stocks, this results in a graph of $n(n-1)/2$ edges, which are indeed the number of different entries in the cross-correlation matrix, which is a symmetric $n \times n$ matrix with all entries on the diagonal. However, we have seen previously that the correlation matrix contains a portion of non-significant entries and a large amount of redundant information that we want to filter out in order to extract only the subset of interactions that is important and makes the significant 'backbone network' of the system [14]–[18]. This is a nontrivial task that unavoidably involves some degree of arbitrariness. Here, we discuss a recently proposed method [19, 20] based on the use of topological constraints on filtered graphs.

3.1. Disentangling the network: minimum spanning tree

We want to build a connected graph whose topological structure represents the correlation among the different elements but that is greatly reduced in the number of edges with respect to the complete graph. In such a network, all the important relations must be represented, but the network should be kept as 'simple' as possible.

The simplest connected graph is a spanning tree (a graph with no cycles that connects all vertices). It is therefore natural to choose as the representative network a spanning tree that retains the maximum possible number of correlations [14]. Such a graph is called a minimum spanning tree (MST) [21]–[23].[5] Let us here briefly describe a very simple intuitive (non-optimal) algorithm to construct the MST. This helps us to clarify the concept and to introduce a methodology that will be then applied to other graph filtering methods.

A general approach to the construction of the MST is to connect the most correlated pairs while constraining the graph to be a tree, as follows:

Step 1: Make an ordered list of edges i, j, ranking them by decreasing correlation $\rho_{i,j}$ (first the largest and last the smallest).

Step 2: Take the first element in the list and add the edge to the graph.

Step 3: Take the next element and add the edge if the resulting graph is still a forest or a tree; otherwise discard it.

Step 4: Iterate the process from step 3 until all pairs have been exhausted.

The resulting MST has $n-1$ edges and it is the spanning tree that maximizes the sum of the correlations over the connected edges.

The resulting MST for the present database of 395 stocks is shown in figure 5(a), where the following stocks have been highlighted: FORD, *General Electric* (GE), *Lehman Brothers Holdings* (LEHBR) and *American International Group* (AIG).

Let us here briefly show that this network is meaningful in providing a structure of interconnections that reflects the market structure and interaction between firms. We can, for instance, observe from figure 5(a) that in such a network the stock *Ford Motor* (FORD) is at the periphery of the network lying at the end of a branch. By following the branch towards the main trunk (see figure 5(b)), we find directly attached to *Ford Motor* the other large USA car manufacturer, *General Motors* (GM). Next, going down the branch, is *Johnson Controls* (JCI),

[5] The use of 'minimum' instead of 'maximum' is historical and is due to the fact that, in the original approach, the edge weight that was *minimized* was a distance such as $d_{i,j} = \sqrt{2(1-\rho_{ij})}$. It is straightforward to see that this same graph also maximizes the alternative weight given by the correlations $\rho_{i,j}$.

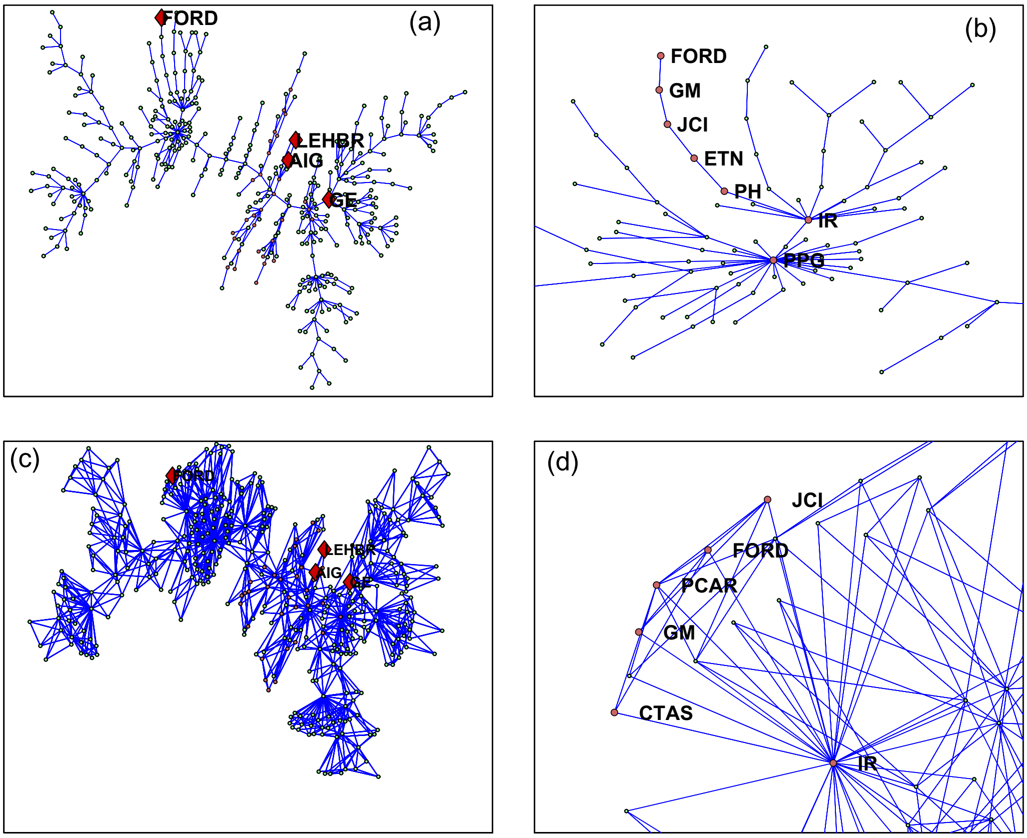

Figure 5. (a) MST built from the cross-correlation matrix of 395 stocks traded in the US equity market over the period from 1/1/1996 to 30/4/2009. (b) Detail of the MST branch ending with FORD. (c) PMFG built for the same cross-correlation data. (d) Detail of the direct connections with FORD.

