PHYSICAL REVIEW E **68**, 056110 (2003)

Dynamics of market correlations: Taxonomy and portfolio analysis

J.-P. Onnela, A. Chakraborti, and K. Kaski
Laboratory of Computational Engineering, Helsinki University of Technology, P.O. Box 9203, FIN-02015 HUT, Finland

J. Kertész
Department of Theoretical Physics, Budapest University of Technology & Economics, Budafoki út 8, H-1111 Budapest, Hungary and Laboratory of Computational Engineering, Helsinki University of Technology, P.O. Box 9203, FIN-02015 HUT, Finland

A. Kanto
Department of Quantitative Methods in Economics and Management Science, Helsinki School of Economics, P.O. Box 1210, FIN-00101 Helsinki, Finland

(Received 26 February 2003; published 13 November 2003)

The time dependence of the recently introduced minimum spanning tree description of correlations between stocks, called the ''asset tree'' has been studied in order to reflect the financial market taxonomy. The nodes of the tree are identified with stocks and the distance between them is a unique function of the corresponding element of the correlation matrix. By using the concept of a central vertex, chosen as the most strongly connected node of the tree, an important characteristic is defined by the mean occupation layer. During crashes, due to the strong global correlation in the market, the tree shrinks topologically, and this is shown by a low value of the mean occupation layer . The tree seems to have a scale-free structure where the scaling exponent of the degree distribution is different for ''business as usual'' and ''crash'' periods. The basic structure of the tree topology is very robust with respect to time. We also point out that the diversification aspect of portfolio optimization results in the fact that the assets of the classic Markowitz portfolio are always located on the outer leaves of the tree. Technical aspects such as the window size dependence of the investigated quantities are also discussed.

DOI: 10.1103/PhysRevE.68.056110 PACS number(s): 89.65.−s, 89.75.−k, 89.90.+n

I. INTRODUCTION

In spite of the traditional wisdom ''Money does not grow on trees,'' here we wish to show that the concept of trees (graphs) has potential applications in financial market analysis. This concept was recently introduced by Mantegna as a method for finding a hierarchical arrangement of stocks through studying the clustering of companies by using correlations of asset returns [1]. With an appropriate metric, based on the correlation matrix, a fully connected graph was defined in which the nodes are companies, or stocks, and the ''distances'' between them are obtained from the corresponding correlation coefficients. The minimum spanning tree (MST) was generated from the graph by selecting the most important correlations and it is used to identify clusters of companies.

In this paper, we study the time dependent properties of the minimum spanning tree and call it a ''dynamic asset tree.'' It should be mentioned that several attempts have been made to obtain clustering from the huge correlation matrix, such as the Potts superparamagnetic method [2], a method based on the maximum likelihood [3] or the comparison of the eigenvalues with those given by the random matrix theory [4]. We have chosen the MST because of its uniqueness and simplicity. The different methods are compared in Ref. [3].

Financial markets are often characterized as evolving complex systems [5]. The evolution is a reflection of the changing power structure in the market and it manifests the passing of different products and product generations, new technologies, management teams, alliances and partnerships, among many other factors. This is why exploring the asset tree *dynamics* can provide us new insights to the market. We believe that dynamic asset trees can be used to simplify this complexity in order to grasp the essence of the market without drowning in the abundance of information. We aim to derive intuitively understandable measures, which can be used to characterize the market taxonomy and its state. A further characterization of the asset tree is obtained by studying its degree distribution [6]. We will also study the robustness of tree topology and the consequences of the market events on its structure. The minimum spanning tree, as a strongly pruned representative of asset correlations, is found to be robust and descriptive of stock market events.

Furthermore, we aim to apply dynamic asset trees in the field of portfolio optimization. Many attempts have been made to solve this central problem from the classical approach of Markowitz [7] to more sophisticated treatments, including spin-glass-type studies [8]. In all the attempts to solve this problem, correlations between asset prices play a crucial role and one might, therefore, expect a connection between dynamic asset trees and the Markowitz portfolio optimization scheme. We demonstrate that although the topological structure of the tree changes with time, the companies of the minimum risk Markowitz portfolio are always located on the outer leaves of the tree. Consequently, asset trees in addition to their ability to form economically meaningful clusters, could potentially contribute to the portfolio optimization problem. Then with a lighter key one could perhaps say that ''some money may grow on trees,'' after all.

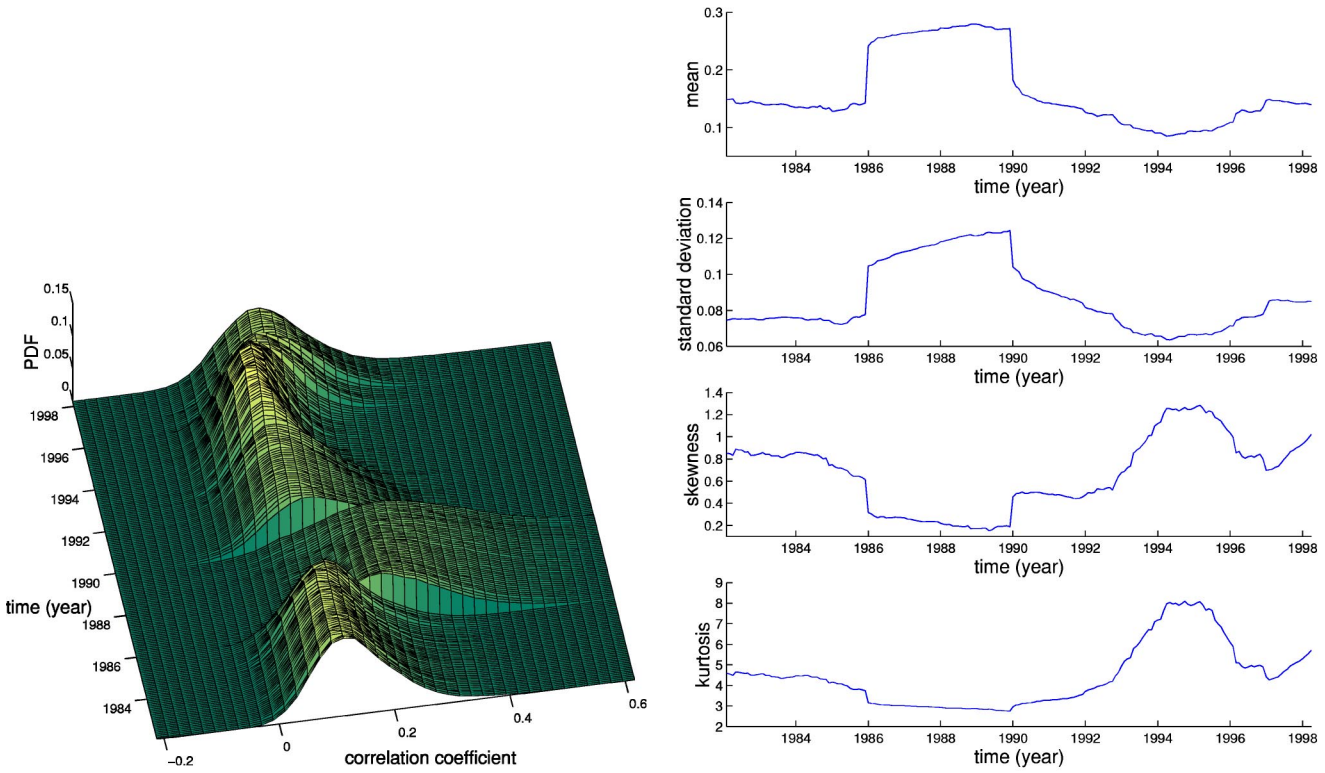

FIG. 1. Left: Plot of the probability density function of the correlation coefficients as a function of time. Right: The mean, standard deviation, skewness, and kurtosis of the correlation coefficients as functions of time.

II. RETURN CORRELATIONS AND DYNAMIC ASSET TREES

The financial market, for the largest part in this paper, refers to a set of data commercially available from the Center for Research in Security Prices (CRSP) of the University of Chicago Graduate School of Business. Here we will study the split-adjusted daily closure prices for a total of $N=477$ stocks traded at the New York Stock Exchange (NYSE) over the period of 20 years, from 02 Jan 1980 to 31 Dec 1999. This amounts to a total of 5056 price quotes per stock, indexed by time variable $\tau=1,2,\ldots,5056$. For analysis and smoothing purposes, the data are divided timewise into M *windows* $t=1,2,\ldots,M$ of width T corresponding to the number of daily returns included in the window. Several consecutive windows overlap with each other, the extent of which is dictated by the window step length parameter δT, describing the displacement of the window, measured also in trading days. The choice of window width is a trade off between too noisy and too smoothed data for small and large window widths, respectively. The results presented in this paper were calculated from monthly stepped four-year windows. Assuming 250 trading days a year, we used $\delta T \approx 20.8$ day and $T=1000$ day. We have explored a large scale of different values for both parameters, and the given values were found optimal [9]. With these choices, the overall number of windows is $M=195$.

In order to investigate correlations between stocks we first denote the closure price of stock i at time τ by $P_i(\tau)$ (Note that τ refers to a date, not a time window.) We focus our attention to the logarithmic return of stock i, given by $r_i(\tau)=\ln P_i(\tau)-\ln P_i(\tau-1)$ which, for a sequence of consecutive trading days, i.e., those encompassing the given window t, form the return vector $\boldsymbol{r}_i^t$. In order to characterize the synchronous time evolution of assets, we use the equal time correlation coefficients between assets i and j defined as

$$\rho_{ij}^t=\frac{\langle \boldsymbol{r}_i^t \boldsymbol{r}_j^t\rangle-\langle \boldsymbol{r}_i^t\rangle\langle \boldsymbol{r}_j^t\rangle}{\sqrt{[\langle \boldsymbol{r}_i^{t2}\rangle-\langle \boldsymbol{r}_i^t\rangle^2][\langle \boldsymbol{r}_j^{t2}\rangle-\langle \boldsymbol{r}_j^t\rangle^2]}}, \tag{1}$$

where $\langle\cdots\rangle$ indicates a time average over the consecutive trading days included in the return vectors. Due to Cauchy-Schwarz inequality, these correlation coefficients fulfill the condition $-1\leqslant\rho_{ij}\leqslant 1$ and form an $N\times N$ correlation matrix $\mathbf{C}^t$, which serves as the basis of dynamic asset trees to be discussed later.

Let us first characterize the correlation coefficient distribution (shown in Fig. 1), by its first four moments and their correlations with one another. The first moment is the *mean correlation coefficient* defined as

$$\overline{\rho}(t)=\frac{1}{N(N-1)/2}\sum_{\rho_{ij}^t\in \mathbf{C}^t}\rho_{ij}^t, \tag{2}$$

where we consider only the nondiagonal $(i\neq j)$ elements ρ_{ij}^t of the upper (or lower) triangular matrix. We also evaluate the higher order normalized moments for the correlation coefficients, so that the variance is

$$\lambda_2(t) = \frac{1}{N(N-1)/2} \sum_{(i,j)} (\rho_{ij}^t - \overline{\rho^t})^2, \tag{3}$$

the skewness is

$$\lambda_3(t) = \frac{1}{N(N-1)/2} \sum_{(i,j)} (\rho_{ij}^t - \overline{\rho^t})^3 / \lambda_2^{3/2}(t), \tag{4}$$

and the kurtosis is

$$\lambda_4(t) = \frac{1}{N(N-1)/2} \sum_{(i,j)} (\rho_{ij}^t - \overline{\rho^t})^4 / \lambda_2^2(t). \tag{5}$$

The mean, standard deviation (square root of the variance), skewness, and kurtosis of the correlation coefficients are plotted as functions of time in Fig. 1.

In this figure the effect and repercussions of Black Monday (October 19, 1987) are clearly visible in the behavior of all these quantities. For example, the mean correlation coefficient is clearly higher than average on the interval between 1986 and 1990. The length of this interval corresponds to the window width T, and Black Monday coincides with the midpoint of the interval [10]. The increased value of the mean correlation is in accordance with the observation by Drozdz *et al.* [11], who found that the maximum eigenvalue of the correlation matrix, which carries most of the correlations, is very large during market crashes. We also investigated whether these four different measures are correlated, as seems clear from the figure. For this we determined the Pearson's linear and Spearman's rank-order correlation coefficients, which between the mean and variance turned out to be 0.97 and 0.90, and between skewness and kurtosis 0.93 and 0.96, respectively. Thus the first two and the last two measures are very strongly correlated.

We now move on to construct an asset tree. For this we use the nonlinear transformation $d_{ij} = \sqrt{2(1-\rho_{ij})}$ to obtain distances with the property $2 \geqslant d_{ij} \geqslant 0$, forming an $N \times N$ distance matrix $\mathbf{D}^t$. At this point an additional hypothesis about the topology of the metric space is required. The working hypothesis is that a useful space for linking the stocks is an *ultrametric space*, i.e., a space where all distances are ultrametric. This hypothesis is motivated *a posteriori* by the finding that the associated taxonomy is meaningful from an economic point of view. The concept of ultrametricity is discussed in detail by Mantegna [1], while the economic meaningfulness of the emerging taxonomy is addressed later in this paper. Out of the several possible ultrametric spaces, the subdominant ultrametric is opted for due to its simplicity and remarkable properties. In practice, it is obtained by using the distance matrix $\mathbf{D}^t$ to determine the MST of the distances, according to the methodology of Ref. [1], denoted by $\mathbf{T}^t$. This is a simply connected graph that connects all N nodes of the graph with $N-1$ edges such that the sum of all edge weights, $\Sigma_{d_{ij}^t \in \mathbf{T}^t} d_{ij}^t$, is minimum. [Here time (window) dependence of the tree is emphasized by the addition of the superscript t to the notation.] Asset trees constructed for different time windows are not independent of each other, but form a series through time. Consequently, this multitude of trees is interpreted as a sequence of evolutionary steps of a single *dynamic asset tree*.

As a simple measure of the temporal state of the market (the asset tree) we define the *normalized tree length* as

$$L(t) = \frac{1}{N-1} \sum_{d_{ij}^t \in \mathbf{T}^t} d_{ij}^t, \tag{6}$$

where t again denotes the time at which the tree is constructed, and $N-1$ is the number of edges present in the MST. The probability distribution function of the $N-1$ distance elements d_{ij} in $\mathbf{T}^t$ as a function of time is plotted in Fig. 2 (cf. Ref. [12]). Also the mean, standard deviation, skewness, and kurtosis of normalized tree lengths are depicted in Fig. 2.

As expected and as the plots show, the mean correlation coefficient and the normalized tree length are very strongly anticorrelated. Pearson's linear correlation between the mean correlation coefficient $\overline{\rho}(t)$ and normalized tree length $L(t)$ is -0.98, and Spearman's rank-order correlation coefficient is -0.92, thus both indicating very strong anticorrelation. Anticorrelation is to be expected in view of how the distances d_{ij} are constructed from correlation coefficients ρ_{ij}. However, the extent of this anticorrelation is different for different input variables and is lower if, say, daily transaction volumes are studied instead of daily closure prices [13].

It should be noted that in constructing the minimum spanning tree, we are effectively reducing the information space from $N(N-1)/2$ separate correlation coefficients to $N-1$ tree edges, in other words, compressing the amount of information dramatically. This follows because the correlation matrix $\mathbf{C}^t$ and distance matrix $\mathbf{D}^t$ are both $N \times N$ dimensional, but due to their symmetry, both have $N(N-1)/2$ distinct upper (or lower) triangle elements, while the spanning tree has only $N-1$ edges. So, in moving from correlation or distance matrix to the asset tree $\mathbf{T}^t$, we have pruned the system from $N(N-1)/2$ to $N-1$ elements of information. If we compare Figs. 1 and 2, we find that distribution of the distance elements contained in the asset tree retain most of the features of the correlation coefficient distribution. Their corresponding moments also bear striking correlation/anticorrelation, e.g., the Pearson's linear correlation between the skewness of the correlation coefficients and the skewness of the edge lengths is -0.85, while the Spearman's rank order correlation is -0.82. Thus one may contemplate that the minimum spanning tree as a strongly reduced representative of the whole correlation matrix, bears the essential information about asset correlations.

As further evidence that the MST retains the salient features of the stock market, it is noted that the 1987 market crash can be quite accurately seen from Figs. 1 and 2. The fact that the market, during crash, is moving together is thus manifested in two ways. First, the ridge in the plot of the mean correlation coefficient in Fig. 1 indicates that the whole

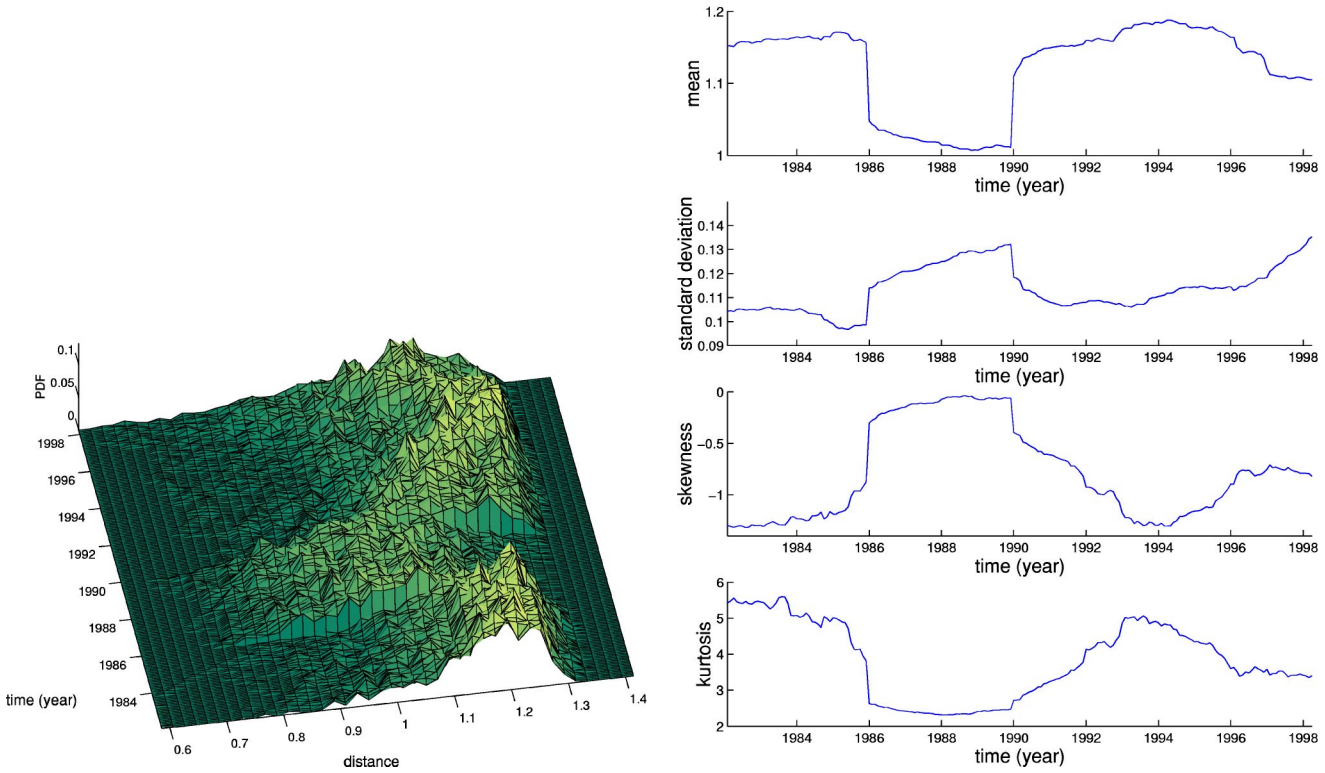

FIG. 2. Left: The probability distribution function of the $(N-1)$ distance elements contained in the asset tree, as a function of time. Right: The mean, standard deviation, skewness, and kurtosis of the normalized tree lengths as functions of time.

market is exceptionally strongly correlated. Second, the corresponding well in the plot of the mean normalized tree length in Fig. 2 shows how this is reflected in considerably shorter than average length of the tree so that the tree, on average, is very tightly packed. Upon letting the window width $T \to 0$, the two sides of the ridge converge to a single date, which coincides with Black Monday [10].

III. TREE OCCUPATION AND CENTRAL VERTEX

Next we focus on characterizing the spread of nodes on the tree. In order to do so, we introduce the quantity of *mean occupation layer* as

$$l(t,v_c) = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(v_i^t), \tag{7}$$

where $\mathcal{L}(v_i)$ denotes the level of vertex v_i. The levels, not to be confused with the distances d_{ij} between nodes, are measured in natural numbers in relation to the *central vertex* v_c, whose level is taken to be zero. Here the mean occupation layer indicates the layer on which the mass of the tree, on average, is conceived to be located.

Let us now examine the central vertex in more detail, as the understanding of the concept is a prerequisite for interpreting mean occupation layer results, to follow shortly. The central vertex is considered the parent of all other nodes in the tree, also known as the root of the tree. It is used as the reference point in the tree, against which the locations of all other nodes are relative. Thus all other nodes in the tree are children of the central vertex. Although there is arbitrariness in the choice of the central vertex, we propose that it is central, or important, in the sense that any change in its price strongly affects the course of events in the market on the whole. We propose three alternative definitions for the central vertex in our studies, all yielding similar and, in most cases, identical outcomes.

The first and second definitions of the central vertex are local in nature. The idea here is to find the node that is most strongly connected to its nearest neighbors. According to the first definition, this is the node with the highest *vertex degree*, i.e., the number of edges which are incident with (neighbor of) the vertex. The obtained results are shown in Fig. 3. The *vertex degree criterion* leads to General Electric (GE) dominating 67.2% of the time, followed by Merrill Lynch (MER) at 20.5%, and CBS at 8.2%. The combined share of these three vertices is 95.9%. The second definition, a modification of the first, defines the central vertex as the one with the highest sum of those correlation coefficients that are associated with the incident edges of the vertex. Therefore, whereas the first definition weighs each departing node equally, the second gives more weight to short edges, since a high value of ρ_{ij} corresponds to a low value of d_{ij}. This is reasonable, as short connections link the vertex more tightly to its neighborhood than long ones (the same principle employed in constructing the spanning tree). This *weighted vertex degree criterion* results in GE dominating 65.6% of the cases, followed by MER at 20.0%, and CBS at 8.7%, the share of the top three being 94.3%.

The third definition deals with the global quantity of *center of mass*. In considering a tree $\mathbf{T}^t$ at time t, the vertex v_i

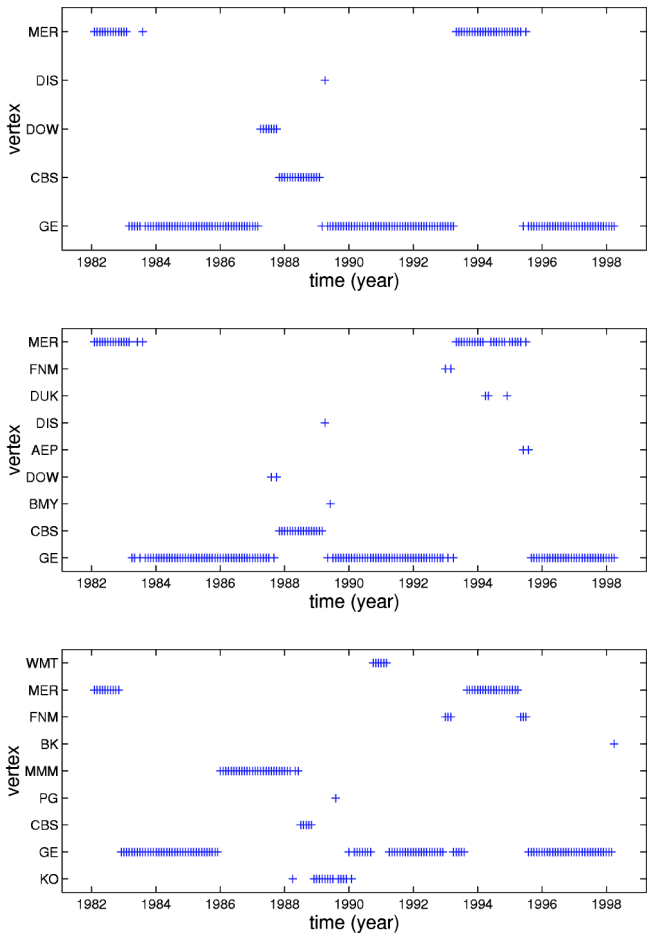

FIG. 3. Central vertices according to (1) vertex degree criterion (top), (2) weighted vertex degree criterion (middle), and (3) center of mass criterion (bottom).

that produces the lowest value for mean occupation layer $l(t,v_i)$ is the center of mass, given that all nodes are assigned an equal weight and consecutive layers (levels) are at equidistance from one another, in accordance with the above definition. With this *center of mass criterion* we find that the most dominant company, again, is GE, as it is 52.8% of the time the center of mass of the graph, followed by MER at 15.4%, and Minnesota Mining & MFG at 14.9%. These top three candidates constitute 83.1% of the total. Should the weight of the node be made proportional to the size (e.g., revenue, profit, etc.) of the company, it is obvious that GE's dominance would increase.

As Fig. 3 shows, the three alternative definitions for the central vertex lead to very similar results. The vertex degree and the weighted vertex degree criteria coincide 91.8% of the time. In addition, the former coincides with center of mass 66.7% and the latter 64.6% of the time, respectively. Overall, the three criteria yield the same central vertex in 63.6% of the cases, indicating considerable mutual agreement. The existence of a meaningful center in the tree is not a trivial issue, and neither is its coincidence with the center of mass. However, since the criteria applied, present a mixture of both local and global approaches, and the fact that they coincide almost 2/3 of the time, does indicate the existence of a well-

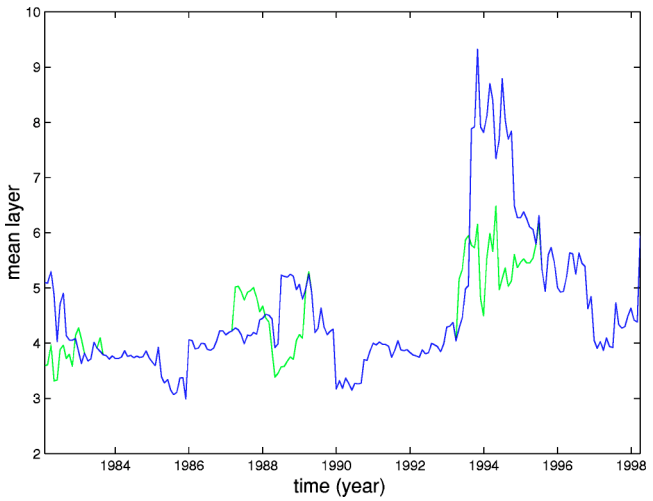

FIG. 4. Plot of mean occupation layer $l(t,v_c)$ as a function of time, with static (solid) and dynamic (dotted) central vertices.

defined center in the tree. The reason for the coincidence of the criteria seems clear, intuitively speaking. A vertex with a high vertex degree, the central vertex, in particular, carries a lot of weight around it (the neighboring nodes), which in turn may be highly connected to others (to their children), and so on. Two different interpretations may be given to these results. One may have either (i) static (fixed at all times) or (ii) dynamic (updated at each time step) central vertex. If the first approach is opted for, the above evidence well substantiates the use of GE as the central vertex. In the second approach, the results will vary somewhat depending on which of the three criteria are used in determining the central vertex.

The mean occupation layer $l(t)$ is depicted in Fig. 4, where also the effect of different central vertices is demonstrated. The solid curve results from the static central vertex, i.e., GE, and the dotted one to dynamic central vertex evaluated using the vertex degree criterion. The two curves coincide where only the solid curve is drawn. This is true most of the time, as the above central vertex considerations lead us to expect. The two dips at 1986 and 1990, located symmetrically at half a window width from Black Monday, correspond to the topological shrinking of the tree associated with the famous market crash of 1987 [10]. Roughly between 1993 and 1997, $l(t)$ reaches very high values, which is in concordance with our earlier results obtained for a different set of data [14]. High values of $l(t)$ are considered to reflect a finer market structure, whereas in the other extreme low dips are connected to market crashes, where the behavior of the system is very homogeneous. The finer structure may result from general steady growth in asset prices during that period as can be seen, for example, from the S&P 500 index.

IV. TREE CLUSTERS AND THEIR ECONOMIC MEANINGFULNESS

As mentioned earlier, Mantegna's idea of linking stocks in an ultrametric space was motivated *a posteriori* by the property of such a space to provide a meaningful economic tax-

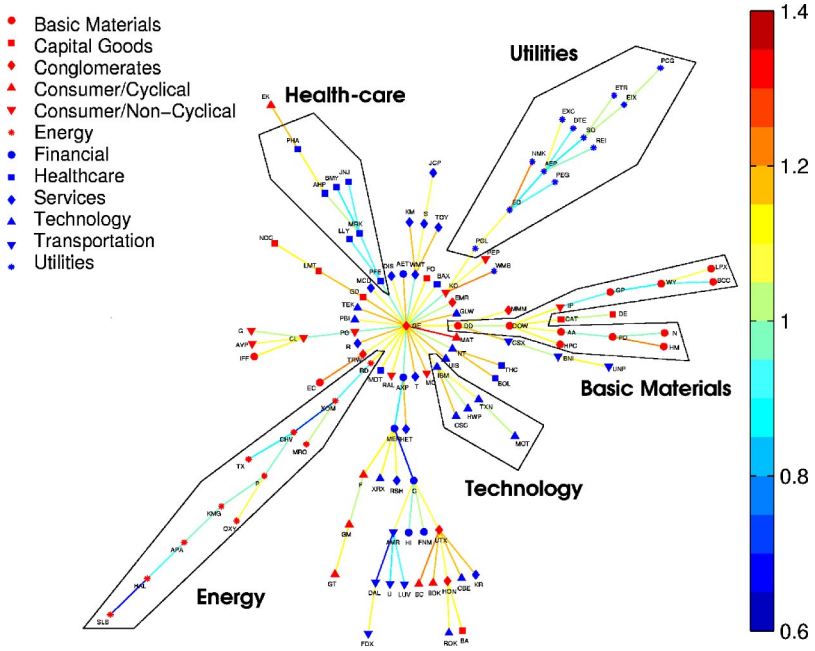

FIG. 5. (Color online) Snapshot of a dynamic asset tree connecting the examined 116 stocks of the S&P 500 index. The tree was produced using four-year window width and it is centered on January 1, 1998. Business sectors are indicated according to Ref. [15]. In this tree, General Electric (GE) was used as the central vertex and eight layers can be identified.

onomy. We will now explore this issue further, as the meaningfulness of the emerging economic taxonomy is the key justification for the use of the current methodology. In Ref. [1], Mantegna examined the meaningfulness of the taxonomy by comparing the grouping of stocks in the tree with a third party reference grouping of stocks by their industry, etc., classifications. In this case, the reference was provided by Forbes [15], which uses its own classification system, assigning each stock with a sector (higher level) and industry (lower level) category.

In order to visualize the grouping of stocks, we constructed a sample asset tree for a smaller dataset [14], shown in Fig. 5. This was obtained by studying our previous dataset [14], which consists of 116 S&P 500 stocks, extending from the beginning of 1982 to the end of 2000, resulting in a total of 4787 price quotes per stock [16].

Before evaluating the economic meaningfulness of grouping stocks, we wish to establish some terminology. We use the term sector exclusively to refer to the given third party classification system of stocks. The term *branch* refers to a subset of the tree, to all the nodes that share the specified common parent. In addition to the parent, we need to have a reference point to indicate the generational direction (i.e., who is who's parent) in order for a branch to be well defined. