Eur. Phys. J. B **11**, 193–197 (1999)

THE EUROPEAN PHYSICAL JOURNAL B

EDP Sciences

Hierarchical structure in financial markets

R.N. Mantegna[a]

Istituto Nazionale per la Fisica della Materia, Unità di Palermo, 90128, Palermo, Italy
Dipartimento di Energetica ed Applicazioni di Fisica, Università di Palermo, Viale delle Scienze, 90128, Palermo, Italy

Received 24 March 1999 and Received in final form 28 June 1999

Abstract. I find a hierarchical arrangement of stocks traded in a financial market by investigating the daily time series of the logarithm of stock price. The topological space is a subdominant ultrametric space associated with a graph connecting the stocks of the portfolio analyzed. The graph is obtained starting from the matrix of correlation coefficient computed between all pairs of stocks of the portfolio by considering the synchronous time evolution of the difference of the logarithm of daily stock price. The hierarchical tree of the subdominant ultrametric space associated with the graph provides a meaningful economic taxonomy.

PACS. 02.50.Sk Multivariate analysis – 89.90.+n Other areas of general interest to physicists

Financial markets are well-defined complex systems. They are studied by economists, mathematicians and, recently, also by physicists. The paradigm of mathematical finance is that the time series of stock returns are unpredictable [1]. Within this paradigm, time evolutions of stock returns are well described by random processes. A key point is if the random processes of stock returns time series of different stocks are uncorrelated or, conversely, if economic factors are present in financial markets and are driving several stocks at the same time. Ross introduced common economic factors in his arbitrage pricing theory model [2].

On the side of modeling of financial markets by using tools and procedures developed to model physical systems [3–11], there is the need to quantify a distance between different stocks traded in a financial markets. In the present analysis, I detect a hierarchical structure present in a portfolio of n stocks traded in a financial market. The observable which is used to detect the hierarchical arrangement of the stocks of a given portfolio is the synchronous correlation coefficient of the daily difference of logarithm of closure price of stocks. The correlation coefficient is computed between all the possible pairs of stocks present in the portfolio in a given time period. The goal of the present study is to obtain the taxonomy of a portfolio of stocks traded in a financial market by using the information of time series of stock prices only.

In this letter, I report results obtained by investigating the portfolio of the stocks used to compute the Dow Jones Industrial Average (DJIA) index and the portfolio of stocks used to compute the Standard and Poor's 500 (S&P 500) index in the time period from July 1989 to October 1995. Both indices describe the performance of the New York Stock Exchange. The starting point of my investigation is to quantify the degree of similarity between the synchronous time evolution of a pair of stock price by the correlation coefficient [12]

$$\rho_{ij} = \frac{\langle Y_i Y_j \rangle - \langle Y_i \rangle \langle Y_j \rangle}{\sqrt{(\langle Y_i^2 \rangle - \langle Y_i \rangle^2)(\langle Y_j^2 \rangle - \langle Y_j \rangle^2)}} \qquad (1)$$

where i and j are the numerical labels of stocks, $Y_i = \ln P_i(t) - \ln P_i(t-1)$ and $P_i(t)$ is the closure price of the stock i at the day t. The statistical average is a temporal average performed on all the trading days of the investigated time period.

For both portfolios, I determine the $n \times n$ matrix of correlation coefficients for daily logarithm price differences (which almost coincides with returns). By definition, ρ_{ij} can vary from -1 (completely anti-correlated pair of stocks) to 1 (completely correlated pair of stocks). When $\rho_{ij} = 0$ the two stocks are uncorrelated.

The matrix of correlation coefficients is a symmetric matrix with $\rho_{ii} = 1$ in the main diagonal. Hence, in each portfolio, $n\ (n-1)/2$ correlation coefficients characterize the matrix completely. An investigation of the statistical properties of the set of correlation coefficients is published elsewhere [13]. In this letter, I investigate the correlation coefficient matrix to detect the hierarchical organization present inside a portfolio of stocks traded in a stock market. In the search for an appropriate taxonomy of stocks of a given portfolio, I first look for a metric. The correlation coefficient of a pair of stocks cannot be used as a distance between the two stocks because it does not fulfill the three axioms that define a metric. However a metric can be defined using as distance a function of the correlation coefficient. An appropriate function is [14]