which is a producer of interior systems for light vehicles, including passenger cars and light trucks. One step forward down the branch is *Eaton* (ETN), which is a large industrial corporation that includes an important vehicle group. Further down in the line we find *Parker Hannifin* (PH), a global leader in the manufacture of motion and control technologies. The following node corresponds to *Ingersoll-Rand* (IR), which, as one can see, is attached to several other nodes (12 in total). Indeed, this company is an international supplier to transportation, manufacturing, construction and agricultural industries. Going forward, we reach the main trunk, where we find the 'hub' *PPG Industries* (PPG), which is attached to 27 other companies. PPG is, indeed, a very influential market player that is a global supplier of paints, coatings, optical products, specialty materials, chemicals, glass and fiber glass with a large amount of activity in automotive coating and refinishing products. It should be quite clear, even to non-experts, that this structure of links that we have extracted with the MST is very meaningful in gathering together some of the main players in the USA automotive industry. What is remarkable is that these links have been extracted solely from the cross-correlation matrix without any other *a priori* information about the system.

Let us stress that the same method can potentially be applied to a very broad class of systems, specifically in all cases where a correlation (or even, more simply, a similarity measure) between a large number of interacting elements can be assigned.

3.2. Disentangling the network: planar maximally filtered graph (PMFG)

Although we have just shown that the MST method is powerful and meaningful, there are some aspects that might be unsatisfactory. In particular, the condition that the extracted network should be a tree is a strong constraint. Indeed, let us, for instance, consider the case where three companies are involved in similar activities and therefore have strongly correlated behaviors in the dynamics of their stock prices. In the MST construction, unavoidably only two of these companies can be directly connected with an edge in the filtered graph because the connection with an extra edge of the third company will form a triangular cycle, a 3-clique, which is not allowed in a tree. Ideally, one would like to be able to maintain the same powerful filtering properties of the MST but also allow the presence of extra links, cycles and cliques in a controlled manner.

A recently proposed solution consists in building graphs embedded on surfaces with a given genus [19]. (Roughly speaking, the genus of a surface is the number of holes in the surface: $g = 0$ corresponds to the embedding on a topological sphere; $g = 1$ on a torus, $g = 2$ on a double torus, etc.) The algorithm to build such a graph is identical to the one for the MST discussed previously, except that at step 3 the condition to accept the link now requires that the resulting graph must be embeddable on a surface of genus g [19, 20, 24]. The resulting graph has $3n - 6 + 6g$ edges and is a triangulation of the surface.

It is known that, for a large enough genus, any graph can be embedded on a surface [19, 25]. From a general perspective, the larger the genus is, the greater the amount of original information that is preserved in the graph is, and the complexity of the embedded triangulation also increases proportionally. When the genus is increased by 1, a new handle connecting two different parts of the surface is added. In the embedded graph, this operation corresponds to the addition of new edges. This increases the complexity of the graph, creating new cycles and making it more interwoven. On the other hand, new added edges shortcut two different parts of the network, making the system more compact.

The simplest graph is the one associated with $g = 0$, which is a triangulation of a topological sphere. Such a planar graph is called a planar maximally filtered graph (PMFG) [20], [26]–[28]. PMFGs have the algorithmic advantage that planarity tests are relatively simple to perform. PMFGs can be viewed as the first incremental step towards complexity after the MST. It has been proved that the MST is always a subgraph of the PMFG [20]. A PMFG has 3 times the number of edges as the MST, cycles are admitted, the number of 3-cliques must be larger or equal to $2n - 4$ and 4-cliques can be present. Let us note that keeping $3n - 6$ entries from the original $n(n - 1)/2$ cross-correlation coefficients is consistent with the required reduction in redundancy that is highlighted from the study of the eigenvalue spectrum in the previous section.

The PMFG for the 395 stocks on the NYSE is reported in figure 5(c). By comparing the PMFG with the MST (reported in figure 5(a)), we can see that the PMFG is a graph richer in links and with a more complex structure than the MST which it preserves, and it expands some of the hierarchical properties. For instance, recalling the previous description of the connection structure in the MST around the node FORD, we can observe from figure 5(d) that now in the PMFG the company FORD has acquired new edges in addition to the one with GM connecting it directly with *Cintas* (CTAS), which is a company delivering specialized services to businesses. It also connects directly with IR, which was the 5th neighbor of FORD in the MST. It also connects with *Paccar* (PCAR), which is the third largest manufacturer of heavy-duty trucks in the world. Finally, it links directly with JCI, which was FORD's third neighbor in the MST branch.

These connections show that PMFG is improving the information content of the MST, gathering together companies in a meaningful way by clustering them according to their sector and their activities without the use of any *a priori* information besides the correlation coefficients measured from the log-returns. This filtering can be conveniently used in portfolio selection where, in order to reduce risk, the investment must be carefully diversified across sectors and activities.

4. Use of the network topology to study structural changes in the market

We have seen so far that graph filtering is a useful and powerful instrument to extract information about the market structure. Let us now investigate how robust such information is and how it changes with time. Our main objective is to implement techniques and methods to highlight structural transitions that might happen in markets before, during or after a crisis.

