Weighted network analysis of high-frequency cross-correlation measures

Giulia Iori*
Department of Economics, City University, Northampton Square, London, EC1V 0HB, United Kingdom

Ovidiu V. Precup†
Department of Mathematics, London School of Economics, Houghton Street, London, WC2A 2AE, United Kingdom

(Received 25 October 2006; revised manuscript received 19 December 2006; published 26 March 2007)

In this paper we implement a Fourier method to estimate high-frequency correlation matrices from small data sets. The Fourier estimates are shown to be considerably less noisy than the standard Pearson correlation measures and thus capable of detecting subtle changes in correlation matrices with just a month of data. The evolution of correlation at different time scales is analyzed from the full correlation matrix and its minimum spanning tree representation. The analysis is performed by implementing measures from the theory of random weighted networks.

DOI: 10.1103/PhysRevE.75.036110 PACS number(s): 89.65.Gh

I. INTRODUCTION

Robust correlation measures are important for derivatives pricing, risk management, portfolio optimization, and understanding market microstructure effects. The conventional method of computing correlation is the Pearson coefficient. This method requires homogeneous time series. In order to apply it to high-frequency data, the time series first need to be homogenized and synchronized through an interpolation scheme. An alternative, nonparametric approach has been suggested in [1] where the variance-covariance matrix estimator of a multivariate process is computed via Fourier analysis. Previous applications of the method can be found in [2–7].

In this paper we compare the performance of the Pearson and Fourier methods by computing returns cross-correlation matrices at different time scales using one month (September 2002) of high-frequency trades in the member stocks of the S&P100 index [8]. The selected stocks are grouped into 12 different industry sectors [9]: technology (16 stocks), basic materials (seven stocks), financial (13 stocks), capital goods (three stocks), conglomerates (five stocks), energy (four stocks), services (16 stocks), transport (four stocks), utilities (seven stocks), health care (ten stocks), noncyclical consumer goods (noncyclical CG) (11 stocks), cyclical consumer goods (CG) (four stocks). Three-quarters of the stocks included in this analysis are very liquid and trade on average at intervals shorter than 14 s. The least liquid stock, Allegheny Technologies, has an average trading time of about a minute.

The estimation of intraday correlations over short periods of time (e.g., a month) is of high practical value for day trading and hedging purposes. In fact, such estimates are more sensitive to short time scale economic factors than correlation measures obtained from averaging over several months. Thus, we choose to investigate a month of tick-by-tick data aiming to compare the quality of the information that can be derived by applying each of the two methods on limited statistics. The Fourier estimates reproduce the structural changes on filtered correlation matrices observed in previous studies [10–18] with much larger data sets. Moreover, we show that the Fourier estimates are sufficiently accurate to reveal further structural changes in the full, unfiltered, correlation matrices.

II. FOURIER CORRELATION MEASURE

The Fourier method is model independent, produces very accurate, smooth estimates, and handles the time series in their original form without imputation or discarding of data. A rigorous proof of the method is given in the original paper by Malliavin and Mancino [1] and only the main results are summarized below.

The method works as follows. Let $S_i(t)$ be the price of asset i at time t and $p_i(t)=\ln S_i(t)$. The physical time interval $[0,T]$ of the asset price series is rescaled to $[0,2\pi]$. In this case T represents the length of the entire time series expressed in the basic time unit of analysis. For example, if we analyze the returns on an intrahourly time scale using one month of trading data, T will have the value 10 560 min, which is the equivalent of 22 trading days with 8 h of trading per day. The variance-covariance matrix Σ_{ij} of log (returns) is derived from its Fourier coefficient $a_0(\Sigma_{ij})$ which is obtained from the Fourier coefficients of dp_i:

$$a_k(dp_i) = \frac{1}{\pi}\int_0^{2\pi} \cos(kt)dp_i(t),$$

$$b_k(dp_i) = \frac{1}{\pi}\int_0^{2\pi} \sin(kt)dp_i(t), \quad k \geqslant 1. \tag{1}$$

In practice, the coefficients are computed through integration by parts. As $p_i(t)$ is not observed continuously but given by unevenly spaced tick-by-tick observations of trade prices, the actual implementation requires the integrals in (1) to be in discrete form:

*Corresponding author. Email address: g.iori@city.ac.uk
†Email address: O.V.Precup@lse.ac.uk

$$a_k(dp_i) = \frac{1}{\pi}\sum_{n=1}^{N} \{[p_i(t_n)\cos(kt_n) - p_i(t'_n)\cos(kt'_n)] - p_i(t'_n) \times [\cos(kt_n) - \cos(kt'_n)]\},$$

$$b_k(dp_i) = \frac{1}{\pi}\sum_{n=1}^{N} \{[p_i(t_n)\sin(kt_n) - p_i(t'_n)\sin(kt'_n)] - p_i(t'_n) \times [\sin(kt_n) - \sin(kt'_n)]\}, \quad (2)$$

where $t'_n = t_{n-1}$.

In (2), N corresponds to the number of trades in the rescaled interval and we set the price $p_i(t) = p_i(t_{n-1})$ to compute the integrals between two consecutive trading times $[t_{n-1}, t_n]$.

The Fourier coefficient of the pointwise variance-covariance matrix Σ_{ij} is

$$a_0(\Sigma_{ij}) = \lim_{\tau \to 0} \frac{\pi\tau}{T}\sum_{k=1}^{T/2\tau} [a_k(dp_i)a_k(dp_j) + b_k(dp_i)b_k(dp_j)]. \quad (3)$$

The integrated value of Σ_{ij} over the time window is defined as $\hat{\sigma}^2_{ij} = 2\pi a_0(\Sigma_{ij})$ which leads to the Fourier correlation matrix $\rho_{ij} = \hat{\sigma}^2_{ij}/(\hat{\sigma}_{ii} \cdot \hat{\sigma}_{jj})$.

The highest wave harmonic $(T/2\tau)$ that can be analyzed is determined by the lower bound of τ (time gap between two consecutive trades) which is 1 s for all S&P100 price series. In this analysis we take $\tau = 3$ min as the shortest time scale and $\tau = 120$ min as the longest one. These values have been chosen to avoid asynchronicity bias at very short τ and statistical errors at longer τ, due to the limited length of the time series.

III. NETWORK ANALYSIS

The correlation matrix can be represented as a network of vertices (stocks) and weighted links (correlations). As a way of filtering information from noise in correlations, work [10–15] has focused on the minimum spanning tree[1] (MST) representation of correlation matrices. In this paper we study both the full correlation matrices and their MST representations using weighted network analysis measures.

Following [19,20] we define the *degree* of a vertex in the network as $k_i = \Sigma_{j \in \mathcal{V}(i)} 1_{ij}$ where the sum runs over the set $\mathcal{V}(i)$ of neighbors of i and 1_{ij} is an indicator function for whether there is a connection between i and j. The *strength* of a vertex is defined as $s_i = \Sigma_{j \in \mathcal{V}(i)} c_{ij}$ where c_{ij} is the correlation between vertices (stocks) i and j. We use the degree k_i as a measure of stock centrality for MSTs and the strength s_i as a measure of stock centrality in the original, unfiltered correlation matrices. For the weighted clustering coefficient we use the definition suggested in [21,22]:

[1]Given a connected, undirected graph, a spanning tree of that graph is a subgraph which is a tree and connects all the vertices together. Here we use the distance $d_{i,j} = \sqrt{2(1-c_{i,j})}$ as the weight of each edge. A minimum spanning tree or minimum weight spanning tree is then a spanning tree with weight less than or equal to the weight of every other spanning tree.

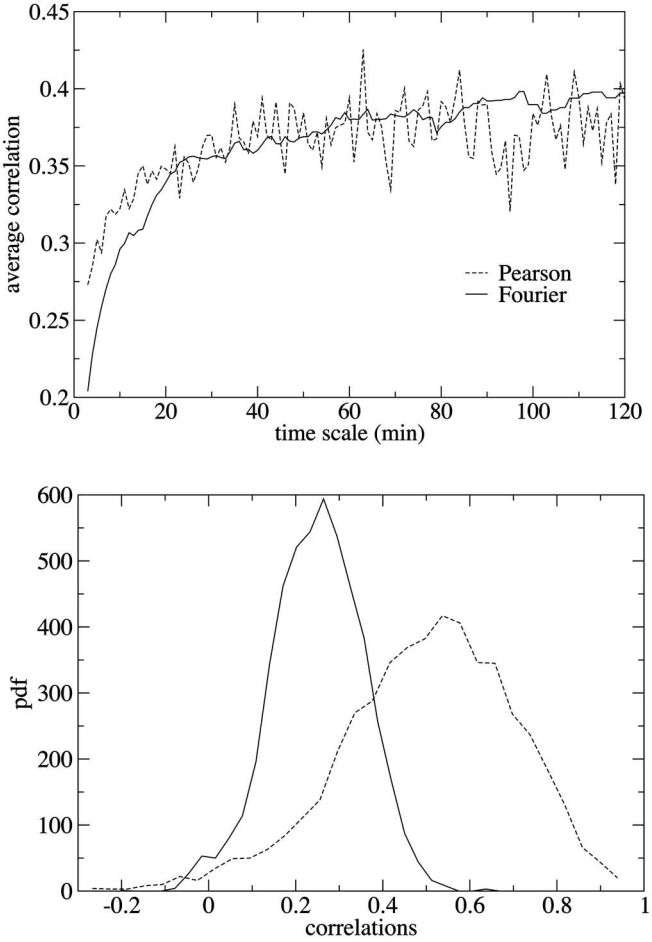

FIG. 1. (Top) Average correlation across all stocks increases with the time scale for both the Fourier (continuous line) and the Pearson (dashed line) methods. (Bottom) Correlation density function for two different time scales: 10 (continuous line) and 100 min (dashed line).

$$C_i^w = \frac{\sum_{j,h} c_{ij}c_{ih}c_{jh}}{\sum_{j,h} c_{ij}c_{ih}}. \quad (4)$$

This definition reduces to the standard clustering coefficient in the binary case and retains the property $0 \leq C_i^W \leq 1$. For alternative definitions of the clustering coefficient see, for example, [23,24].

For our analysis we consider only the positive elements of the correlation matrices. This is to avoid a spurious effect on the clustering coefficient and the strength resulting from negative correlations. Less than 2% of the correlations are negative as can be seen (for $\tau = 10$ and 100 min) from Fig. 1 (bottom). Even with this choice the correlation matrix remains almost fully connected.

When analyzing intraday data, the choice of time scale on which to measure correlations becomes crucial. In Fig. 1 we plot the average correlation at different time scales, from 3 min to 2 h. The average correlation increases with the time scale, a result known as the Epps effect [25]. A possible

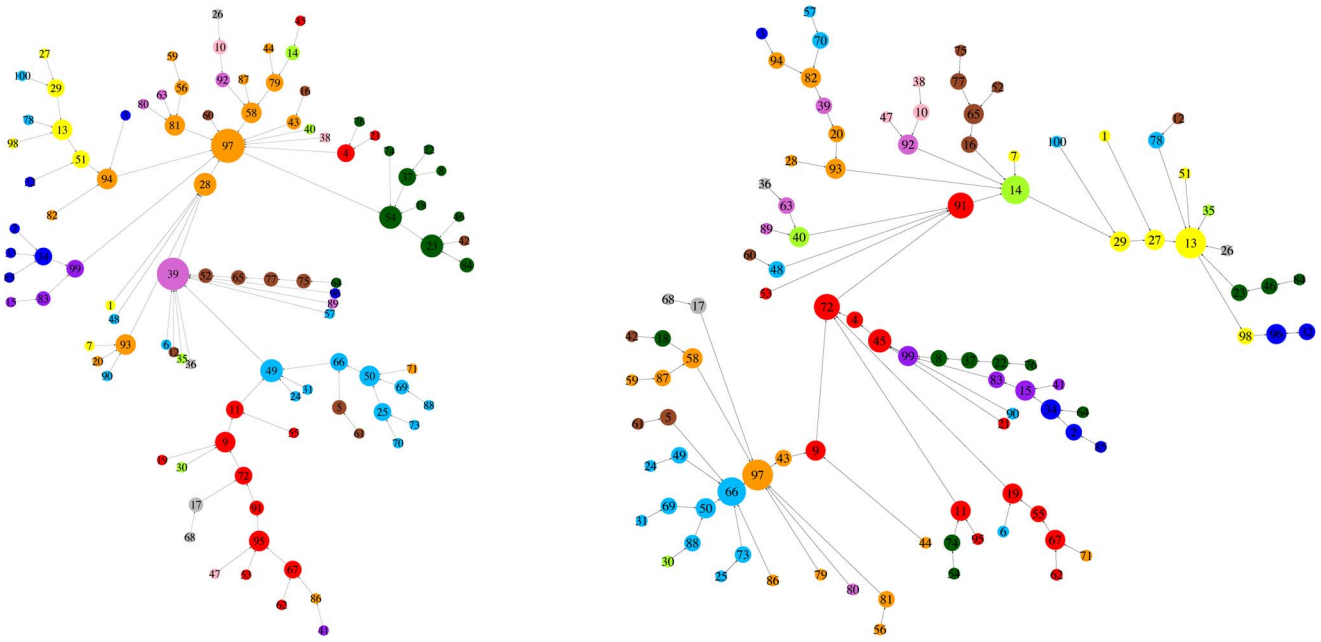

FIG. 2. (Color online) MST obtained with the Fourier method at 10 (left) and 90 min (right). WMT is node 97 and GE is node 39. The size of the dots representing the different stocks is proportional to the number of links. The color codes of the different industry sectors are given in Table I below.

explanation for this effect has been recently proposed [4,6,26] not in terms of economic factors but as a consequence of an asynchronicity bias which is particularly relevant when correlation is measured between illiquid stocks. Nonetheless, the average correlation increase with time scale is accompanied by a structural change in the correlation matrix as shown in [10–15] and more recently using the planar maximally filtered graph method in [16–18]. This fact is difficult to explain purely in terms of the asynchronicity bias, as this would imply that at short time scales the most central sector is the most liquid one (i.e., technology stocks), which is not the case. On the contrary, the cluster of technology stocks is on the periphery of the graphs at both time scales shown in Fig. 2.

The above mentioned studies demonstrate that the shape of the MST changes substantially with the time scale. On very short time scales the MSTs are centralized graphs with a few vertices that collect a large number of connections. On longer time scales the graph structure becomes significantly more dispersed with no obvious hubs. Figure 2 shows the same type of qualitative result in our case as well. On the left we plot the MST, obtained with the Fourier method on 10 min time scales and on the right at time scales of 90 min.

Before proceeding with the analysis we point out that the robustness of the Fourier method with respect to the length of the time series is an open question as an asymptotic theory for this method has not yet been developed. Here we attempt to quantify the estimation error by calculating the percentage differences between two correlation matrices (and MSTs) calculated at successive time scales. Differences are defined as

$$d(\tau) = \frac{1}{N^2} \sum_{i,j} \frac{|c_{i,j}(\tau) - c_{i,j}(\tau+1)|}{c_{i,j}(\tau)}.$$

It seems reasonable to assume that once correlations have stabilized (i.e., after the first 30 min during which we observe the Epps effect), the variation in correlation at consecutive 1 min difference intervals (for example 36 and 37 min returns) is not due to economic factors but is in fact attributable to estimation errors. In Fig. 3 we plot the percentage differences for the MST and the full correlation matrix. The MST is, by construction, more noisy than the correlation matrices and this is reflected in the distance fluctuations. While the fluctuations settle around 2% for the full correlation matrices, in the MST they are as high as 10%. We obtain the same figures when calculating fluctuations of an individual vertex degree, strength, and clustering. Thus, in the rest of the paper we assume a Fourier error of 10% for the MST degree and of 2% for the strength and clustering coefficient. Furthermore, we note that while the error associated with the Pearson measure increases with the time scale (dashed line), as a consequence of the interpolation procedure which inevitably discards more and more of the available data, the Fourier estimates are unaffected (at least up to the 2 h time scale) by the length of the series, due to the fact that all observations are used to estimate correlations at all time scales.

In order to quantify the structural change of the MST in Fig. 2 (left) we plot the evolution, up to a 2 h time scales, of the maximum degree in the MST. This is measured by the degree of the most connected stock at any given time (not necessarily the same one at each time scale). We compare the

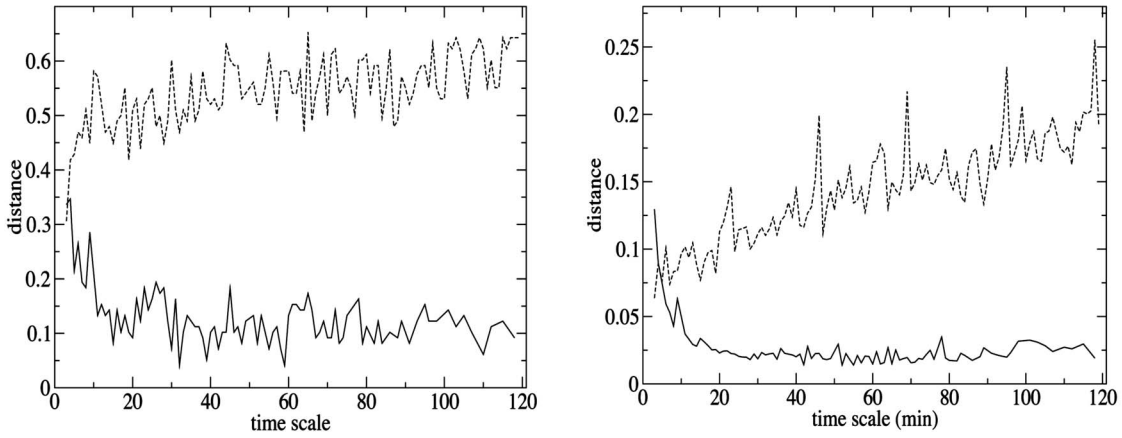

FIG. 3. Distance between MSTs (left) and correlation matrices (right) at consecutive time scales of analysis for the Fourier (continuous line) and the Pearson (dashed line) methods.

results for the Pearson (dashed line) and Fourier (solid line) correlation estimates. We notice that while the Pearson estimate gives very noisy results on this small data set (also visible in Fig. 1), the Fourier estimator provides much more consistent results across different time scales.

Figure 4 (right) shows the evolution of the maximum degree for Wal-Mart Stores (WMT) and General Electric (GE) obtained from the Fourier MST matrix. For both stocks the degree rises quickly and remains high at time scales between 10 and 20 min. We find an average degree, for $3<\tau<30$, of 7.36 ± 2.94 for Wal-Mart and 5.96 ± 2.99 for General Electric. In [18] General Electric and Wal-Mart are reported as the most connected stocks in 2002 at 5 min time scales, with the degree of GE decreasing as the time scale increases. Our results are in agreement with these previous findings. Nonetheless, when averaging at all time scales up to 2 h, WMT appears to be the most connected stock in the MST, with an average degree of 6.19 ± 2.05 versus an average degree of 2.73 ± 2.46 for GE.

We look at the strength and clustering measures (as previously defined) in order to analyze the structural changes in the full correlation matrix. While some analysis in this direction has been performed in previous studies, this was based on filtered correlation matrices (either planary filtered graphs [16–18] or graphs constructed by including only the strongest $N-1$ links, with N being the number of stocks [13]).

In Fig. 5 (left) we plot the evolution, across time scales, of the normalized strength of the most connected vertex in the full correlation matrix calculated with both the Pearson (dashed line) and the Fourier (continuous line) methods. The normalized strength, at time scale τ, is defined as $\tilde{s}_i(\tau)=s_i(\tau)/\hat{c}(\tau)$, where $\hat{c}(\tau)=\Sigma_{i,j}c_{ij}(\tau)$ is the total correlation. Without this normalization the strength would trivially increase with time scale as a result of the Epps effect. By normalizing we can quantify the way the most correlated stock is central to the network, in terms of its proportional contribution to the total correlation. We notice again [Fig. 5 (left)] that the Pearson estimator is very noisy while the Fourier estimator is significantly smoother. The Fourier estimator also indicates a rise in the most correlated stocks relative strengths at progressively shorter time scales under 20 min, analogous to the increasing degree of the most connected stock in the MST on short time scales.

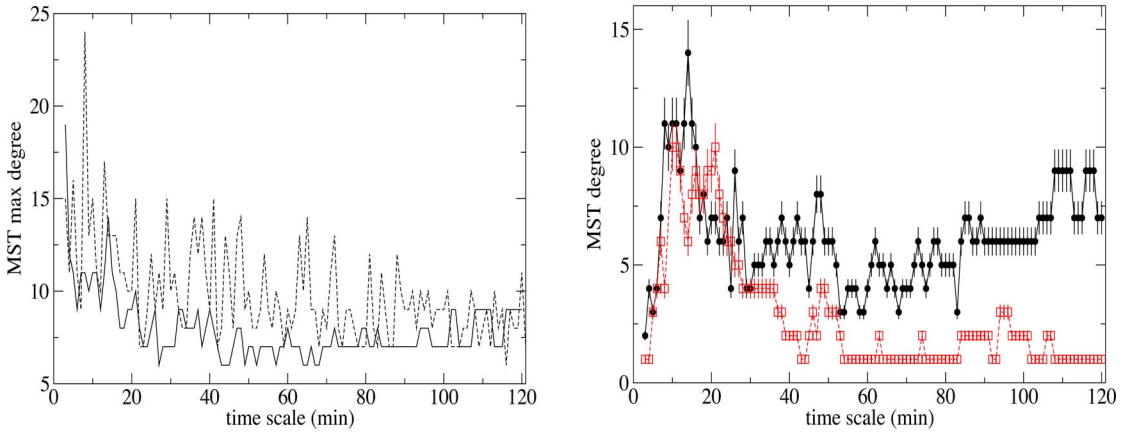

FIG. 4. (Color online) (Left) Maximum degree in the MST for Fourier (continuous line) and Pearson (dashed line) methods. (Right) Degree of GE (red, square) and WMT (black, circles) as a function of the time scale, obtained with the Fourier method.

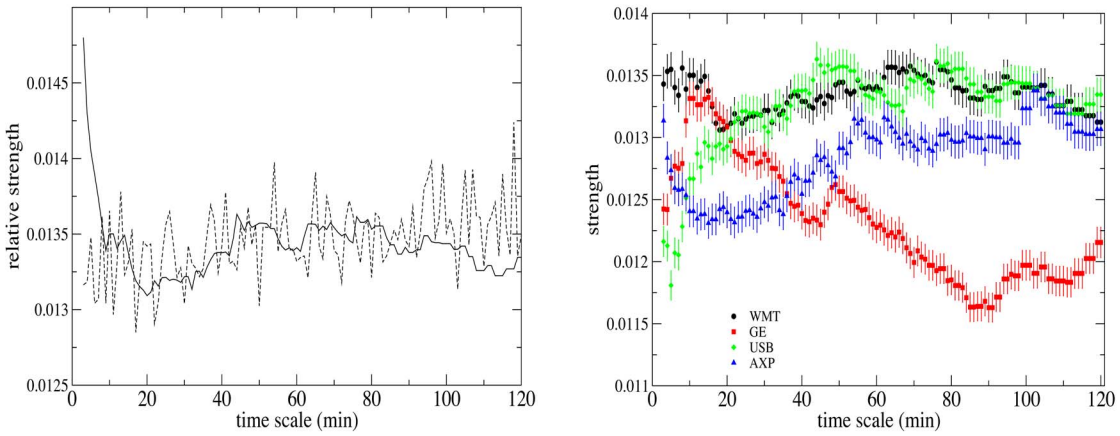

FIG. 5. (Color online) (Left) Greatest normalized strength in the correlation matrix determined with the Fourier (continuous line) and the Pearson (dashed line) methods. (Right) Fourier normalized strength of WMT (circle, black), GE (square, red), USB (diamond, green), and AXP (triangle, blue) as a function of time scale.

The stocks that contribute the most, on average, to the total correlation on time scales shorter than 30 min are again WMT and GE, with relative strengths, respectively, of $(0.0133\pm1.6)\times10^{-4}$ and $(0.0130\pm2.5)\times10^{-4}$. We also show, in Fig. 5 (right), that while the normalized strength of GE decreases with the time scale the normalized strength of WMT fluctuates around the same level across time scales. For time scales up to 2 h, WMT remains one of the most central stocks to the network, along with US Bancorp (USB) and American Express (AXP).

In Fig. 6 we plot the evolution, across time scales, of the relative weighted clustering coefficient of the most clustered stock in the full correlation matrix calculated with both the Pearson (dashed line) and Fourier (continuous line) methods. The relative weighted clustering coefficient is defined as $\tilde{C}_i^w(\tau)=C_i^w(\tau)/\bar{C}^w(\tau)$, where $\bar{C}^w(\tau)$ is the scale τ average clustering coefficient. The normalization is also necessary in this case as the clustering coefficient, defined in Eq. (4), would trivially rise purely as a consequence of the general correlation level increase with the time scale. Again the Fourier method provides smooth results which reveal that the relative clustering coefficient of the most clustered stock increases as the time scale falls below 20 min. Here, instead of identifying the stock with the highest clustering coefficient, we shift the analysis to the industry sectors. In Table I we report the average of the intrasector relative strength and intrasector relative clustering on time scales shorter than 30 min. A relative clustering (strength) larger than 1 implies that intrasector clustering (strength) is larger than the average clustering (strength) in the network. We first note that for some sectors there is a significant difference between intrasector strength and clustering, revealing that not all the stocks in that sector are highly correlated with each other. This effect is particularly evident for the cyclical consumer goods and the capital goods sectors. The most clustered sector at all time scales, up to 2 h, is the financial one. At short time scales this is followed by services, technology, energy, and noncyclical consumer goods. The study in [17] uses the same sector classification but a different selection of stocks (100 highly capitalized stocks instead of the member stocks of the S&P100) and finds that the financial and the energy sectors have the highest intrasector clustering, on planary filtered graphs, on a daily time scale. This is in agreement

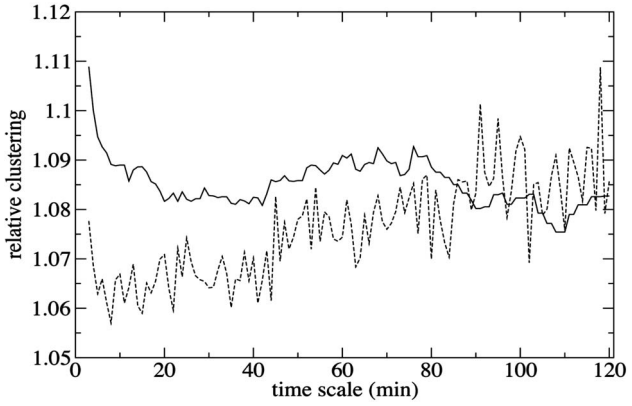

FIG. 6. Relative clustering coefficient of the highest cluster coefficient stocks by Fourier (continuous line) and Pearson (dashed line) methods.

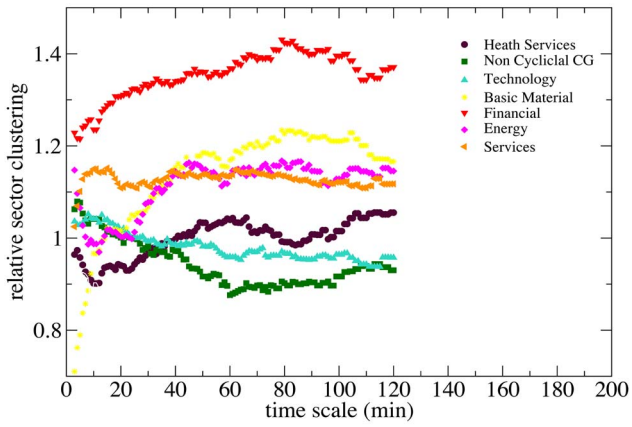

FIG. 7. (Color online) Relative clustering coefficient of the seven most clustered industrial sectors.

TABLE I. Intrasector relative strength and clustering coefficients at time scales shorter than 30 min.

Sector	Size	Intrasector strength	Intrasector clustering
Technology	16	1.13	1.02
Basic materials	7	1.04	0.97
Financial	13	1.33	1.28
Capital goods	3	0.86	0.41
Conglomerates	5	1.01	0.87
Energy	4	0.97	1.02
Services	16	1.28	1.12
Transport	4	0.99	0.74
Utilities	7	0.84	0.80
Health care	10	1.07	0.94
Noncyclical consumer goods	11	1.08	1.02
Cyclical consumer goods	4	1.10	0.66

with our results, even though we include only 13 stocks in the financial sector while in [17] 24 stocks were selected. A recent paper [27] analyzes the time evolution of the daily clustering coefficient among industrial sectors between 1984 and 2000 and (even though the paper uses a different stock classification) identifies the financial and energy sectors as the most clustered ones after 1995.

Nonetheless our analysis leads to clearly different results for the services sector (which includes Wal-Mart) which [17] report as being poorly intraconnected. A possible reason for this disagreement could be that the composition of this sector is very different in the two studies (in [17] this sector is composed of 20 stocks while in our study only seven stocks are present). Another possible reason may be the difference in time scale at which the correlations are measured. In Fig. 7 we plot the relative clustering coefficient for the seven most clustered sectors as a function of the time scale. We can see that the ranking of sectors in terms of their relative clustering coefficients changes considerably over time, and in particular the services sector, which is the second most clustered at short time scales, becomes only the fourth most clustered on 2 h time scales. It may well be that the relative clustering of this sector decreases even further on daily time scales. The high clustering coefficient of some sectors is reflected in the MST. For example, the MST at 10 min in Fig. 2 (left) identifies very clearly the clusters associated with the financial (red, triangle down), services (orange, triangle left), noncyclical consumer goods (green, square), and technology (turquoise, triangle up) sectors. In contrast, at 90 min both services and noncyclical consumer goods clusters are broken while in addition to the financial and technology groups, the basic materials sector (yellow, star) (the second most clustered sector at this time scale) can be clearly identified.

IV. CONCLUSIONS

The analysis carried out in this paper provides further evidence that the Fourier method of computing the correlation matrix from high-frequency data is better than the traditional Pearson alternative in terms of generating smooth estimates from small sample data sets. Unfortunately, while work is ongoing to establish this, no asymptotic theory for the robustness of the Fourier method is currently available.

The Fourier MST representation of the correlation matrix exhibits similar characteristics to those found in previous studies. The graph is centralized on a very short time scale and becomes more dispersed on longer time scales. The analysis of the entire correlation matrix provides additional evidence of the structural changes that affect the correlation matrix at different time scales. As a result of our analysis we find that Wal-Mart Stores and General Electric are the two most central stocks both in the MST and in the full correlation network on time scales shorter than 20 min. Furthermore, Wal-Mart has one of the highest centrality scores at all time scales up to 2 h. At aggregate level we have identified the financial, energy, and services sectors as the most intraconnected at short time scales with the financial being the most intraconnected at all time scales up to 2 h.

ACKNOWLEDGMENTS

We are very grateful to Roberto Renó, Vanessa Mattiussi, Anirban Chakraborti, and Rosario Mantegna for stimulating discussions.

[1] P. Malliavin and M. Mancino, Finance Stoch. **6**, 49 (2002).
[2] E. Barucci and R. Renó, J. Int. Financial Markets, Institutions Money, 12, 182 (2002).
[3] E. Barucci and R. Renó, Econ. Lett. **74**, 371 (2002).
[4] R. Renó, Int. J. Theor. Appl. Finance **6**, 87 (2003).
[5] O. V. Precup and G. Iori, Physica A **344**, 252 (2004).
[6] O. V. Precup and G. Iori, Eur. J. Finance (to be published).
[7] V. Mattiussi and G. Iori, *Debt, Risk and Liquidity in Futures Markets*, edited by B. A. Goss (Routledge, London, in press), Chap. 5.
[8] NYSE Trades and Quotes (TAQ) database. www.ngsedata.com
[9] According to the classification provided by finance.yahoo.com
[10] G. Bonanno, F. Lillo, and R. N. Mantegna, Quant. Finance **1**, 96, (2001).
[11] B. Bonanno, G. Caldarelli, F. Lillo, and R. N. Mantegna, Phys. Rev. E **68**, 046130 (2003).
[12] G. Bonanno, G. Caldarelli, F. Lillo, S. Micciché, N. Vandewalle, and R. N. Mantegna, Eur. Phys. J. B **38**, 363 (2004).
[13] J.-P. Onnela, A. Chakraborti, K. Kaski, J. Kertesz, and A. Kanto, Phys. Scr., T **106**, 48 (2003).
[14] J.-P. Onnela, A. Chakraborti, K. Kaski, and J. Kertesz, Physica A **324**, 247 (2003).
[15] J.-P. Onnela, K. Kaski, and J. Kertesz, Eur. Phys. J. B **38**, 353 (2004).

[16] M. Tumminello, T. Aste, T. Di Matteo, and R. N. Mantegna, Proc. Natl. Acad. Sci. U.S.A. **102**, 10421 (2005).
[17] C. Coronnello, M. Tumminello, F. Lillo, S. Micciché, and R. N. Mantegna, e-print physics/0609036.
[18] M. Tumminello, T. Di Matteo, T. Aste, and R. N. Mantegna, e-print physics/0605251.
[19] A. Barrat, M. Barthélemy, and A. Vespignani, Phys. Rev. Lett. **92**, 228701 (2004).
[20] M. E. J. Newman, Phys. Rev. E **70**, 056103 (2004).
[21] Grindrod, Phys. Rev. E **66**, 066702 (2002).
[22] B. Zhang, and S. Horvath, Stat. Appl. Genet. Mol. Biol. **4**, 17 (2005).
[23] J.-P. Onnela, J. Saramki, J. Kertsz, and K. Kaski, Phys. Rev. E **71**, 065103(R) (2005).
[24] A. Barrat, M. Barthelemy, R. Pastor-Satorras, and A. Vespignani, Proc. Natl. Acad. Sci. U.S.A. **101**, 3747 (2004).
[25] T. Epps, J. Am. Stat. Assoc. **74**, 291 (1979).
[26] Lan Zhang, http://ssrn.com/abstract=885438.
[27] G. Tibely *et al.*, Physica A **370**, 145 (2006).