# Topology of correlation-based minimal spanning trees in real and model markets

Giovanni Bonanno,[1,2] Guido Caldarelli,[1,3] Fabrizio Lillo,[4] and Rosario N. Mantegna[2,4]
[1]*Unità INFM di Roma 1, piazzale Aldo Moro 2, I-00185 Roma, Italy*
[2]*Dipartimento di Fisica e Tecnologie Relative, Università di Palermo, viale delle Scienze, I-90128 Palermo, Italy*
[3]*Dipartimento di Fisica, Università "La Sapienza," piazzale Aldo Moro 2, I-00185 Roma, Italy*
[4]*Unità INFM di Palermo, viale delle Scienze, I-90128 Palermo, Italy*
(Received 6 May 2003; published 28 October 2003)

We compare the topological properties of the minimal spanning tree obtained from a large group of stocks traded at the New York Stock Exchange during a 12-year trading period with the one obtained from surrogated data simulated by using simple market models. We find that the empirical tree has features of a complex network that cannot be reproduced, even as a first approximation, by a random market model and by the widespread one-factor model.

DOI: 10.1103/PhysRevE.68.046130 PACS number(s): 89.75.Fb, 89.75.Hc, 89.65.Gh

The study of topological properties of networks has recently received much attention. In particular, it has been shown that many natural and social systems display an unexpected amount of correlation [1] and cannot therefore be described in terms of random graphs [2]. The topological properties of several graphs describing physical and social systems have been recently investigated. Examples are the WWW [3,7], Internet [4,5], social networks [6], and financial markets [11]. In the last cited case, the investigated graph is a spanning tree. Spanning trees are particular types of graphs. They connect all the vertices in a graph without forming any loops. Therefore, if the number of vertices [8] is $n$, one has $n-1$ arcs to connect them. There are several examples of spanning trees in nature and several observables have been proposed in order to classify them and study the possible optimization with respect to some external cost function [9].

In this paper we compare the topological properties of the minimal spanning tree (MST) of empirical data recorded at the New York Stock Exchange (NYSE) with MSTs obtained from simple models of the portfolio dynamics. Specifically we consider a model of uncorrelated Gaussian return time series and the widespread one-factor model. This last model is the starting point of the Capital Asset Pricing Model [10]. Our comparison shows that the random and the one-factor model fail to describe the topological properties of the MST extracted from a portfolio of stocks simultaneously traded in a financial market. Topological properties of this financial system can be therefore used to falsify widespread financial models.

The topological characterization of the correlation-based MST of real data was originally investigated in Ref. [11]. In their study, authors investigated a portfolio of $\approx 6000$ stocks by estimating the correlation coefficient on a yearly time period by using $\approx 250$ daily data. In the present study, we use a smaller number of stocks $N$ and a larger number of daily records $T$. Our choice is motivated by the request that the correlation matrix be positive definite. In fact, when the number of variables is larger than the number of time records the covariance matrix is only positive semidefinite [12].

The variable under investigation is the daily price return $r_i(t)$ of asset $i$ on day $t$. Given a portfolio composed of $N$ assets traded simultaneously in a time period of $T$ trading days, we extract the $N \times N$ correlation matrix. Each correlation coefficient $\rho_{i,j}$ can be associated with a metric distance $d_{i,j}$ between assets $i$ and $j$ through the relation $d_{i,j} = \sqrt{2(1-\rho_{i,j})}$ [13,14]. The distance matrix is then used to determine the MST connecting all the assets. The method of constructing the MST linking $N$ objects is known in multivariate analysis as the nearest neighbor single linkage cluster algorithm [12]. In a previous study three of us showed that the structure of the MST changes with the time horizon used to compute price returns [15].

The dataset used here consists of daily closure prices for 1071 stocks traded at the NYSE and continuously present in the 12-year period 1987–1998 (3030 trading days). The ratio $T/N \simeq 2.83$ is significantly larger than one and the correlation matrix positive definite. Figure 1 shows the MST of the real data. The color code is chosen by using the main industry sector of each firm according to the Standard Industrial Classification system and the correspondence is reported in the figure caption. Regions corresponding to different sectors are clearly seen. Examples are clusters of stocks belonging to the financial sector (purple); to the transportation, communications, electric gas, and sanitary services sector (green); and to the mining sector (red). The mining sector stocks are observed to belong to two subsectors, one containing oil companies (located on the right side of the figure) and the other containing gold companies (left side of the figure).