4.1. Network significance

Both MSTs and PMFGs extract a subset of links. We have discussed in the previous subsection that these links are meaningful in reproducing bonds between firms with similar or related economic activities. It must be pointed out that these are non-thresholding filtering methods that are not simply extracting the highest correlations but are instead extracting the correlation structure, keeping also some weak correlations that are, however, relevant in the local structure description. Indeed, correlation structures in complex systems often have cross-scale properties with, for instance, some firms forming highly correlated clusters and others instead interacting weakly with each other. Thresholding will keep only the strong interaction, disconnecting from the system the part that is less correlated than the threshold. A graph-filtering method can instead describe the entire structure by also keeping the weak correlations but simultaneously filtering out redundancies in the highly correlated part.

We have already shown in section 2 with figure 3 that the average correlation calculated over a moving window of 250 working days (∼1 year) changes significantly with time showing, in particular, a strong increase during the unfolding of the 2008–2009 crisis. In figure 3, the behavior of the average correlation coefficients in the whole cross-correlation matrix (the complete graph) is compared with the ones associated with edges in the MST and PMFG. We can see that the three averages follow very similar trends and variations, revealing that both the MST and the PMFG capture the main evolving features of the whole correlation structure.

In section 2, it has also been pointed out that a large portion of the correlation coefficients might not be significant when the correlations are estimated over a short time window size (such as 3 months). This has been shown in figure 2, where the histograms of the correlation coefficients are reported. For instance, we have seen that only 35% of the correlation coefficients calculated from the time window between 1/10/2000 and 31/12/2000 (figure 2(b)) are statistically significant. An efficient filtering method should automatically extract the significant coefficients only. Indeed, we have verified that, in the same time window, 100% of the edges of the MST are associated with significant coefficients, and ∼96% of the edges of the PMFG are associated with significant coefficients. For all the time windows analyzed, we have observed very similar proportions, with the MST always having 100% significant edges and the PMFG having between 94% and 100% significant edges. A small presence of edges for non-significant coefficients in the PMFG is a consequence of the largely increased number of coefficients retained by the PMFG with respect to the MST.

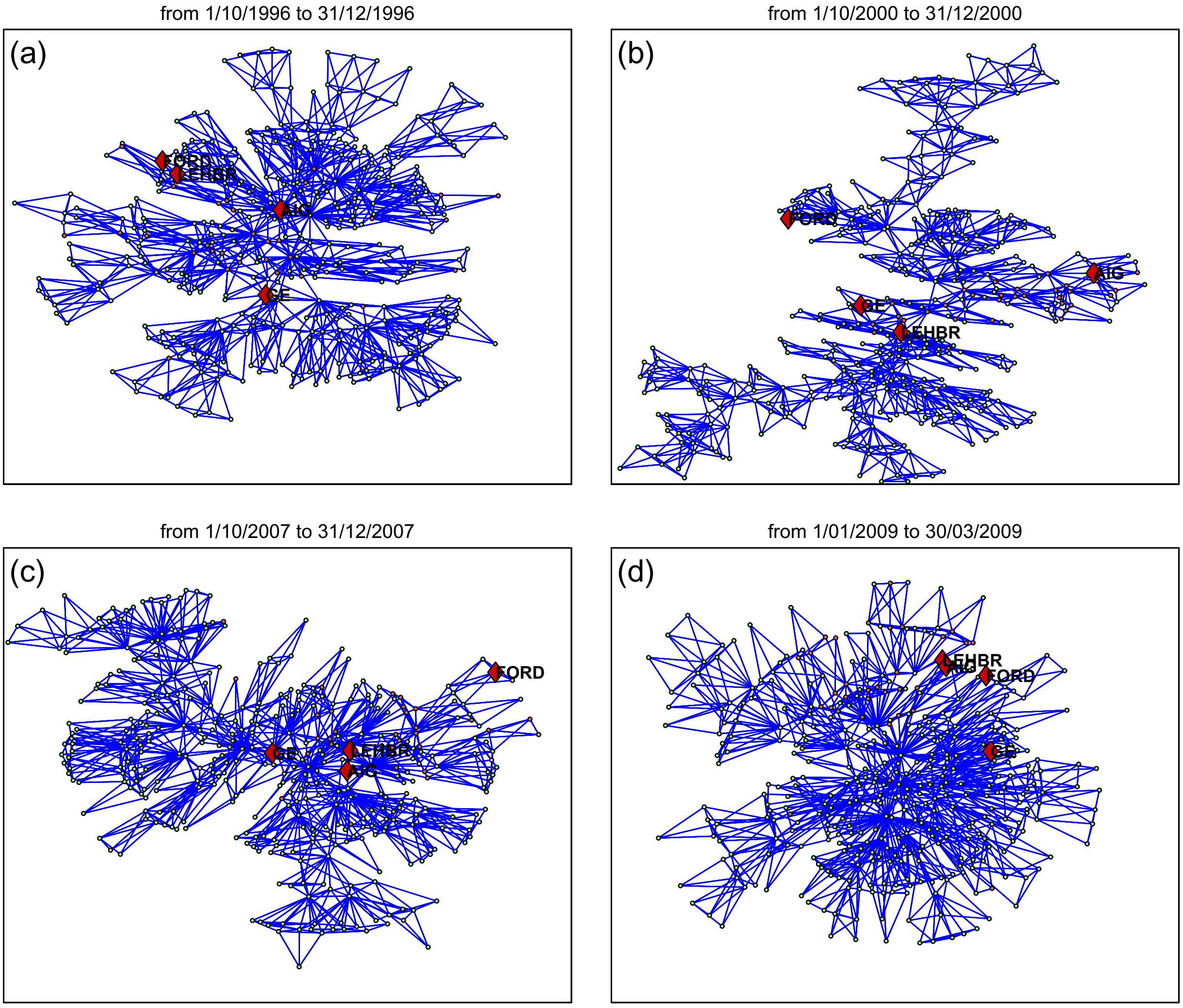
from 1/10/1996 to 31/12/1996
(a)
FORD
LEHBR
AIG
GE
from 1/10/2000 to 31/12/2000
(b)
FORD
AIG
GE
LEHBR
from 1/10/2007 to 31/12/2007
(c)
FORD
GE
LEHBR
AIG
from 1/01/2009 to 30/03/2009
(d)
LEHBR
AIG
FORD
GE