$$d(i,j) = \sqrt{2\ (1 - \rho_{ij})}. \qquad (2)$$

[a] e-mail: mantegna@ifata1.deaf.unipa.it

With this choice $d(i,j)$ fulfills the three axioms of a metric distance – (i) $d(i,j) = 0$ if and only if $i = j$; (ii) $d(i,j) = d(j,i)$ and (iii) $d(i,j) \leq d(i,k) + d(k,j)$ [15]. The first axiom is valid because $d(i,j) = 0$ if and only if the correlation is total ($\rho = 1$, namely only if the two stocks perform the same stochastic process). The second axiom is valid because the correlation coefficient matrix and hence the distance matrix $\mathbf{D}$ is symmetric by definition. The third axiom is valid because equation (2) is equivalent to the Euclidean distance between two vectors $\tilde{\mathbf{Y}}_i$ and $\tilde{\mathbf{Y}}_j$ which are obtained from the time series Y_i and Y_j by considering each record of the time series a component of the vector. The vector obtained must have a unitary norm, namely it is obtained by subtracting to each record the average value and by normalizing it to its standard deviation.

The distance matrix $\mathbf{D}$ is then used to determine the minimal spanning tree [16] connecting the n stocks of the portfolio. The method of constructing a MST linking a set of n objects is direct [17]. The MST of a set of n elements is a graph with $n-1$ links. Here I shortly describe this method by illustrating the determination of some links of the MST obtained for the DJIA portfolio of stocks in the investigated time period. In the following I will indicate each stock of the portfolio with its tick symbol. For example I indicate with CHV the Chevron Corp. and with XON the Exxon Corp. The correspondence between each tick symbol and the corresponding name of the company may be easily find in several web pages, one possibility is `www.forbes.com`. The MST associated with an Euclidean distance matrix can be obtained as follows. One first orders the nondiagonal elements of the distance matrix $\mathbf{D}$ in increasing order. For example, in the present case the shortest 10 distances observed for the portfolio of stocks of the DJIA index are CHV–TX $d = 0.949$, TX–XON $d = 0.962$, CHV–XON $d = 0.982$, KO–PG $d = 1.040$, GE–KO $d = 1.063$, AA–IP $d = 1.074$, GE–MMM $d = 1.078$, KO–MCD $d = 1.084$, GE–T $d = 1.090$ and DD–GE $d = 1.095$. The MST is progressively built up by linking all the elements of the set together in a graph characterized by a minimal distance between stocks. One starts with the pair of elements with the shortest distance. In our case CHV and TX ($d = 0.949$). At this stage the part already identified of the MST, which I address here as growing MST is just composed by these two elements. The next-smallest distance is between TX and XON ($d = 0.962$). Hence, at this step one can link XON at TX and modify the growing MST as CHV–TX–XON. The next pair of stocks in the ordered list of distances is CHV and XON, which need not to be inserted in the growing MST because both stocks have been already sorted to it. By continuing, one next has the KO and PG pair ($d = 1.040$). None of these stocks is present in the growing MST. Hence both must be sorted to it. At this step, the growing MST results to be composed by two distinct regions, which are CHV–TX–XON and KO–PG. The next pair of stocks comprises GE and KO ($d = 1.063$). The growing MST is then modified as CHV–TX–XON and GE–KO–PG. Then we have the pair of stocks AA–IP with distance $d = 1.074$. They are not in the growing MST and they need to be sorted to it. Hence the growing MST assumes the form CHV–TX–XON, GE–KO–PG and AA-IP. In other words, at this step three distinct groups of stocks are observed in the growing MST. By looking at the other distances in the list one notes that all the new elements (MMM,MCD, T and DD) need to be linked to the MST and in particular to the group of stocks GE–KO–PG. Specifically MMM, T and DD are to be linked to GE (which has at this stage 4 links) while MCD is to be linked to KO (which has at this stage 3 links). At a given stage a pair of stocks with both stocks already sorted to the growing MST but in different groups is detected in the ordered list of distances. For example in the present case at the distance $d = 1.110$ the pair of stocks composed by IP and MMM is detected. In the growing MST the two groups containing MMM and IP are then linked together through the connection MMM–IP. By following the above illustrated procedure for all the $n\,(n-1)/2$ distances one eventually obtain the final MST. In Figure 1a I show the complete MST obtained for the 30 stocks used to compute the DJIA in the investigated time period. The minimal spanning tree (MST) is attractive because provides an arrangement of stocks which selects the most relevant connections of each element of the set. Moreover the minimal spanning tree gives, in a direct way, the subdominant ultrametric [18] hierarchical organization of the points (stocks) of the investigated portfolio. The subdominant ultrametric can be obtained as follows. The knowledge of the MST allows us to determine the subdominant ultrametric distance matrix $\mathbf{D}^<$. This ultrametric matrix is obtained by defining the subdominant ultrametric distance $d^<(i,j)$ between i and j as the maximum value of any Euclidean distance $d(k,l)$ detected by moving in single steps from i to j through the shortest path connecting i and j in the MST. For example the ultrametric distance between XON and CHV is $d^< = 0.962$ because the maximum Euclidean two point distance detected by moving from XON to CHV in the MST is the Euclidean distance between XON and TX which is $d = 0.962$. It is worth noting that the Euclidean distance between XON and CHV is $d = 0.982$. Hence in the subdominant ultrametric space XON has the same ultrametric distance from TX and CHV (while this statement it is of course not true in the Euclidean space). This is shown in the hierarchical tree by linking XON to the CHV and TX branch at an ultrametric distance $d^<(i,j) = 0.962$. This group of three stocks is then linked to the other group of stocks in the hierarchical tree (right part of the tree in Fig. 1b) at an ultrametric distance $d^< = 1.131$ (which is determined by the link of the MST between DD and XON). The determination of the hierarchical tree of a subdominant ultrametric is then completely controlled by the ultrametric distance matrix [18]. It is worth pointing out that by using the detected subdominant ultrametric space it is possible to obtain a taxonomy of the investigated elements which is *uniquely* defined without any further assumption.

