Clustering Financial Time Series by Network Community Analysis

Article *in* International Journal of Modern Physics C · January 2011

DOI: 10.1142/S012918311101604X

CITATIONS
17

READS
596

3 authors, including:

Lisa Calatroni
Politecnico di Milano
3 PUBLICATIONS **78** CITATIONS
SEE PROFILE

Fabio Bertoni
SKEMA Business School
80 PUBLICATIONS **1,436** CITATIONS
SEE PROFILE

International Journal of Modern Physics C

CLUSTERING FINANCIAL TIME SERIES BY NETWORK COMMUNITY ANALYSIS

CARLO PICCARDI
DEI, Politecnico di Milano
Piazza Leonardo da Vinci 32
20133 Milano, Italy
carlo.piccardi@polimi.it

LISA CALATRONI
DEI, Politecnico di Milano
Piazza Leonardo da Vinci 32
20133 Milano, Italy

FABIO BERTONI
DIG, Politecnico di Milano
Via Lambruschini 4/b
20156 Milano, Italy

Received Day Month Year
Revised Day Month Year

In this paper we describe a method for clustering financial time series which is based on community analysis, a recently developed approach for partitioning the nodes of a network (graph). A network with N nodes is associated to the set of N time series. The weight of the link (i, j), which quantifies the similarity between the two corresponding time series, is defined according to a metric based on symbolic time series analysis, which has recently proved effective in the context of financial time series. Then, searching for network communities allows one to identify groups of nodes (and then time series) with strong similarity. A quantitative assessment of the significance of the obtained partition is also provided. The method is applied to two distinct case-studies concerning the U.S. and Italy stock exchange, respectively. In the U.S. case, the stability of the partitions over time is also thoroughly investigated. The results favorably compare with those obtained with the standard tools typically used for clustering financial time series, such as the minimal spanning tree and the hierarchical tree.

Keywords: Time series; clustering; network; communities.

PACS Nos.: 89.75.Hc, 89.65.Gh

1. Introduction

Clustering of time series is an active area of data mining research, with plenty of applications in almost all fields of science and technology[1,2]. Two of the most important ingredients of a time series clustering method are: (a) the definition

of a suitable similarity/distance measure, and (b) the clustering algorithm. When analyzing financial time series, the standard choices are (a) distances based on the Pearson correlation coefficient and (b) the use of aggregating tools such as the minimal spanning tree (MST) or the hierarchical tree (HT)[3,4].

In this paper, we adopt alternative approaches to (a) and (b) above. As far as the distance is concerned, we follow Brida and Risso[5,6] in using a metric based on symbolic dynamics, which they proved to be effective in the specific context of financial time series. Moreover, we propose a non conventional clustering algorithm which is in fact an application of a recent development of the theory of complex networks, namely *community analysis*[7,8]. Given a graph defined by suitable sets of nodes and links, the analysis is aimed at discovering subnetworks (the communities) characterized by a density of intra-community links which is much larger than the inter-community density. If, more in general, a weight is associated to each link, communities are such that the total weight of the intra-community links is significantly larger than the total inter-community weight.

Obviously, identifying communities is nothing but a way of clustering network nodes. It is not surprising, therefore, that community analysis can straightforwardly be transferred into a simple but effective method for clustering time series. Indeed, a network with N nodes can be associated to the set of N time series, with the weight of the link (i, j) quantifying the similarity between the two corresponding time series. Then, the community analysis identifies groups of nodes (i.e., time series) whose similarity is strong. Besides, the method provides a quantitative indicator (called *modularity*) of the significance of the obtained partition.

This paper, after recalling the required notions on networks and community analysis, presents the method in detail and illustrates its application to the analysis of two case-studies, namely the 30 time series of the Dow Jones Industrial Average index, and the 31 time series of the S&P/MIB Milan Stock Exchange index. They are exactly the two cases considered by Brida and Risso[5,6], and the results favorably compare to their work since community analysis, more directly than the MST and HT, evidences in a clear manner the clusters of time series having similar patterns, if any. At the same time, the modularity value allows one to check whether the resulting partition is truly significant. As a matter of fact, we will see that the community analysis evidences a significant partition in the U.S. case, whereas a much weaker diversification emerges in the Italian case.

With reference to the the U.S. case, we also investigate the stability of the obtained communities over time. As a matter of fact, several dramatic changes in real and financial markets occurred during the time span under analysis. Two different segmentation schemes are proposed: in the first one, the entire period under scrutiny is divided into three, consecutive subperiods corresponding to well distinct economic phases. In the second scheme, instead, a more sophisticated approach is adopted to assess to what extent the community structure is influenced by volatility and turbulence in financial markets. Somewhat surprisingly, all the results denote

a fundamental stability of the community partition: this can be interpreted as the footprint of deeper connections among the companies included in our sample, which are not significantly affected by stock market conditions.

2. Networks and Communities

In this section, we briefly introduce the minimal terminology on networks and communities needed in the paper. Detailed surveys can be found in Refs. 9, 10, 11.

We consider undirected networks composed of N nodes ($\mathbf{N} = \{1, 2, \ldots, N\}$ is the set of nodes) and L links, and denote by $A = [a_{ij}]$ the (symmetric) $N \times N$ *connectivity matrix*, where $a_{ij} = a_{ji} = 1$ if there exists the link $i \leftrightarrow j$, and $a_{ij} = 0$ otherwise. The *degree* k_i of node i, i.e., the number of links incident to i, is given by $k_i = \sum_j a_{ij} = \sum_j a_{ji}$, and the *degree sequence* is the list $\{k_1, k_2, \ldots, k_N\}$ of the node degrees. The network is *connected* if, for every pair (i, j) of distinct nodes, there exists a path from i to j; it is *complete* if any node i has a link to any other node j (thus $k_i = N - 1$ for all i).