The empirical MST of real data can be compared with the results obtained from simple models of the simultaneous dynamics of a portfolio of assets. The simplest model assumes that the return time series is uncorrelated Gaussian time series, i.e., $r_i(t) = \epsilon_i(t)$, where $\epsilon_i(t)$ are Gaussian random variables with zero mean and unit variance. This type of model has been considered in Refs. [17,18] as a null hypothesis in the study of the spectral properties of the correlation matrix. In the cited references it has been shown that the spectrum of the real correlation matrix has a very large eigenvalue corresponding to the collective motion of the assets. A random model does not explain this empirical observation and therefore this fact clarifies why a better modeling of the portfolio dynamics is obtained by using the one-factor model. The

 

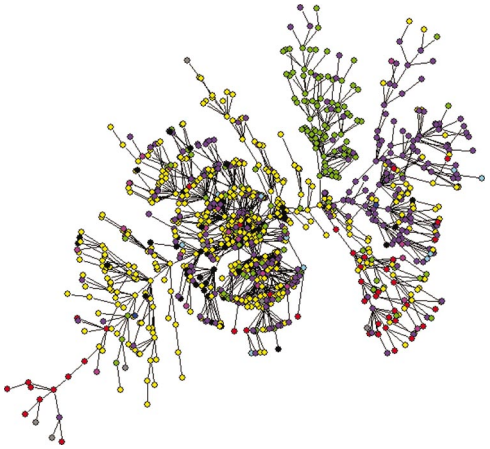

FIG. 1. (Color) Correlation-based minimal spanning tree of real data from daily stock returns of 1071 stocks for the 12-year period 1987–1998 (3030 trading days). The node color is based on Standard Industrial Classification system. The correspondence is red for mining, cyan for construction; yellow for manufacturing; green for transportation, communications, electric, gas, and sanitary services; magenta for wholesale trade; black for retail trade; purple for finance, insurance and real estate; orange for service industries; light blue for public administration.

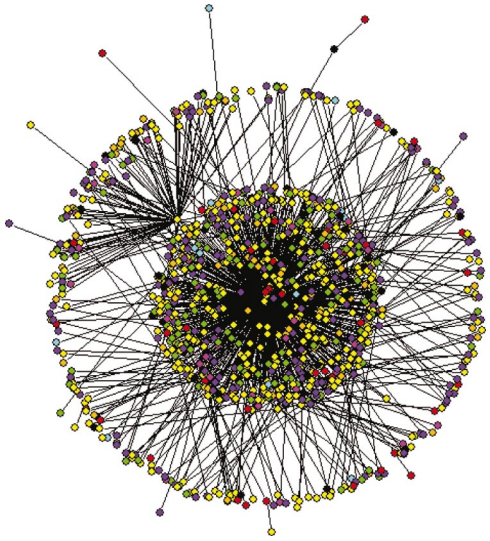

FIG. 2. (Color) Correlation-based minimal spanning tree of a numerical simulation of the one-factor model of Eq. (1). The color code is the same as used in Fig. 1.

one-factor model assumes that the return of assets is controlled by a single factor (or index). Specifically, for any asset $i$ we have

$$r_i(t) = \alpha_i + \beta_i r_M(t) + \epsilon_i(t), \tag{1}$$

where $r_i(t)$ and $r_M(t)$ are the returns of the asset $i$ and of the market factor at day $t$, respectively, $\alpha_i$ and $\beta_i$ are two real parameters, and $\epsilon_i(t)$ is a zero mean noise term characterized by a variance equal to $\sigma^2_{\epsilon_i}$. Our choice for the market factor is the Standard & Poor's 500 index and we assume that $\epsilon_i = \sigma_{\epsilon_i} w$, where $w$ is a random variable distributed according to a Gaussian distribution.