In the rest of this letter, I will show that the taxonomy found through the minimal spanning tree associated with the distance matrix $\mathbf{D}$ is of great interest from an

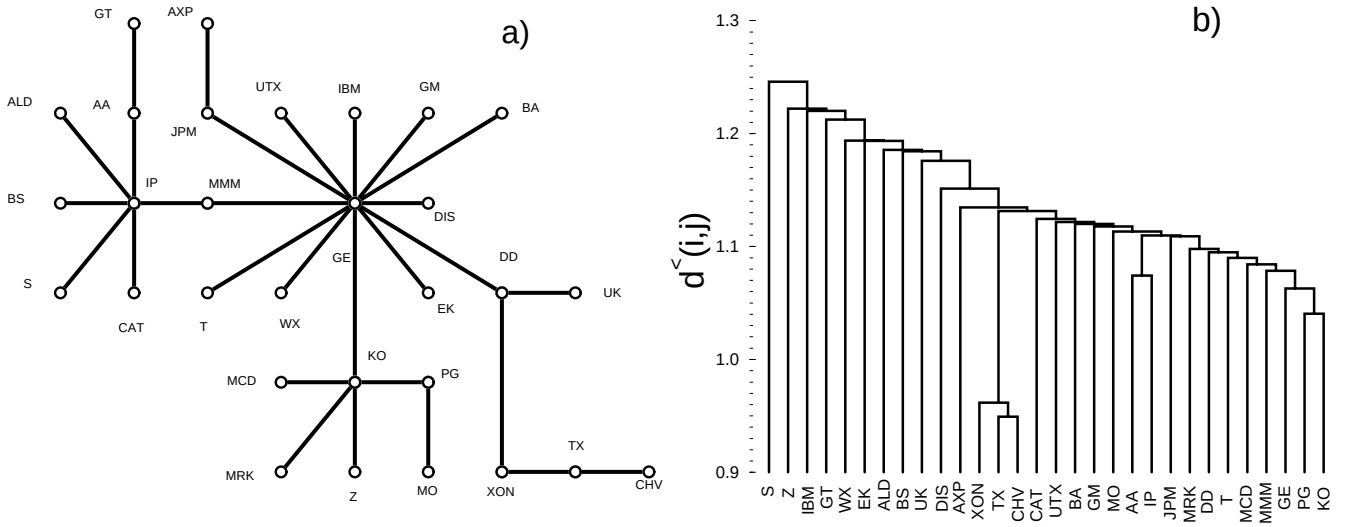

Fig. 1. (a) Minimal spanning tree connecting the 30 stocks used to compute the Dow Jones Industrial Average. The 30 stocks are labeled by their tick symbols. The distance between the stocks is bounded as: CHV-TX $0.90 < d(i,j) \leq 0.95$; XON–TX $0.95 < d(i,j) \leq 1.00$; KO-PG $1.00 < d(i,j) \leq 1.05$; MMM–GE–KO, DD–GE–T, AA–IP and MRK–KO–MCD $1.05 < d(i,j) \leq 1.10$; CAT–IP–MMM, AXP–JPM–GE–GM, BA–GE–UTX, DD–XON and MO–PG $1.10 < d(i,j) \leq 1.15$; DIS–GE–EK, DD–UK, BS–IP–ALD and GE–WX $1.15 < d(i,j) \leq 1.20$; AA–GT, GE–IBM, KO–Z and IP–S $1.20 < d(i,j) \leq 1.25$. (b) Hierarchical tree of the subdominant ultrametric space associated with the minimal spanning tree of a). In the hierarchical tree, several groups of stocks homogeneous with respect to the economic activities of the companies are detected: (i) oil companies (Exxon (XON), Texaco (TX) and Chevron (CHV)); (ii) raw material companies (Alcoa (AA) and International paper (IP)) and (iii) companies working in the sectors of consumer nondurable products (Procter & Gamble (PG)) and food and drinks (Coca Cola (KO)). The ultrametric distance at which individual stocks are branching from the tree is given by the y axis.

economic point of view. In particular, by assuming this kind of hierarchical organization, I am able to isolate groups of stocks, which make sense from an economic point of view by starting from the information carried by the time series of price only. The classification of the groups of stocks obtained with my analysis of the correlation coefficients is performed by using the industry and subindustry sectors reported in the Forbes 49th annual report on American industry. In Figure 1a, I show the minimal spanning tree for the DJIA portfolio of stocks. Each circle represents a stock, labeled by its tick symbol. Segments are linking connected stocks. The approximate distance between stocks is given in the figure caption. In Figure 1b the hierarchical tree of the subdominant ultrametric [18] associated to the MST is shown. An inspection of the MST and of the associated hierarchical tree shows the existence of three groups of stocks. The observed grouping has a direct economic explanation. The more evident and strongly connected group is the group of stocks CHV, TX and XON