Consider now a connected network. Roughly speaking, a subset $\mathbf{C}_h \subset \mathbf{N}$ is called a *community* if the density of links internal to $\mathbf{C}_h$ is much larger than the density of links connecting $\mathbf{C}_h$ to the rest of the network. A precise, quantitative formulation of this notion has been put forward by Newman and Girvan[7], and it relies on the notion of *modularity* Q. Given a network with nodes $\mathbf{N}$ and a partition $\mathbf{C}_1, \mathbf{C}_2, \ldots, \mathbf{C}_q$ (i.e., $\bigcup_h \mathbf{C}_h = \mathbf{N}$ and $\mathbf{C}_h \cap \mathbf{C}_k = \oslash$ for all h, k), the modularity is given by

$$Q = \frac{1}{2L} \sum_{h=1}^{q} \sum_{i,j \in \mathbf{C}_h} \left[a_{ij} - \frac{k_i k_j}{2L} \right]. \tag{1}$$

Q is obtained by summing up, through all the sets $\mathbf{C}_h$, the difference between the actual number of links internal to the set ($\frac{1}{2} \sum_{ij} a_{ij}$) and the value expected if links were created at random but preserving the node degrees (which can be proved to be $\frac{1}{2} \sum_{ij} \frac{k_i k_j}{2L}$). Thus Q is large (i.e., it tends to 1, due to the proper normalization) when the density of links internal to the sets $\mathbf{C}_h$ (the communities) is surprisingly large with respect to a random distribution of links in the network.

Analyzing the community structure of a network amounts, first of all, at finding the partition $\mathbf{C}_1, \mathbf{C}_2, \ldots, \mathbf{C}_q$ to which the largest modularity $Q = Q_{\max}$ is associated (note that the number q of communities is not *a priori* defined). At this point, a high value of $Q_{\max}$ denotes the existence of a true, significant community structure in the network, whereas a small value of $Q_{\max}$ means that the link distribution in the network is not far from being random. In order to assess whether the obtained value of $Q_{\max}$ is significantly high, one can consider the ensemble of networks having the same degree sequence as the original one, and extract a large number M of random networks from this ensemble (see Ref. 12 for a comparative analysis of generation methods). Then, the maximum modularity $\overline{Q}_i$, $i = 1, 2, \ldots, M$, is computed for each one of them. At this point, denoting by μ and σ the mean and standard deviation

of the $\overline{Q}_i$-s, a large value of the z-score

$$z = \frac{Q_{\max} - \mu}{\sigma} \tag{2}$$

indicates that the maximum modularity obtained for the original network is significantly high.

The notion of modularity can be extended to weighted networks[13]. If we denote by $w_{ij} > 0$ the weight of the link $i \leftrightarrow j$, then the modularity Q of a given partition is defined as

$$Q = \frac{1}{2w} \sum_{h=1}^{q} \sum_{i,j \in \mathbf{C}_h} \left[w_{ij} - \frac{w_i w_j}{2w} \right], \tag{3}$$

where $w_i = \sum_j w_{ij}$ is the *strength* of node i, i.e., the total weight of the links incident to i, and $w = \frac{1}{2} \sum_i w_i$ is the total weight of the links in the network. Equation (3) is the natural extension of (1) in the case of integer weights, since w_{ij} can be interpreted as the number of links connecting i to j (in this case the network is a *multigraph*). Otherwise, one can (approximately) transform the weights of the original network into integers by discretization, namely by measuring them with respect to a sufficiently small unit which can be, e.g., the minimum weight existing in the network[14].

After Ref. 7, the modularity approach to analyze network communities has been widely applied and proved to be very effective in capturing the structure of many real-world networks (see, e.g., Refs. 8, 15, 16, 17, 18, 19), including economic networks describing the connections among companies due to common directors and/or shareholding[20]. In parallel, a great effort has been devoted in devising efficient algorithms for finding the best network partition (i.e., $Q = Q_{\max}$). Since it has been proved that the exhaustive optimization of Q is a computationally hard problem[21], a large number of practical, sub-optimal methods have been proposed. We have used the aggregative, hierarchical method devised by Blondel et al.[22], which appears to outperform the others both in terms of $Q_{\max}$ (i.e., in the capability of finding a partition with higher modularity) and in computational requirements[23].

3. From Time Series to Networks

Suppose we have N time series each of length T, and denote the i-th time series by $\mathbf{X}_i = \{x_{i1}, x_{i2}, \ldots, x_{iT}\}$. Assume that a suitable distance $d_{ij} = d_{ji}$ between any pair (i, j) is defined. In the literature of time series clustering, many alternative definitions of distance are used, including Euclidean distance, correlation-based distance, and many others[1]. We will specify our choice in the next section.

We associate an N-node, undirected, weighted, complete network to the set of time series. To conform to standard network models, the weight $w_{ij} > 0$ will quantify the *similarity* between the time series i and j. Thus, the function f defining the distance-to-weight transformation $w_{ij} = f(d_{ij})$ will be a decreasing one, in order to assign a larger weight to a smaller distance, i.e., to a larger similarity.

In addition, our choice of f will be strongly nonlinear, in order to remarkably differentiate the most important links (i.e., the highly similar time series) from the others. As a matter of fact, while in sparse networks (i.e., few links per each node) communities typically emerge thanks to the topology only, this cannot happen in our complete network, unless weights are strongly differentiated. In our case studies, and with our distance definition (but the same happens with the other standard distances above recalled), the d_{ij}-s are roughly bell-shape-distributed in a quite narrow interval $[d_{min}, d_{max}]$. A strongly nonlinear transformation from distances to weights is thus needed. We note that such a weight-differentiation procedure is not unusual in the literature of graph clustering. For example, in the Markov Cluster (MCL) algorithm[24] ("one of the most used clustering algorithms in bioinformatics", according to Ref. 11) the weights are iteratively modified on the basis of an artificial Markov chain process, until the less important links are eliminated and the communities eventually emerge.

With this aims, we firstly denote by $\Pi(d) : [d_{min}, d_{max}] \rightarrow [0, 1]$ the (empirically derived) cumulative probability distribution of the distances d_{ij}, i.e.,

$$\Pi(d) = \Pr\{d_{ij} \leq d\} = \frac{\text{number of } d_{ij} \leq d}{N(N-1)}. \tag{4}$$

Then, the network is created by assigning the weight to the link (i, j) as follows:

$$w_{ij} = f(d_{ij}) = \begin{cases} 1, & \text{if } \Pi(d_{ij}) \leq 0.025; \\ 0.1, & \text{if } 0.025 < \Pi(d_{ij}) \leq 0.05; \\ 0.01, & \text{if } 0.05 < \Pi(d_{ij}) \leq 0.1; \\ 0.001, & \text{otherwise.} \end{cases} \tag{5}$$

According to (5), the largest weight (i.e., 1) is assigned to the smallest 2.5% distances; then, a considerably smaller weight (one tenth of the previous value) to the following 2.5% (i.e., up to the 5th percentile), and so on. Note that, as required, the transformation (5) yields a network with strong differentiation in the weights but, nonetheless, connected (in fact, the resulting graph is complete).

Needless to say, other transformations $w \rightarrow d$ could be devised. For example, one may try to use a transformation similar to (5) in spirit but more straightforward, e.g., the exponential function:

$$w_{ij} = f(d_{ij}) = 10^{-\gamma D_{ij}}, \; D_{ij} = \frac{d_{ij} - d_{min}}{d_{max} - d_{min}}, \; \gamma > 0. \tag{6}$$

Transformations (5) and (6) clearly resemble each other when $\gamma = 3$ (the resulting weights span exactly the same range $0.001 \leq w \leq 1$ when $d_{min} \leq d \leq d_{max}$). However, (5) is more refined as it takes into account the specific distribution of d_{ij} within the interval $[d_{min}, d_{max}]$. We will see, in the next section, that this allows one to obtain better results in the community analysis.

Having defined the weights, clustering the N time series simply amounts at executing the community analysis described in Sec. 2. Time series (i.e., nodes) will tend to cluster into communities with total intra-community weight significantly

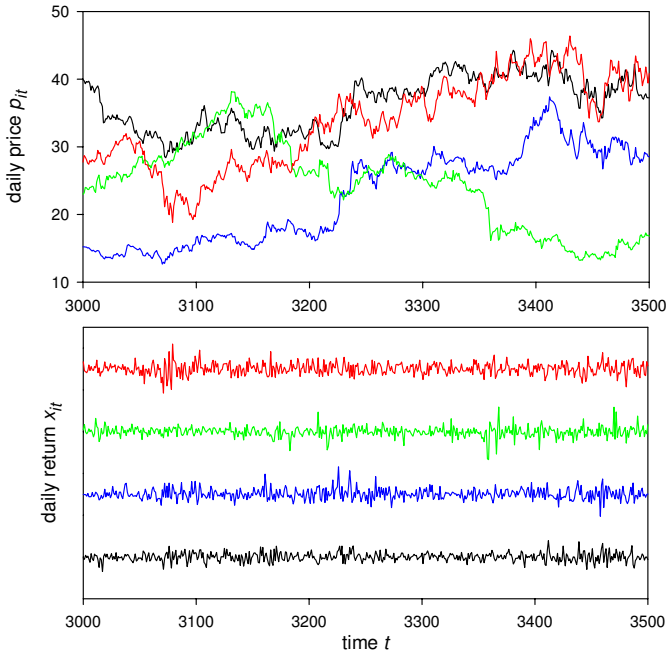

Fig. 1. Four of the time series of the DJIA companies in a 500-day interval. The daily returns x_{it} (below) are computed from the daily prices p_{it} (above). The time series of the daily returns (below) have been vertically shifted for readability.

larger than the inter-community weight, i.e., in such a way that a time series is significantly more similar to any other one inside the same community than to anyone outside it. The value of the optimal modularity $Q_{\max}$, together with the z-score, will inform whether the resulting partition is significant or, on the contrary, there is no meaningful differentiation among the time series.

4. Examples of Application: U.S. and Italian stock returns

We first analyze a set of $N = 30$ time series of the daily stock price data of the companies forming the Dow Jones Industrial Average (DJIA) index ($T = 5184$ days, from 10 July 1986 to 26 January 2007), which includes the most important companies of the New York Stock Exchange. From the sequence of daily prices p_{it} for the i-th stock, we derive the sequence of *daily returns*, defined as $x_{it} = \log(p_{it}/p_{i,t-1})$. Figure 1 displays a 500-day-long piece of the time series for four different companies. Clustering the daily return time series appears to be absolutely far from trivial.

We follow Brida and Risso[5,6] in adopting a metric based on time series symbolization, which is generally considered to be fairly robust in case of noisy time series[25,26] and, in the specific application context, suitable to better identify and separate different economic regimes[27]. In practice, for each return time

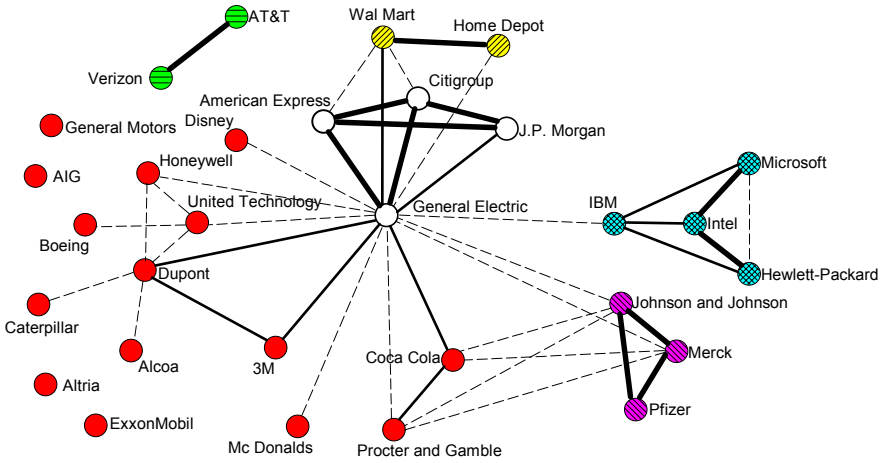

Fig. 2. The network derived from the 30 time series of daily returns of the DJIA companies. Links with $w_{ij} = 1$, 0.1, 0.01 are represented with decreasing thickness, whereas the remaining links ($w_{ij} = 0.001$) are not visualized. The filling patterns of the nodes (and colors online) denote the identified communities.

series $\mathbf{X}_i = \{x_{i1}, x_{i2}, \ldots, x_{iT}\}$, the cumulative probability distribution is first computed, i.e., $P(x) = \Pr\{x_{it} \leq x\}$. Then, a discretized (symbolic) time series $\mathbf{S}_i = \{s_{i1}, s_{i2}, \ldots, s_{iT}\}$ is defined by letting

$$s_{it} = \begin{cases} 1, \text{ if } P(x_{it}) \leq 1/3; \\ 2, \text{ if } 1/3 < P(x_{it}) \leq 2/3; \\ 3, \text{ otherwise.} \end{cases} \tag{7}$$

At this point, the distance d_{ij} between the time series i and j is defined as the Euclidean distance between the vectors $\mathbf{S}_i$ and $\mathbf{S}_j$, i.e., $d_{ij} = (\sum_t (s_{it} - s_{jt})^2)^{1/2}$. Finally, the network is derived by defining the weights as in (5). A graphic visualization of the resulting network is given in Fig. 2.

The analysis of this network reveals the existence of 6 communities, with a modularity $Q_{\max} = 0.68$. To analyze the statistical significance of such a large value, we compute the z-score on a sample of $M = 100$ randomized networks (see Sec. 2), obtaining a value as high as $z = 13.2$. Figure 3 compares the histogram of the maximum modularities $\bar{Q}_i$-s of the randomized networks, with the $Q_{\max}$ values of the network under study. Their clear separation (which reflects in the large z-score) testifies in favor of the significance of the obtained communities, i.e., a clear cluster structure exists in this set of time series. By looking inside the communities, it is immediately realized that the partition largely reflects the *a priori* classification of the industrial sectors the listed companies belong to. Indeed, five communities out of six are formed by companies belonging to well defined sectors: retailing (Home Depot, Wal Mart); telecommunications (AT&T, Verizon); pharmaceutical products (Johnson and Johnson, Merck, Pfizer); finance (American Express, Citigroup,

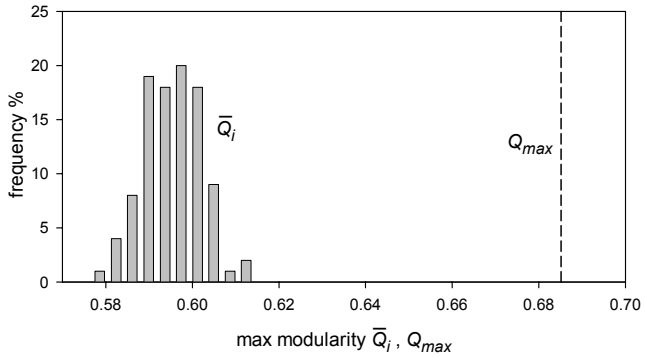

Fig. 3. Distribution of the maximum modularities $\bar{Q}_i$'s of the networks obtained by randomization of the DJIA network of Fig. 2. The dashed line evidences the $Q_{\max}$ value of the network under study.

J.P. Morgan, plus General Electric); computer hardware and software (IBM, Intel, Hewlett-Packard, Microsoft); whereas the sixth and last community is formed by the remaining companies, belonging to a miscellaneous of sectors. This validates the method, to some extent, and confirms that, in the U.S. stock market, the behavior of prices is basically clustered according to the industrial sector, as already pointed out by previous studies[3,6].

We repeat now the same analysis by defining the weights according to transformation (6) instead of (5). The network clustering turns out to be scarcely significant (i.e., low $Q_{\max}$ with low z-score) for γ up to $4-5$. For γ equal or larger than 6, on the contrary, the partition appears to be robust, but only 4 communities are identified: those related to telecommunication (AT&T, Verizon), pharmaceutical (Johnson and Johnson, Merck, Pfizer), and computer companies (IBM, Intel, Hewlett-Packard, Microsoft), plus a fourth one including all the rest. However, contrarily to above, the algorithm is not able to isolate the clusters of financial and retailing companies, that were instead clearly revealed by the use of weights (5) (see again Fig. 2). This result testifies in favor of the use of the $d \to w$ transformation (5).

The same procedure (with weights defined by (5)) is repeated on the $N = 31$ companies of the S&P/MIB index ($T = 1398$ days, from 6 December 2001 to 17 April 2007), defined on the Milano Stock Exchange, Italy. The resulting network is visualized in Fig. 4. Here, however, the community analysis obtains a modularity value as low as $Q_{\max} = 0.22$, with $z = 4.95$. This result denotes the lack of a strong community structure, i.e., the absence of a significant diversification in the time patterns of the stock prices. This is not unexpected, if one compares the networks of Figs. 2 and 4 even only visually. The former contains at least 5 groups of nodes with strong internal links, and comparably weaker connections to the rest of the network. No such structure can be identified in the latter, where connections are more uniformly distributed. This phenomenon can be considered as a consequence of

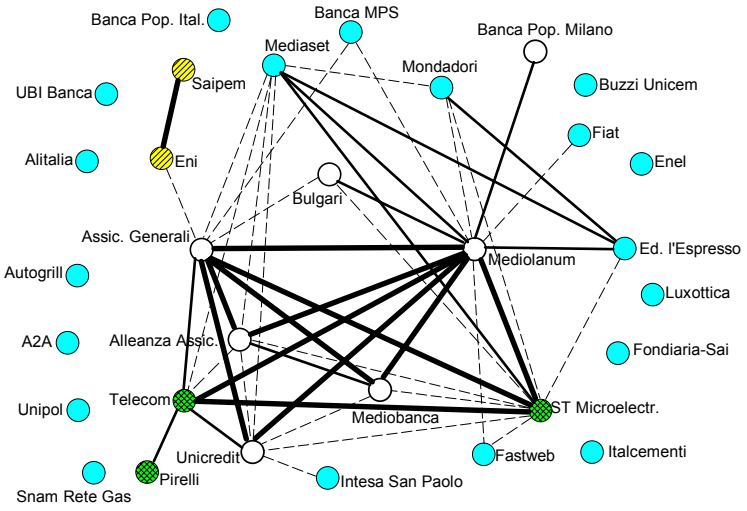

Fig. 4. The network derived from the 31 time series of daily returns of the S&P/MIB companies. Links with $w_{ij} = 1$, 0.1, 0.01 are represented with decreasing thickness, whereas the remaining links ($w_{ij} = 0.001$) are not visualized. The filling patterns of the nodes (and colors online) denote the identified communities.

the peculiar feature of the Italian stock market, where the ties among firms mainly derive from a common ownership (typically, a common holding company) rather than from the industrial sector[28].

5. Stability of communities

The results for the U.S. DJIA index reported in the previous section are obtained by pooling observations over a very long time period (10 July 1986-26 January 2007) during which several dramatic changes in real and financial markets occurred. It is then worth analyzing to what extent the community structure we find has been affected by these aggregation.

Specifically, we study two conceptually different segmentations of the sample. First, in Sec. 5.1 we study how the community structure differs across three economic periods: 1987-1994, 1995-2000 and 2001-2006[a]. Second, in Sec. 5.2 we assess to what extent the community structure is influenced by volatility and turbulence in financial markets.

For each segmentation of the sample we report the number of communities identified, the related modularity Q_{max}, the z-score computed according to the procedure illustrated in Sec. 2, and two measures of (dis)similarity with the original partition, i.e., two indexes quantifying to what extent the partition of the node set $\mathbf{N} = \{1, 2, \ldots, N\}$ obtained on the entire data period is similar to the partition

[a]In this part of the analysis we neglect the years 1986 and 2007 which are incomplete in the pooled sample.

obtained on a subset of data. The first indicator we use is the *Rand Index* R, which expresses similarity as the fraction of node pairs that are treated in the same way (i.e., put in the same community or in different ones) in both partitions[29]. R ranges from 0 (the two partitions are maximally different) to 1 (the two partitions are identical). The second indicator we use is the (normalized) *Variation of Information* V, which is an information-theory based distance first introduced by Meilă[30] and rapidly adopted by network scholars (e.g., Refs. 18, 11). V ranges from 0 to 1 too but, as it is a distance, 0 corresponds to the perfect identity between the two partitions which are compared, whereas 1 denotes maximal dissimilarity. We refer the reader to Ref. 30 for a critical review of these and other indexes of (dis)similarity between partitions.

5.1. *Evolution of community structure over time*

During the time period we analyze, at least three economic phases can be identified: 1987-1994, 1995-2000 and 2001-2006. First, between the mid 1980s and the mid 1990s a huge organizational shakeout occurred among the largest U.S. corporations. Since the 1950s large corporations had begun an unrelated growth process which resulted, in the early 1980s, in the widespread presence of large conglomerate firms[31] (several of these huge conglomerates are included in our sample, e.g., DuPont and General Electric), during the 1980s a dramatic deconglomeration process took place[32] and by the mid 1990s virtually all large conglomerates had restructured and refocused on their core business. We identify the decline of conglomerates to occur, in our sample, between 1987 and 1994.

Since 1995 a major technological shift drove the growth of the U.S. economy and its productivity, the so-called "IT revolution". Studies at firm-level[33], industry-level[34], and State-level[35], all confirm that the IT has been the engine in the recovery of the U.S. economy since the mid 1990s. This period was also characterized by an unprecedented surge of the NASDAQ stock exchange, which resulted in the (in)famous "dot.com bubble" which reached its peak in March 2000. Accordingly, the second period of our analysis, the IT revolution, is between 1995 and 2000.

The burst of the "dot.com bubble" was followed by a long period of monetary stabilization[36]. The economic recovery was gradual but solid and lasted until another asset bubble burst, giving rise to the subprime crisis in late 2007. This latter event is however outside our observation period; accordingly, our third period, the post-bubble period, is between 2001 and 2006.

We analyze the network using the procedure outlined in Sec. 2 and weights defined by equation (5), separately for each of the three subperiods. Results are reported in Table 1.

As a first observation, we note that results on 1987-1994 seems to be less clear-cut than those on the other two time periods. Modularity is far smaller in the first period ($Q_{max} = 0.367$) than in the second ($Q_{max} = 0.659$) and third ($Q_{max} = 0.679$); the number of communities singled out is smaller (4 against 5 and 7 respectively).

Table 1. Community analysis in different time periods: number of communities identified, maximum modularity, z-score, and Rand Index and Variation of Information with respect to the community structure obtained in the pooled sample.

	1987-1994	1995-2000	2001-2006
# of communities	4	5	7
Q_{max}	0.367	0.659	0.679
z-score	6.39	12.42	11.97
Rand Index	0.641	0.867	0.871
Variation of Information	0.387	0.148	0.170

Similarly, the Rand Index and the Variation of Information with respect to the community structure obtained on the pooled sample show that the first period is substantially (albeit not totally) different.

The economic interpretation of results in Table 1 is pretty interesting. First, the creation and burst of the IT bubble does not seem to have modified substantially the topology of the network. This is partly attributable to the nature of the companies we are considering: large and relatively mature industrial companies. Albeit our sample includes some IT-related companies (e.g., Microsoft), most sample companies operate in mature industries which were only marginally affected by the IT-revolution. Second, and more interesting, conglomerates may explain why the first period exhibits a weaker community structure. As we already noticed in Sec. 4, communities are strongly overlapped with industrial categories. Conglomerates are diversified across different industries and, accordingly, they do not easily fit in a community structure which is strongly industry-driven. Since the second half of the 1990s (i.e., in our second and third periods) the sectoral diversification of large corporation was dramatically reduced, which explains why community structure is more pronounced.

5.2. *Community structure, volatility, and turbulence*

Stock prices and, more generally, financial markets are well-known to follow non-stationary stochastic processes and, specifically, to go through persistent periods in which their returns are "perturbed". What we want to analyze in this section is whether the community structure that we find on the pooled sample is somewhat affected by these perturbations in the dynamics of the stock prices.

The most commonly adopted measure of the extent to which a market is perturbed is the level of return *volatility*, namely the standard deviation of the return distribution. A vast literature has shown that volatility changes over time and that it is persistent (i.e., a period of high volatility is likely to be followed by another period with high-volatility)[37]. One of the most prominent tools for characterizing changes in volatility is the Generalized Autoregressive Conditional Heteroskedasticity (GARCH) model[38]. To test whether volatility affects the community structure we estimate a standard GARCH(1,1) model on an equally weighted index including

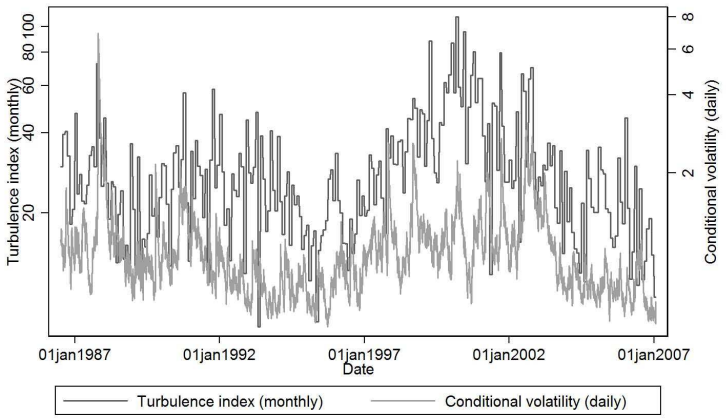

Fig. 5. Conditional volatility (expressed in %) computed using GARCH(1,1) model (8) on daily returns, and turbulence index computed using model (9) on monthly returns. Both y-axes are log-scaled.

stocks in our sample[b]. The daily return of the equally weighted index in day t is given by the scalar $x_t = \frac{1}{N}\sum_i x_{it}$, where x_{it} is the return of firm i in day t. In a GARCH(1,1) framework, this return is assumed to follow the stochastic process:

$$\begin{aligned} x_t &= m_t + \sigma_t \varepsilon_t, \\ \sigma_t^2 &= \omega + \alpha \varepsilon_{t-1}^2 + \beta \sigma_{t-1}^2, \end{aligned} \tag{8}$$

where $\varepsilon_t \sim N(0,1)$. Estimating parameters α and β in equation (8) gives us the daily conditional volatility of returns for each trading day in our sample. The estimated conditional volatility is depicted in Fig. 5 (light grey line). Consistently with expectations, days with the highest conditional volatility in our samples are clustered around well-known "perturbed" periods in the U.S. stock market (e.g., October 1987, September 2001). We split conditional returns in quartiles to identify periods with low, medium-low, medium-high and high volatility (respectively the first, second, third and fourth quartile of estimated conditional volatility). We analyze the network using the procedure outlined in Section 2 and weights defined by equation (5) separately for each of the four subsamples. Results are reported in Table 2.

Comparing the results obtained in subsamples with different levels of volatility, we find no evidence of any significant impact of volatility on community structure. Modularity is only slightly decreasing moving from the subsample with the lowest volatility ($Q_{max} = 0.722$) to the subsample with the highest volatility ($Q_{max} = 0.603$) and z-scores are all well beyond customary rejection levels. The similarity of the partitions with the one obtained using the pooled sample is also very high (the

[b]The GARCH(1,1) model, despite its relative simplicity, is found to be a good predictor of volatility[39].

Table 2. Community analysis at different volatility levels: number of communities identified, maximum modularity, z-score, and Rand Index and Variation of Information with respect to the community structure obtained in the pooled sample.

	Low volatility	Medium-low volatility	Medium-high volatility	High volatility
# of communities	6	5	5	5
Q_{max}	0.722	0.654	0.617	0.603
z-score	12.09	9.94	10.62	10.84
Rand Index	0.959	0.784	0.885	0.862
Variation of Information	0.059	0.244	0.120	0.153

Rand index ranging between 0.784 and 0.959) and shows no significant decrease as volatility increases. In a nutshell: community structure does not seem to be any different in periods of high volatility. This is a non-trivial result, given the well-known tendency of correlations among stock returns to be higher (in absolute value) in times of high volatility, partly as a consequence of sample selection and partly because of structural breaks in the underlying dynamics[40]. Apparently, the algorithm of community analysis on stock returns that we adopted is robust to these changes in correlations.

To further test the robustness of the community structure, we use another indicator of "perturbation" in stock returns: *turbulence*. This indicator is based on a simple measure of distance between a vector of observed returns and the average return vector, rather than on a conditional model. Despite its simplicity, this model is very effective in identifying outliers in returns distribution[41] and has significant applications in risk management[42]. We compute the turbulence index as follows. First, let $\mathbf{y}_\tau$ be the $1 \times N$ vector of stock returns in month τ,[c] $\bar{\mathbf{y}} = \langle \mathbf{y}_\tau \rangle$ be the $1 \times N$ vector of average stock returns $\mathbf{y}_\tau$ throughout the whole observation period, and $\mathbf{\Sigma}$ be the $N \times N$ sample covariance matrix of historical returns $\mathbf{y}_\tau$. Then the turbulence at month τ is computed as

$$d_\tau = (\mathbf{y}_\tau - \bar{\mathbf{y}})\, \mathbf{\Sigma}^{-1}\, (\mathbf{y}_\tau - \bar{\mathbf{y}})' . \tag{9}$$

Equation (9) measures the extent to which returns observed in month τ are anomalous with respect to their average distribution. While, clearly, the idea itself of "average" distribution is ill-posed if the stochastic process underlying stock prices is non-stationary, the benefit of the turbulence index in equation (9) is that it gives a simple and aggregate measure of deviations computed using a richer information set than what done with the GARCH(1,1) model used above, where only volatility (i.e., the diagonal of the covariance matrix) is considered. Turbulence and volatility, which somewhat capture the same construct (perturbation), are obviously closely related. The most turbulent months in our sample (March 2000, September 2001 and October 1987) also exhibit days with very high conditional volatility and, more

[c]Monthly returns are normally adopted in computing turbulence, rather than daily returns. Let Γ_τ be the set of days in a generic month τ, then $\mathbf{y}_\tau = \sum_{t \in \Gamma_\tau} \mathbf{x}_t$, where $\mathbf{x}_t = (x_{1t}, x_{2t}, \ldots, x_{Nt})$ is the $1 \times N$ vector of stock returns in day t.

Table 3. Community analysis at different turbulence levels: number of communities identified, maximum modularity, z-score, and Rand Index and Variation of Information with respect to the community structure obtained in the pooled sample.

	Low turbulence	Medium-low turbulence	Medium-high turbulence	High turbulence
# of communities	5	5	6	6
Q_{max}	0.692	0.609	0.701	0.712
z-score	13.71	10.55	13.84	14.11
Rand Index	0.807	0.844	0.880	0.959
Variation of Information	0.235	0.150	0.115	0.060

generally, the Pearson correlation between conditional volatility and the turbulence index (extended daily) is positive (0.384) and significant. The two measures however do not overlap entirely, as shown in Fig. 5. Again, we divide the sample according to quartiles of turbulence and analyze the network using the procedure outlined in Sec. 2 and weights defined by equation (5), separately for each of the subperiods. Results are reported in Table 3.

It is interesting to observe, in Table 3, that community structure is significant in each turbulence quartile (Q_{max} ranges between 0.609 and 0.712, with z-scores between 10.55 and 14.11) and, moreover, there is no symptom that it is weakened when markets are turbulent.

Summarizing, financial market conditions, expressed either in terms of conditional volatility or by turbulence, do not appear to affect significantly the community structure found using the algorithm illustrated in Sec. 2. Accordingly, we may interpret the partition we obtain as the result of a deeper connection among companies included in our sample, which is not significantly affected by stock market conditions.

6. Concluding Remarks

Clustering financial time series provides a fundamental aid for the comprehension of an economic system and, more specifically, is crucial in portfolio selection, namely in allocating wealth among alternative assets. In the literature of financial time series, the quantification of the similarity/distance between two time series has typically been based on the Pearson correlation coefficient, and aggregating tools such as the minimal spanning tree or the hierarchical tree have mainly been used[3,4]. In this paper we propose a rather different approach: following Refs. 5, 6, the distance is defined as the result of a symbolization of the time series, while the clustering method is an application of a recent development of network theory, that is community analysis[7,8].

The results favorably compare to those previously obtained on the same datasets[5,6]. Thanks to a suitable weighting scheme, strong and weak ties are clearly differentiated in the network. Then, community analysis is able to identify groups of time series with similar behavior, if any. The analysis of the U.S. case has shown, as already pointed out by previous analyses, that clusters reflect, to a large extent,

the industrial sectors. Furthermore, and somewhat surprisingly, the obtained partition appears to be stable over time, although real and financial markets went across very different economic phases during the time span under analysis. In the Italian case, on the other side, much less differentiation (i.e., clustering) emerges, as the connections among companies are more homogeneously distributed.

The proposed method can be applied, with no modifications, to datasets much larger than those analyzed in this paper. In fact, the current available algorithms for modularity optimization are able to effectively analyze networks with thousands of nodes (see, e.g., Ref. 22). Therefore, there seems to be no computational obstacles in analyzing the entire stock market of a given country, or even to consider cross-country sets of assets. In this respect, one of the possible extensions of the proposed method is that of considering algorithms for graph clustering alternative to modularity optimization[11], to check whether, in this specific context, they provide better results. This could possibly be associated to new weighting schemes, another aspect that is worth of deeper investigation.

References

1. T. W. Liao, Pattern Recognit. **38**, 1857 (2005).
2. H. Ding, G. Trajcevski, P. Scheuermann, X. Wang, and E. J. Keogh, in *Proc. 34th Int. Conf. on Very Large Data Bases (VLDB08)* (Auckland, New Zealand, 2008), pp. 1542–1552.
3. R. Mantegna, Eur. Phys. J. B **11**, 193 (1999).
4. R. N. Mantegna and H. E. Stanley, *An Introduction to Econophysics: Correlations and Complexity in Finance* (Cambridge University Press, 2000).
5. J. G. Brida and W. A. Risso, Int. J. Mod. Phys. C **18**, 1783 (2007).
6. J. G. Brida and W. A. Risso, Comput. Econ. **35**, 85 (2010).
7. M. E. J. Newman and M. Girvan, Phys. Rev. E **69**, 026113 (2004).
8. M. E. J. Newman, Proc. Natl. Acad. Sci. U.S.A. **103**, 8577 (2006).
9. M. E. J. Newman, SIAM Rev. **45**, 167 (2003).
10. S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D. H. Hwang, Phys. Rep. **424**, 175 (2006).
11. S. Fortunato, Phys. Rep. **486**, 75 (2010).
12. R. Milo, N. Kashtan, S. Itzkovitz, M. E. J. Newman, and U. Alon, On the uniform generation of random graphs with prescribed degree sequences, http://arxiv.org/abs/cond-mat/0312028, 2004.
13. M. E. J. Newman, Phys. Rev. E **70**, 056131 (2004).
14. V. Zlatic, G. Bianconi, A. Diaz-Guilera, D. Garlaschelli, F. Rao, and G. Caldarelli, Eur. Phys. J. B **67**, 271 (2009).
15. M. E. J. Newman, Phys. Rev. E **74**, 036104 (2006).
16. R. Guimera, M. Sales-Pardo, and L. A. N. Amaral, Phys. Rev. E **76**, 036102 (2007).
17. M. J. Barber, Phys. Rev. E **76**, 066102 (2007).
18. B. Karrer, E. Levina, and M. E. J. Newman, Phys. Rev. E **77**, 046119 (2008).
19. S. Fortunato and C. Castellano, in *Encyclopedia of Complexity and System Science*, edited by Meyers, R.A. (Springer-Verlag Berlin, 2009), pp. 1141–1163.
20. C. Piccardi, L. Calatroni, and F. Bertoni, Physica A **389**, 5247 (2010).
21. U. Brandes, D. Delling, M. Gaertler, R. Goerke, M. Hoefer, Z. Nikoloski, and D. Wagner, IEEE Trans. Knowl. Data Eng. **20**, 172 (2008).

22. V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, J. Stat. Mech.-Theory Exp. P10008 (2008).
23. A. Lancichinetti and S. Fortunato, Phys. Rev. E **80**, 056117 (2009).
24. S. Van Dongen, SIAM J. Matrix Anal. Appl. **30**, 121 (2008).
25. C. S. Daw, C. E. A. Finney, and E. R. Tracy, Rev. Sci. Instrum. **74**, 915 (2003).
26. C. Piccardi, Chaos **16**, 043115 (2006).
27. L. Molgedey and W. Ebeling, Eur. Phys. J. B **15**, 733 (2000).
28. M. Faccio and L. H. P. Lang, J. Financ. Econ. **65**, 365 (2002).
29. W. M. Rand, J. Am. Stat. Assoc. **66**, 846 (1971).
30. M. Meilă, J. Multivar. Anal. **98**, 873 (2007).
31. G. F. Davis, K. A. Diekmann, and C. H. Tinsley, Am. Sociol. Rev. **59**, 547 (1994).
32. S. Bhagat, A. Shleifer, and R. W. Vishny, in *Brookings Papers on Economic Activity: Microeconomics*, edited by M. N. Baily and C. Winston (Brookings Institution, 1990), pp. 1–84.
33. E. Brynjolfsson and L. M. Hitt, Review of Economics and Statistics **85**, 793 (2003).
34. K. J. Stiroh, Am. Econ. Rev. **92**, 1559 (2002).
35. F. Daveri and A. Mascotto, Rev. Income Wealth **52**, 569 (2006).
36. A. Greenspan, Am. Econ. Rev. **94**, 33 (2004).
37. T. Bollerslev, R. Y. Chou, and K. F. Krone, J. Econom. **52**, 5 (1992).
38. T. Bollerslev, J. Econom. **31**, 307 (1986).
39. P. R. Hansen and A. Lunde, Journal of Applied Econometrics **20**, 873 (2005).
40. B. H. Boyer, M. S. Gibson, and M. Loretan, International Finance Discussion Papers n. 597, Board of Governors of the Federal Reserve System (1999).
41. G. Chow, E. Jacquier, M. Kritzman, and K. Lowry, Financ. Anal. J. **1999**, 65 (1999).
42. M. Kritzman and Y. Li, Financ. Anal. J. **66**, 30 (2010).