Figure 6. PMFGs built from correlations calculated from four different time periods of 3 months. (a) From 1/10/1996 to 31/12/1996; (b) from 1/10/2000 to 31/12/2000; (c) from 1/10/2007 to 31/12/2007 and (d) from 1/01/2009 to 30/03/2009.

4.2. Network dynamics and stability

When the PMFGs and MSTs are calculated from correlations over a moving time window, their topological structures can change dynamically with the window position, reflecting the changes occurring in the market structure during such a period of time. In figure 6, we report an example concerning four PMFGs calculated over four different time periods starting, respectively, at 01/10/1996 (a), 01/10/2000 (b), 01/10/2007 (c) and 01/01/2009 (d) with a time window size of 3 months. As one can observe, there are substantial changes between the different sub-periods which involve the whole graph structure. A qualitative view of these changes is provided by looking at the relative positions of four stocks, namely FORD, GE, LEHBR and AIG. In order to quantify these changes, we have followed the dynamics of the graph topology within different sub-periods. This has been done by considering a time window of 3 months

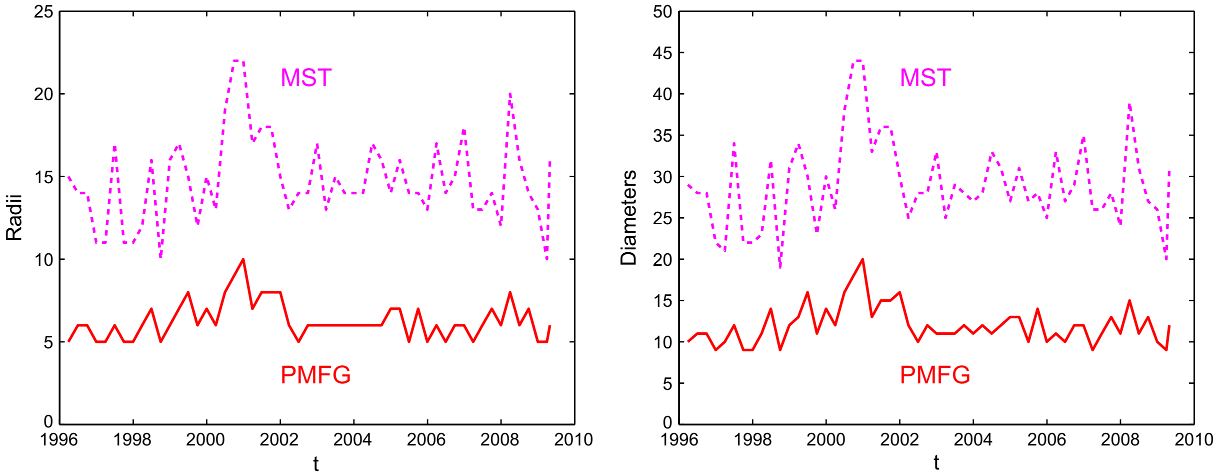

Figure 7. Radii (left) and diameters (right) for the MST (magenta) and PMFG (red) built from data over non-overlapping time windows of $\Delta = 3$ months.

and rolling it over the whole period by shifting it 3 months per time, dividing in this way the time period from 1/1/1996 to 30/4/2009 into 54 non-overlapping sub-periods of about 63 working days each. For each of these sub-periods, we have computed the MST and the PMFG. A quantitative measure of the changes occurring in the overall graph structures can be inferred by calculating the radius and diameter of the graphs. Let us recall that the *diameter* of a graph is the largest shortest path between any couple of nodes, whereas the *radius* of a graph is calculated by taking the smallest among the set of longest shortest paths between each node and all other nodes (typically the radius is about half the diameter). These are global measures that quantify the graph interconnections and its elongation. For instance, it should be rather clear from figures 6(a) and (b) that the PMFG in figure 6(a) must have a smaller diameter and radius than the more elongated one in figure 6(b). Figure 7 shows the evolution of diameters and radii for the graph computed from the 54 non-overlapping sub-periods of 3 months from 1/1/1996 to 30/4/2009. We can observe two interesting large peaks, respectively, around 2001 and 2008 that precede the unfolding of market instabilities.

4.3. T1 distance

One can think of these sequences of graphs (MST or PMFG) as evolving networks that change and adapt accordingly with the changes in the market structure. In this way, two successive graphs Γ_1 and Γ_2 get transformed into each other by a dynamical process where links in Γ_1 are re-wired according to local moves that eventually lead to Γ_2. It can be shown that there is a simple move known as the T1 move [29] that can be used to transform any PMFG into another. This move consists of switching edges by connecting two vertices that are second neighbors in the PMFG, and simultaneously disconnecting two adjacent first neighbors in order to maintain planarity [30]–[32]. A similar T1 move can also be implemented in the MST where two second neighbors vertices can be directly connected and one edge between these vertices can be removed in order to avoid the creation of cycles. We can therefore introduce a topological distance that is the minimum number of moves required for a dynamical adaptation from one graph to the following one. We call this number the 'T1 distance' [29, 32] between previously

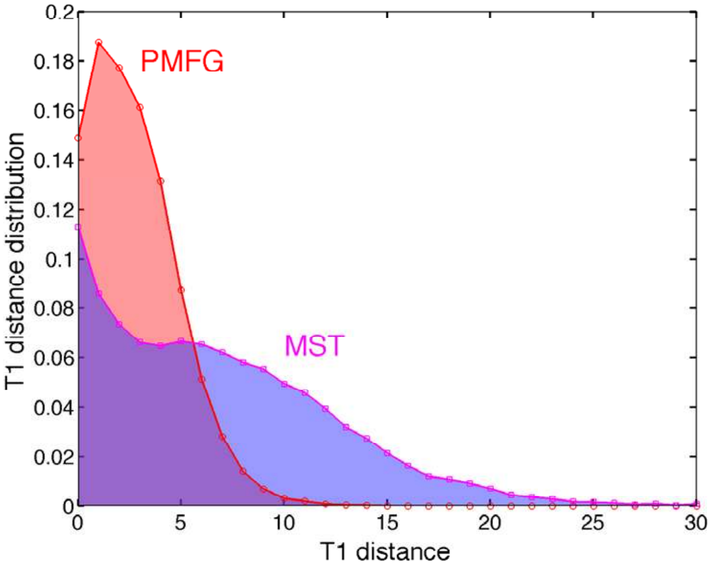

Figure 8. The distribution of the T1 distances between nodes in graphs generated from the data in the time window between t and $t + \Delta$ (with $\Delta = 3$ months) for couples of nodes that were connected in the previous window ($t - \Delta$ and t). Magenta squares refer to MST and red circles to PMFG.

connected nodes. For instance, if two nodes that were connected in the previous sub-period are still connected in the present sub-period, they are given a T1 distance equal to zero. On the other hand, if they have now become second neighbors, they are considered to have a T1 distance equal to one. If they are instead third neighbors, then they are associated with a T1 distance equal to two, etc. The distributions of these T1 distances for both the MSTs and the PMFGs are reported in figure 8. Such distributions are the average relative frequency of T1 distances and they represent an empirical estimate of the probability of a couple of nodes, which are first neighbors (zero T1 distance) in the graph computed over the time window between $t - \Delta$ and $t - 1$, becoming separated by a given T1 distance in the graph computed over the following time-window between t and $t + \Delta$. The frequencies are calculated over a set of 54 graphs computed using a time window of $\Delta = 3$ months over the time period from 1/1/1996 to 30/4/2009. We can see that in the PMFG, 15% of the edges rest connected as first neighbors (zero T1 distance), whereas a lower fraction of 11% is measured in the MST. We can also see that the distribution of the T1 distance in the PMFG drops much faster than the distribution in the MST, indicating that the occurrence of large T1 distances is less likely in the PMFG than in the MST. We can, for instance, measure that, on average, in the PMFG, almost 70% of the previous neighbors rest within three T1 moves in the following graph, whereas in the MST only 30% of the previous neighbors are within three T1 moves. A similar difference is found in the average T1 distances measured between graphs computed over subsequent time windows. This is shown in figure 9, where we can observe that subsequent PMFGs have average T1 distances fluctuating around 3, whereas the MSTs have significantly larger values around 7. All these measures indicate that the PMFG is more stable than the MST, with graphs computed over subsequent time windows possessing a larger number of edges in common (T1 distance equal to zero) and showing an overall shorter distance between the vertices that become separated. However, if we divide

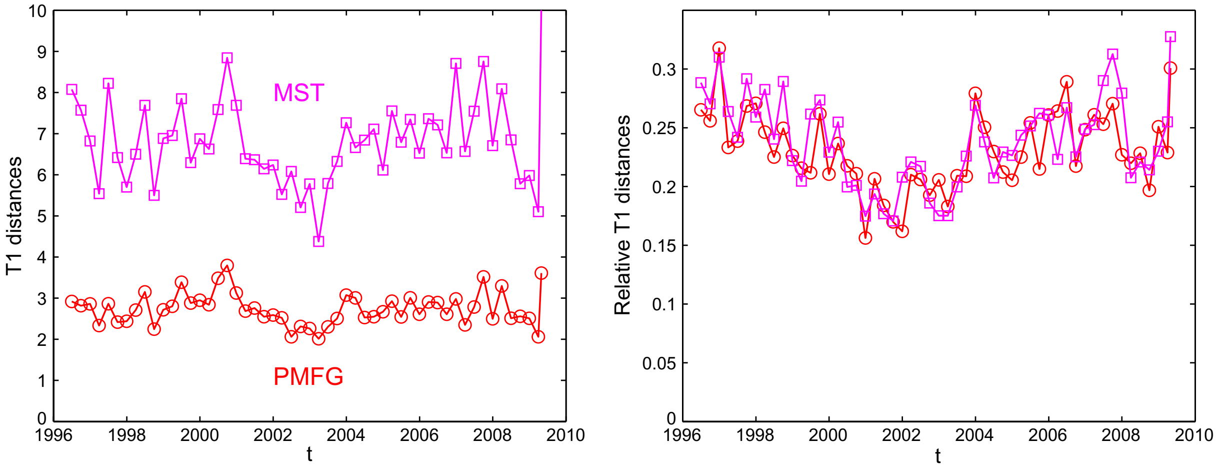

Figure 9. Left: average topological distance in the graphs calculated from the window between time t and $t + \Delta$ (with $\Delta = 3$ months) for the connected nodes in the previous window between time $t - \Delta$ and t. Blue squares refer to MST and red circles to PMFG. The horizontal dashed lines are the average distances. Right: the above distances divided by the diameters of the relative graphs.

these distances by the respective graph diameters, we see that the data become very similar, with values ranging from 0.15 to 0.35, revealing therefore that, even if in absolute terms the PMFG is more stable than the MST, in relative terms they are comparably stable.

5. Fall of the financial sector

It is rather intuitive that stocks belonging to the financial sector must have an important role within financial markets. This was, for instance, recently highlighted in [28, 33]. Furthermore, the financial sector has been indicated as a central player in the ripening and unfolding of the 'credit crunch' crisis. Here, we make use of our filtered graphs to quantify how relevant and central the financial firms have been in the financial market in the past 13 years and how the 2008–2009 financial crisis has changed the scenario. To this end, we have computed graph-based topological measures that quantify the relative position of a vertex (a stock) within the PMFG (the market). The simplest of these quantities is the *degree*, which counts the number of edges connected to each vertex. This is a local measure of centrality with the more 'popular' nodes having a larger degree. A non-local measure of centrality can be inferred by computing the number of shortest paths that pass through a given vertex: the larger the number of paths, the more central the node. The *betweenness-centrality* of vertex i is computed by summing over the ratio between the number of shortest paths $\sigma_{k,j}(v)$ that pass through vertex v with respect to the total number of shortest paths $\sigma_{k,j}$ between each couple of vertices $k, j \neq v$ and $k \neq j$ in the graph: $C_b(v) = \sum_{k,j \neq i} \sigma_{k,j}(v)/\sigma_{k,j}$. An opposite kind of measure is the *eccentricity*, which is the largest distance between a vertex v and any other vertex in the graph. Similarly, the *closeness* measures the average distance between a vertex v and all other vertices in the graph. These last two are measures of 'peripherality' of a vertex. Indeed, vertices with larger eccentricity or closeness are more in the outside part of the graph and are more distant from the other vertices, and presumably less 'influential' on the whole system.

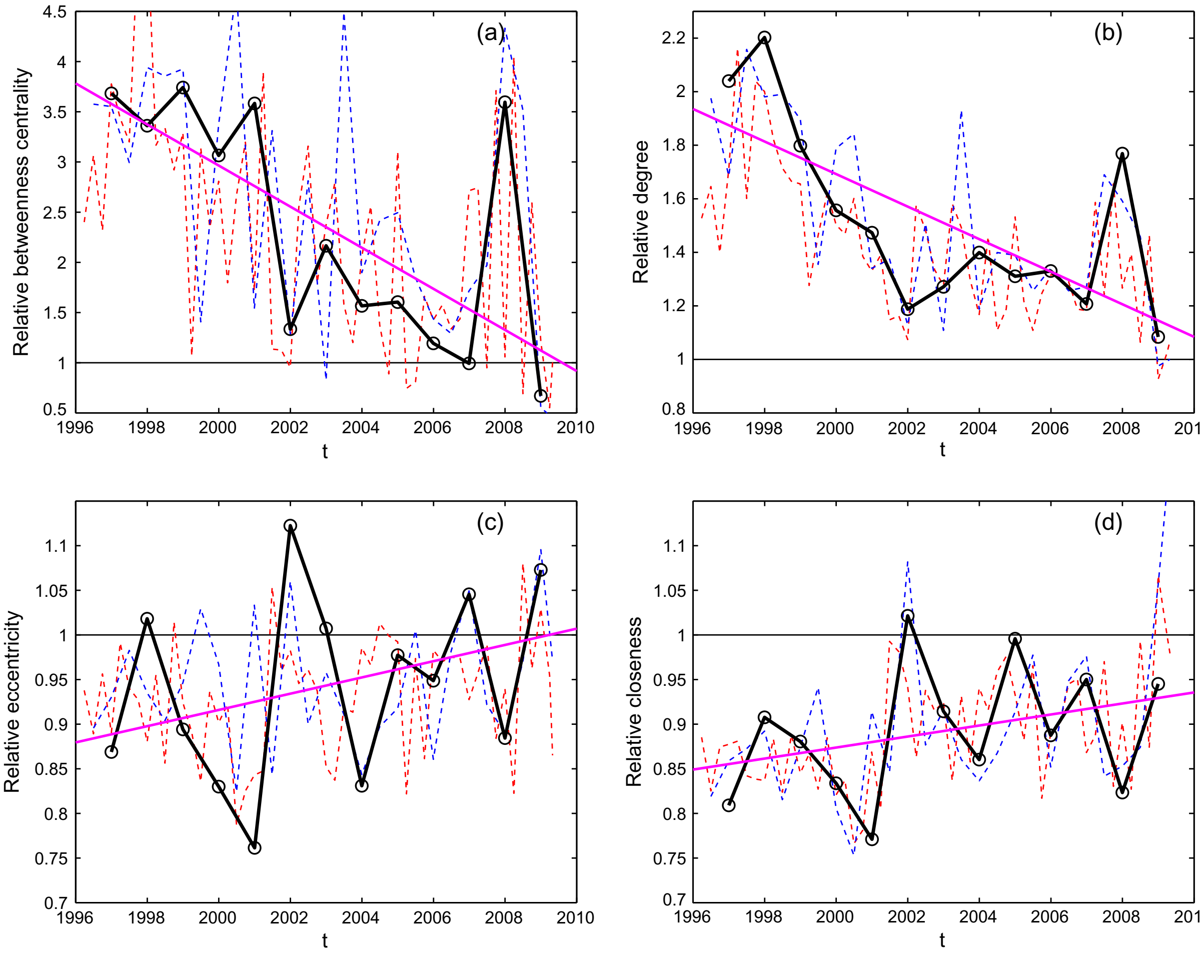

Figure 10. Changes in the relative average *betweenness centrality* (a), *degree* (b), *eccentricity* (c) and *closeness* (d), measured for a set of 34 stocks belonging to the financial sector (stock names listed in the appendix). These results refer to PMFGs computed over moving time windows of 3 months (dashed red lines), 6 months (dashed blue lines) and 1 year (full black lines with ○ symbol). The line in magenta is the trend of each of the measures calculated from the linear best fit of the data for 1 year window size.

We have computed these four measures of centrality/peripherality on the PMFG for all 395 stocks. In order to assess the *relative* positioning of the financial sector within the PMFG, we have considered a subset of 34 stocks belonging to the financial sector (stock names listed in the appendix), and we have compared the average value of each of these measures with respect to the averages for the entire system of 395 stocks. In figure 10(a), we report the values of the average betweenness-centrality for the subset of 34 stocks belonging to the financial sector divided by the average over the whole graph. One can see from this figure that the financial sector plays a very central role in the market with the average betweenness-centrality of the financial firms being up to four times as large as the average betweenness-centrality over all firms. However, we observe a clear declining trend of the central positioning of this set of stocks over time, with values of the relative betweenness-centrality passing from about four times the average in 1996 to values below average in the first months of 2009. The relative average degree, reported in figure 10(a), shows a very similar trend. These measures of decreased centrality are consistently supported by the relative average eccentricity and closeness (figures 10(c) and (d)), which show an overall increase in peripherality over the same period. It is interesting to note that this trend is rather continuous and it seems to indicate a change in the market structure that has been developing over the last 13 years. Meaningfully, the overall declining centrality has a significant peak in 2008 at the threshold of the unfolding financial crisis where the financial sector has become for a while central again, but for the 'wrong' reasons being the player that was dragging down the rest of the market. One can note that a similar short increase in centrality followed by an overall drop is observed during the 2001–2002 period. In figure 10, the same measures from graphs computed over time windows of different sizes, respectively, of 3 months, 6 months and 1 year are also reported. We can see that, despite sizable differences in the fluctuation, the overall trend is very well reproduced by all measures. Let us also note that weighted measures (where the edges are counted with the weight of the associated distance $d_{i.j} = \sqrt{2(1 - \rho_{i,j})}$) give compatible results.

6. Conclusions

In this paper, we have reported an empirical study of the evolution of stock prices for 395 firms quoted on the US equity market in the time period from 1 January 1996 to 30 April 2009. The collective dynamics of the stock prices have been analyzed by looking at the cross-correlation matrix between the log-returns observing fluctuations in the average correlations that are significantly related to market cycles. Signatures of market changes affecting the correlation structure have also been observed in the distribution of the cross-correlation coefficients, which reveals the occurrence of important changes in the shape of the distribution in relation to different market periods. The significance of such a measure of dependency has been extensively analyzed by comparing the measured correlation coefficients with surrogated coefficients obtained from randomly shuffled data. We have introduced a graph-based approach to filter out relevant information from the cross-correlation matrices, eliminating the redundancies and keeping a backbone network of significant links between firms. This approach consists in a hierarchical construction of planar graphs with maximal correlation weight: PMFGs. We have shown that these emerging graphs are very meaningful from an economic perspective.

The evolution and stability of these networks have been investigated by systematically comparing networks generated over a set of subsequent non-overlapping time intervals of 3 months. We have verified that the links selected by the PMFG are almost all associated with significant correlation coefficients. We have also introduced the 'T1 distance', a measure of distance between first neighbor nodes that might become separated by longer paths in the graph from subsequent time periods. This analysis reveals that, in the PMFG, there is a large portion of nodes that either remain linked or get separated by only a few links (three on average). A comparison with the corresponding MST shows that the PMFG is more stable.

Finally, we have used this graph filtering approach to monitor the relative evolution of 34 stocks belonging to the financial sector. Intriguingly, we have observed that two measures of centrality of the stocks belonging to this sector with respect to all the stocks in the system (relative betweenness-centrality and relative degree) change from large to small values, indicating a change from a very predominant position in 1996 to a marginal position in 2009. Consistently, two measures of peripherality of the stocks belonging to this sector with respect to all the stocks in the system (relative eccentricity and relative closeness) steadily increase during the same period. All these measures consistently highlight a declining influence of the financial sector. This might be a consequence of the increasingly speculative nature of the financial activity that since the late 1990s has relatively reduced its original role of service and support to other firms.

Acknowledgments

We thank the Quantitative Analysis Group at Nomura International plc for the provision of the data on which this research is based. This work was partially supported by the COST MP0801 project.

Appendix. Financial sector firms

1. Aflac Inc [AFL]
2. The Allstate Corporation [ALL]
3. American Express Co [AXP]
4. American International Group Inc [AIG]
5. AON Corp [AOC]
6. Bank of New York Mellon Corp/THE [BK]
7. Chubb Corp [CB]
8. Cigna Corp [CI]
9. Cincinnati Financial Corp [CINF.]
10. Comerica Inc [CMA]
11. Freddie Mac [FRE]
12. Fannie Mae [FNM]
13. Franklin Resources Inc [BEN]
14. Keycorp [KEY]
15. Lincoln National Corp [LNC]
16. Loews Corp [L]
17. Marsh & McLennan Cos Inc [MMC]
18. Bank of America Corp [BAC]
19. Wells Fargo & Co [WFC]
20. PNC Financial Services Group Inc [PNC]
21. Citigroup Inc [C]
22. Progressive Corp/THE [PGR]
23. Travelers Cos Inc/THE [TRV]
24. Suntrust Banks Inc [STI]
25. Torchmark Corp [TMK]
26. MBIA Inc [MBI]

27. ACE Ltd [ACE]
28. XL Capital Ltd [XL]
29. Lehman Brothers Holdings Inc [LEHMQ]
30. Ambac Financial Group Inc [ABK]
31. Capital One Financial Corp [COF]
32. MGIC Investment CP [MTG]
33. Synovus Financial [SNV]
34. Hartford Financial Services Group Inc [HIG]

References

[1] Lux T and Marchesi M 1999 Scaling and criticality in a stochastic multi-agent model of financial market *Nature* **397** 498–500

[2] Schweitzer F, Battiston S and Tessone C J (eds) 2009 *Topical Issue on the Physics Approach to Risk: Agent-Based Models and Networks Eur. Phys. J.* B **71**(4) 439–648

[3] http://uk.reuters.com/

[4] Mantegna R N and Stanley H E 2000 *An Introduction to Econophysics: Correlations and Complexity in Finance* (Cambridge: Cambridge University Press)

[5] Platen E and Heath D 2006 *A Benchmark Approach to Quantitative Finance* (Berlin: Springer)

[6] Sengupta A M and Mitra P P 1999 Distributions of singular values for some random matrices *Phys. Rev.* E **60** 3389–92

[7] Laloux L, Cizeau P, Bouchaud J-P and Potters M 1999 Noise dressing of financial correlation matrices *Phys. Rev. Lett.* **83** 1467–70

[8] Plerou V, Gopikrishnan P, Rosenow B, Amaral L A N and Stanley H E 1999 Universal and nonuniversal properties of cross correlations in financial time series *Phys. Rev. Lett.* **83** 1471–4

[9] Barabási A L and Albert R 1999 Emergence of scaling in random networks *Nature* **286** 509–12

[10] Amaral L A N, Scala A, Barthelemy M and Stanley H E 2000 Classes of small-world networks *Proc. Natl Acad. Sci. USA* **97** 11149–52

[11] Newman M E J 2003 The structure and function of complex networks *SIAM Rev.* **45** 167–256

[12] Caldarelli G 2007 *Scale-Free Networks Complex Webs in Nature and Technology* (Oxford: Oxford Univesity Press)

[13] Amaral L and Ottino J 2004 Complex networks *Eur. Phys. J.* B **38** 147–62

[14] Mantegna R N 1999 Hierarchical structure in financial markets. *Eur. Phys. J.* B **11** 193–7

[15] Bonanno G, Lillo F and Mantegna R N 2001 High-frequency cross-correlation in a set of stocks *Quant. Finance* **1** 96–104

[16] Onnela J-P, Chakraborti A, Kaski K and Kertész J 2002 Dynamic asset trees and portfolio analysis *Eur. Phys. J.* B **30** 285–8

[17] Onnela J-P, Chakraborti A, Kaski K, Kertész J and Kanto A 2003 Dynamics of market correlations: taxonomy and portfolio analysis *Phys. Rev.* E **68** 056110

[18] Onnela J-P, Chakraborti A, Kaski K and Kertész J 2003 Dynamic asset trees and black monday *Physica* A **324** 247–52

[19] Aste T, Di Matteo T and Hyde S T 2005 Complex networks on hyperbolic surfaces *Physica* A **346** 20–6

[20] Tumminello M, Aste T, Di Matteo T and Mantegna R N 2005 A tool for filtering information in complex systems *Proc. Natl Acad. Sci. USA* **102** 10421–6

[21] Prim R C 1957 Shortest connection networks and some generalizations *Bell Syst. Tech. J.* **36** 1389–401

[22] Kruskal J B 1956 On the shortest spanning subtree of a graph and the traveling salesman problem *Proc. Am. Math. Soc.* **7** 48–50

[23] Chazelle B 2000 A minimum spanning tree algorithm with inverse-Ackermann type complexity *J. ACM* **47** 1028–47

[24] Aste A 2010 An algorithm to compute maximally filtered planar graphs http://www.mathworks.com/matlabcentral/fileexchange/27360

[25] Ringel G 1974 *Map Color Theorem* (Berlin: Springer)

[26] Aste T and Di Matteo T 2006 Dynamical networks from correlations *Physica* A **370** 156–61

[27] Tumminello M, Aste T, Di Matteo T and Mantegna R N 2007 Correlation based networks of equity returns sampled at different time horizons *Eur. Phys. J.* B **55** 209–17

[28] Di Matteo T, Pozzi F and Aste T 2010 The use of dynamical networks to detect the hierarchical organization of financial market sectors *Eur. Phys. J.* B **73** 3–11

[29] Aste T, Boosé D and Rivier N 1996 From one cell to the whole froth: a dynamical map *Phys. Rev.* E **53** 6181–91

[30] Aste T, Szeto K Y and Tam W Y 1996 Statistical properties and shell analysis in random cellular structures *Phys. Rev.* E **54** 5482

[31] Ohlenbusch H M, Aste T, Dubertret B and Rivier N 1998 The topological structure of 2d disordered cellular systems *Eur. Phys. J.* B **29** 211–20

[32] Aste T and Sherrington D 1999 Glass transition in self organizing cellular patterns *J. Phys. A: Math. Gen.* **32** 7049–56

[33] Pozzi F, Di Matteo T and Aste T 2008 Centrality and peripherality in filtered graphs from dynamical financial correlations *Adv. Complex Syst. (ACS)* **11** 927–50