namely Chevron, Texaco and Exxon. These three companies are working in the same industry (energy) and in the same subindustry (international oils). AA and IP, namely Alcoa (working in the subindustry sector of nonferrous metals) and International Paper (working in the subindustry sector of paper and lumber) form a second group. Both companies provide raw materials. The third group involves companies which are in industry sectors which deals with consumer nondurables (Procter & Gamble, PG) and food drink and tobacco (Coca Cola, KO).

The same investigation is repeated for the set of stocks used to compute the S&P 500 index. In this case the larger size of the portfolio allows to perform a more refined test of the detected hierarchical structure of stocks. In my analysis, I considered only the companies which were present in the S&P 500 index for the entire period investigated. With this constrain the portfolio is composed of 443 stocks. Due to the size of the portfolio investigated, the obtained minimal spanning tree cannot be shown in a single figure in a legible way. As an illustrative example, I show a part of the MST in Figure 2. A group of financial services, capital goods, retailing, food drink & tobacco and consumer nondurables companies is observed in this strongly connected group of stocks. The portfolio of stocks used to compute the S&P 500 index is characterized by a hierarchical structure of stocks which is much more detailed than the one observed in the case of the DJIA portfolio. The structure of the minimal spanning tree of the portfolio of stocks of the S&P 500 index shows many groups of stocks which are homogeneous from an economic point of view. A detailed inspection of the hierarchical tree associated to the MST provides a large amount of economic information. It is impossible to put in a single legible figure the complete hierarchical tree of a so broad portfolio. In Figure 3, I then show only the branching of the tree up to the level of homogeneous groups. This means lines in the hierarchical tree shown in Figure 3 are always ending in a group of stocks which contains at least 2 stocks (but usually more). The branches of single stocks departing from the tree are not shown to make the figure readable. In the caption of Figure 3, I give details about industry sectors and/or subsectors of stocks belonging to the groups shown in the figure. With only a few exceptions the groups are homogeneous with respect to industry and often also subindustry

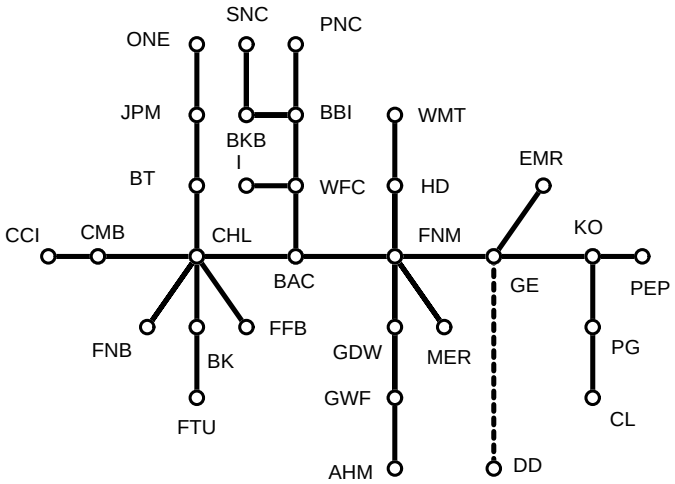

Fig. 2. A partial region of the minimal spanning tree of the portfolio of stocks used to compute the S&P 500 index. The figure shows a strongly connected large group of stocks observed for $d^<(i,j) < 1.10$. Circles represent stocks, which are labeled by their stock exchange tick symbols. In this region of the MST financial service companies (AHM, BAC, BBI, BK, BKB, BT, CCI, CHL, CMB, FFB, FNB, FNM, FTU, GDW, GWF, I, JPM, MER, ONE, PNC, SNC and WFC), capital goods companies (EMR and GE), retailing companies (HD and WMT), consumer nondurables companies (CL and PG) and food and drinks companies (KO and PEP) are found. Du Pont company (DD) is joining this group of stocks to the group of oil companies (not shown here).

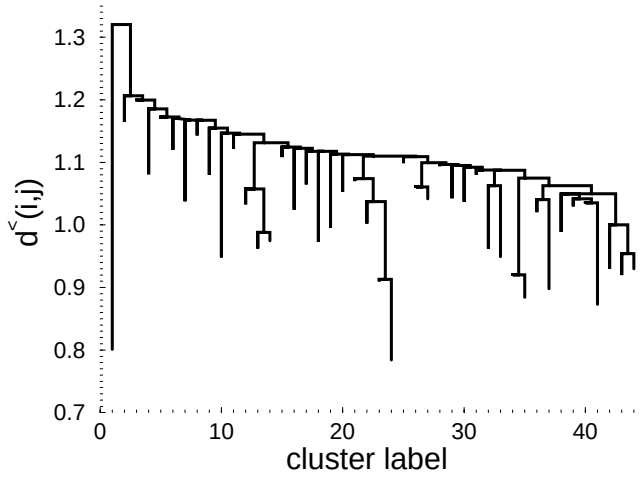

Fig. 3. Main structure of the hierarchical tree of the portfolio of stocks used to compute the S&P 500 index. Each line ending in the bottom corresponds to a group of stocks composed by at least two stocks. Lines are ending when the first bifurcation inside the group is observed. Individual stocks departing from the main tree are not shown for the sake of clarity. Groups are labeled with integers ranging from 1 to 44. The branching of each group from the main tree and inside the group are occurring at a distance given by the $d^<(i,j)$ scale. Below, I report for each group detected in the MST the observed common industry sector and, in parenthesis, the subindustry sector used in the 49th Forbes annual report of American industry (accessible on the web at the address `www.forbes.com`). 1. Metals (nonferrous metals, gold); 2. Construction (residential builders); 3. No common industry sector; 4. Travel and transport (trucking and shipping); 5. Consumer nondurables (photography and toys); 6. No common industry sector; 7. Metals (steel); 8. Consumer durables (automotive parts); 9. Travel and transport (airlines); 10. Entertainment and information (broadcasting and cable); 11. Financial services (lease and finance); 12. Energy (oilfield services); 13. Energy (international oils); 14. No common industry sector. 15. Capital goods (heavy equipment); 16. Business services and supplies (environmental and waste); 17. Construction (commercial builders); 18. Consumer durables (automobiles and trucks); 19. Food drink and tobacco (tobacco); 20. Entertainment and information (publishing); 21. Forest products and packaging (paper and lumber); 22. Metals (nonferrous materials); 23. Metals (nonferrous materials); 24. Metals (nonferrous materials); 25. Computer and communications (peripherals & equipment or software); 26. Electric utilities (regional area); 27. Computer and communications (telecommunications); 28. Retailing (department stores and drug & discount); 29. no common industry sector; 30. Travel and transport (railroads); 31. Food drink and tobacco (food processors); 32. no common industry sector; 33. Insurance (property & casualty and diversified); 34. Health (drugs); 35. Health (drugs); 36. Consumer nondurables (personal products); 37. Food drink and tobacco (beverages); 38. Retailing (no common subindustry sector (SS)); 39. Capital goods (electrical equipment); 40. Financial services (no common SS); 41. Financial services (thrift institutions); 42. Financial services (multinational banks); 43. Financial services (regional banks); 44. Financial services (multinational banks).

sectors suggesting that set of stocks working in the same industry and subindustry sectors respond, in a statistical way, to the same economic factors.

In some cases, my analysis, based on the statistical analysis of correlation coefficients between pairs of stock returns, refines the classification in sectors and subsectors used by Forbes. For example, ores, aluminum and copper are all classified metals as industry and nonferrous metals as subindustry. From my analysis, I detect that they respond to quite different economic factors. Specifically, ores companies are grouped in a cluster, which is the most distant from all the others groups of stocks of the tree, while aluminum and copper companies constitute a subgroup of the group containing raw materials companies.

The detection of a hierarchical structure in a broad portfolio of stocks traded in a financial market is consistent with the assumption that the time series of returns of a stock is affected by a number of economic factors. The analysis shows that the number and the relative influence of these factors is specific to each stock. In general, stocks or groups of stocks departing early from the tree (at high values of the distance $d^<(i,j)$) are mainly controlled by economic factors which are specific to the considered group (for example gold price for the stocks of the group 1 of the tree (see Fig. 3) which is composed only by companies involved in gold mining). When departure occurs

for (moderately) low values of $d^<$, the stocks are affected either by economic factors which are common to all stocks and by other economic factors which are specific to the considered set of stocks. The relative relevance of these factors is quantified by the length of the segment (or segments) observed for each group from one branching to the successive one.

In conclusion, the main aim of this paper is to announce a method of selecting a topological space (the subdominant ultrametric space) linking stocks traded in a financial market, which has associated a meaningful economic taxonomy. In fact, the present study shows that it is possible to determine a MST and the associated subdominant ultrametric hierarchical tree, starting from the distance matrix of equation (2). The detected hierarchical structure might be useful in the theoretical description of financial markets and in the search of economic factors affecting specific groups of stocks. The taxonomy associated with the obtained hierarchical structure is obtained by using information present in the time series of stock prices only. This result shows time series of stock prices are carrying valuable (and detectable) economic information.

I wish to thank Didier Sornette for kindly suggesting me the rigorous definition of distance given in equation (2). I wish also to acknowledge financial support from INFM and MURST.

References

1. P.A. Samuelson, Ind. Manag. Rev. **6**, 41 (1965); reproduced as Chapter 198 in P.A. Samuelson, *Collected Scientific Papers*, Vol. III (M.I.T. Press, Cambridge, 1972).
2. S.A. Ross, J. Econ. Theo. **13**, 341 (1976).
3. B.B. Mandelbrot, J. Business **36**, 394 (1963).
4. L.P. Kadanoff, Simulation **16**, 261 (1971).
5. R.N. Mantegna, Physica A **179**, 232 (1991).
6. P. Bak, K. Chen, J.A. Scheinkman, M. Woodford, Ric. Econ. **47**, 3 (1993).
7. J.-P. Bouchaud, D. Sornette, J. Phys. I France **4**, 863 (1994).
8. R.N. Mantegna, H.E. Stanley, Nature **376**, 46 (1995).
9. S. Ghashghaie, W. Breymann, J. Peinke, P. Talkner, Y. Dodge, Nature **381**, 767 (1996).
10. P. Bak, M. Paczuski, M. Shubik, Physica A **246**, 430 (1997).
11. A. Arnéodo, J.-F. Muzy, D. Sornette, Eur. Phys. J. B **2**, 277 (1998).
12. W. Feller, *An Introduction to Probability Theory and Its Applications* (Wiley, New York, 1971).
13. R.N. Mantegna, *Proc. of the ANDM 97 International Conference, San Diego*, 1997, edited by J.B. Kadtke, A. Bulsara (AIP, Woodbury, 1997), p. 197.
14. D. Sornette (private communication).
15. The introduction of a distance between a synchronous evolving pair of stocks was first proposed by R.N. Mantegna in `cond-mat/9802256` where a distance numerically verifying properties (i) to (iii) but theoretically incorrect was used. Even if the distance used was only approximately correct, the results obtained in that study essentially coincides with the results presented here with a rigorous definition of distance.
16. D.B. West, *Introduction to Graph Theory* (Prentice-Hall, Englewood Cliffs NJ, 1996).
17. C.H. Papadimitriou, K. Steiglitz, *Combinatorial Optimization* (Prentice-Hall, Englewood Cliffs, 1982).
18. R. Rammal, G. Toulouse, M.A. Virasoro, Rev. Mod. Phys. **58**, 765 (1986).