We estimate the model parameters for each asset from real time series with ordinary least squares method [10] and we use the estimated parameters to generate an artificial market according to Eq. (1). A consequence of this equation is that the variance (the squared volatility) of asset $i$ can be written as the sum of a term depending on the market factor and an idiosyncratic term. The fraction of variance explained by the factor $r_M$ is approximately described by an exponential distribution with a characteristic scale of about 0.16. The random model can be considered as the limit of the one-factor model when the fraction of variance explained by the factor goes to zero.

In the MST obtained with the random model, few nodes have a degree larger than few units. This implies that the MST is composed of long files of nodes. These files join at nodes of connectivity equal to few units. The MST obtained with the one-factor model is very different from the one obtained with the random model. In Fig. 2 we show the MST obtained in a typical realization of the one-factor model performed with the control parameters obtained as described above. The structure of this MST is also very different from the one obtained from real data. It is evident that the structure of sectors of Fig. 1 is not present in Fig. 2. In fact the MST of the one-factor model has a starlike structure with a central node. The largest fraction of node links directly to the central node and a smaller fraction is composed by the next-nearest neighbors. Very few nodes are found at a distance of three links from the central node. The central node corresponds to General Electric and the second most connected node is Coca Cola. It is worth noting that these two stocks are the two most highly connected nodes in the real MST also.

The MSTs obtained by simulating the models are different in each realization. However, a statistical characterization of MST is possible. In order to characterize quantitatively the structure of the MST we make use of two topological quantities. The first one is the distribution of the degree $k$. In random graph this quantity is distributed according to a binomial distribution which for large networks tends to a Poisson distribution. In many real networks it has been shown that the degree is distributed according to power law distribution signaling the presence of long range correlation. The second topological quantity is frequently used for oriented graphs. For any vertex $i$ in the tree we count the total number of vertices $a$ in the uphill subtree whose root is $i$. This quantity is called drainage basin area in oriented graphs of river networks [19], whereas it is usually referred as the in-degree component in graph theory. To calculate the in-degree component in a correlation-based MST, we orient the MST according to the number of steps each node is far from the most connected node (sink). When more than one sink is present in the MST a preferential one is randomly chosen among them.

We show in Fig. 3 the frequency distribution for the degree $k$ for the real data and for the average of over 100 realizations of the random model and of the one-factor model. The degree distribution for the MST of the real data is approximated by a power law behavior with exponent

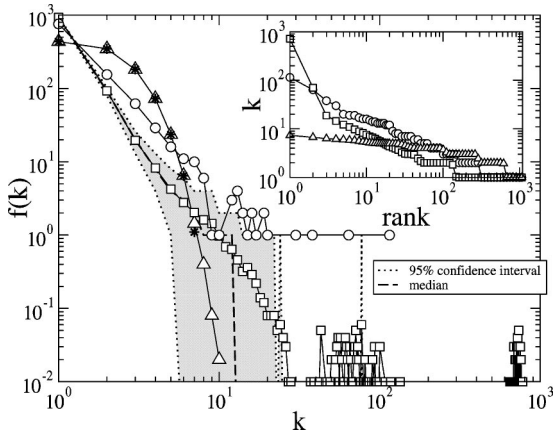


FIG. 3. Frequency distribution of the degree of the MST of real data (circle). We also show the mean degree distribution of random (triangle) and one-factor (square) model averaged over 100 numerical realizations of the MST. In the main panel, the gray region indicates the 95% confidence interval of the simulations of the one-factor model. The stars are the theoretical values of the degree frequency for the random model in mean field limit. The inset shows the corresponding rank plot of the degree in the three cases. Here, the 95% confidence interval of the simulations of the one-factor model is of the order of the size of the symbol used.

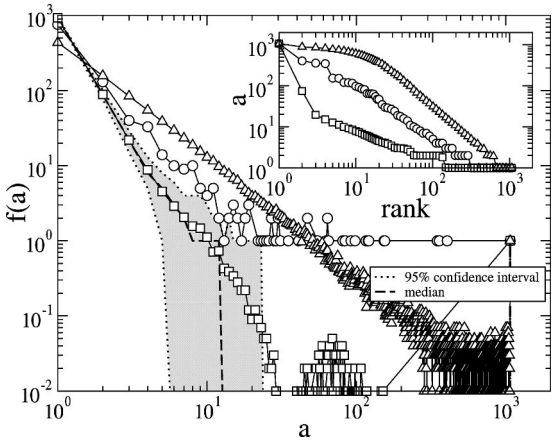


FIG. 4. Frequency distribution of the in-degree component of the MST of real data (circle). We also show the mean in-degree component distribution of random (triangle) and one-factor (square) model averaged over 100 numerical realizations of the MST. In the main panel, the gray region indicates the 95% confidence interval of the simulations of the one-factor model. The inset shows the corresponding rank plot of the in-degree component for the three cases. Here, the 95% confidence interval of the simulations of the one-factor model is of the order of the size of symbol used.

$-2.6$ for one decade followed by a set of isolated points with high degree. A power law behavior with a similar exponent has been observed in Ref. [11] and in another recent study [20]. The highest degree $k_{max}=115$ is observed for the General Electric, one of the most capitalized company in the NYSE. As we pointed out in a previous work [15], some important companies clearly emerge for its high degree value indicating that they act as a reference for other companies. The random model displays an approximately exponential decay of the degree distribution. The value of the maximum degree is small, $k_{max}=7.34\pm0.92$, showing that no asset plays a central role in the MST. The correlation-based MST of the random model can be considered as the MST of a set of $N$ points randomly distributed in a Euclidean space with $d=T$ dimension [16]. The $N$ points have independent identically Gaussian distributed coordinates $\mathbf{r_i}=(r_i(1),r_i(2),\ldots,r_i(T))$ with $i=1,2,\ldots,N$. It has been shown that the distribution of degree of the random MST in Euclidean space converges to a specific distribution in the mean field limit $d\rightarrow\infty$ [21]. The numerical values of the degree of frequency obtained from this mean field limit are shown as a star in Fig. 3 for $k=1,\ldots,7$. The agreement of theoretical values with the numerical simulations is very good showing that the mean field limit is already a good approximation for our $T$ parameter.

The MST obtained from the one-factor model is characterized by a rapidly decaying degree distribution and by an asset with a very high value of the degree. The value of the maximum degree is $k_{max}=718\pm29$. The corresponding asset is the center of the starlike structure of Fig. 2. The region with the highest value of the degree contains information about the stocks that act as reference for a large set of other stocks. To get more insight into the structure of this high $k$ region we show a rank plot of the degree both for the real data and for the considered models in the inset of Fig. 3. For the random model many nodes have a similar value of the degree which ranges to less than an order of magnitude. This is due to the fact that there is no hierarchy in the random model. The rank plot of the degree of the MST for the one-factor model is essentially different from the one observed in empirical data. We are able to prove that this result is statistically robust by showing in Fig. 3 the 95% confidence interval on the degree distribution computed starting from our numerical simulations of the one-factor model. Figure 3 shows unambiguously that there is a single highly connected node (the center) and a rapidly decaying degree as a function of the rank. This fact corresponds to the simple one-center hierarchy of the MST of the one-factor model. To better appreciate the relevance of this result, it is worth noting that the one-factor model is able to explain more that 80% of the correlation coefficients observed in real data. Nevertheless our results show that the topology of the MST of one-factor model is very different from the MST of real data. This apparent contradiction is due to the fact that the MST filters out the more relevant information about the correlation structure [14], whereas most of the correlation matrix is heavily dressed by noise, as shown in Refs. [17,18].

A discrepancy between real data and models is also observed in the frequency distribution of the in-degree component. Figure 4 shows the frequency distribution of the in-degree component for real and surrogate data. Again here we also show the 95% confidence interval for the one-factor model simulations. The inset of Fig. 4 shows the rank plot of the same data. In all three cases the in-degree component distribution has a power law shape. This is particularly clear for the MST of the random uncorrelated time series where the power law lasts for more than two decades with an exponent of $\approx-1.6$. It is known that for critical random trees the probability distribution of tree size decays as a power law

with an exponent 3/2 [22]. A critical random tree is a tree in which the mean number of sons of each node is one. In a MST the mean degree is exactly equal to $2n/(n-1) \simeq 2$. Hence when we orient the MST from the root to the leaves we have a tree with one son for each node. Our result shows that the in-degree component of the MST arising from random uncorrelated time series has properties similar to that of a critical random tree. This is not the case for the one-factor model where the power law has greater absolute slope due to the starlike structure of the tree. Neither model is actually able to catch the oriented structure of real data whose in-degree component distribution is in between the two models. The same arguments are also valid for the region of high values of $a$ as is evident from the rank plot in the inset.

In summary these results show that the topology of the MST for the real and for the considered artificial markets is different for node with both high and low degrees. If we define the importance of a node as its degree (or its in-degree component), from our analysis it emerges that the real market has a hierarchical distribution of importance of the nodes whereas the considered models are not able to catch such a hierarchical complexity. Specifically, in the random model the fluctuations select randomly few nodes and assign them small values of degree. Thus the MST of the random model is essentially nonhierarchical. On the other hand the MST of the one-factor model shows a simple one-center hierarchy. The MST of real market shows a more structured hierarchy of the importance of the stocks which is not captured by the considered models. The topology of stock return correlation-based MST shows large scale correlation properties characteristic of complex networks in the native as well as in an oriented form. Such properties cannot be reproduced at all, even as a first approximation, by simple models as a random model or the widespread one-factor model.

The authors acknowledge partial support from FET Open Project No. COSIN IST-2001-33555 and G.C. acknowledges European Commission Contract No. FMRXCT980183. F.L. and R.N.M. acknowledge partial support from INFM and MIUR.

[1] D.J. Watts and S.H. Strogatz, Nature (London) **393**, 440 (1998).
[2] P. Erdős and A. Rényi, Bull. Inst. Int. Stat. **38**, 343 (1961).
[3] R. Albert, H. Jeong, and A.L. Barabási, Nature (London) **401**, 130 (1999).
[4] G. Caldarelli, R. Marchetti, and L. Pietronero, Europhys. Lett. **52**, 386 (2000).
[5] R. Pastor-Satorras, A. Vazquez, and A. Vespignani, Phys. Rev. Lett. **87**, 258701 (2001).
[6] M.E.J. Newman, D.J. Watts, and S.H. Strogatz, Proc. Natl. Acad. Sci. U.S.A. **99**, 2566 (2002).
[7] A.L. Barabási and R. Albert, Science **286**, 509 (1999).
[8] G. Caldarelli, A. Capocci, P. De Los Rios, and M.A. Muñoz, Phys. Rev. Lett. **89**, 258702 (2002).
[9] A. Maritan, F. Colaiori, A. Flammini, M. Cieplak, and J. Banavar, Science **272**, 984 (1996).
[10] Y.J. Campbell, A.W. Lo, and A.C. Mackinlay, *The Econometrics of Financial Markets* (Princeton University Press, Princeton, 1997), and references therein.
[11] N. Vandewalle, F. Brisbois, and X. Tordoir, Quant. Finance **1**, 372 (2001).
[12] K.V. Mardia, J.T. Kent, and J.M. Bibby, *Multivariate Analysis* (Academic, San Diego, CA, 1979).
[13] J.C. Gower, Biometrika **53**, 325 (1966).
[14] R.N. Mantegna, Eur. Phys. J. B **11**, 193 (1999).
[15] G. Bonanno, F. Lillo, and R.N. Mantegna, Quant. Finance **1**, 96 (2001).
[16] R.N. Mantegna and H.E. Stanley, *An Introduction to Econophysics: Correlations and Complexity in Finance* (Cambridge University Press, Cambridge, 2000).
[17] L. Laloux, P. Cizeau, J.P. Bouchaud, and M. Potters, Phys. Rev. Lett. **83**, 1467 (1999).
[18] V. Plerou, P. Gopikrishnan, B. Rosenow, L.A. Nunes Amaral, and H.E. Stanley, Phys. Rev. Lett. **83**, 1471 (1999).
[19] I. Rodriguez-Iturbe and A. Rinaldo, *Fractal River Basins* (Cambridge University Press, Cambridge, 1997).
[20] J.-P. Onnela, A. Chakraborti, K. Kaski, J. Kertesz, and A. Kanto, Phys. Rev. E (to be published).
[21] M.D. Penrose, Ann. Appl. Probab. **24**, 1903 (1996).
[22] T.E. Harris, *The Theory of Branching Processes* (Dover, New York, 1989).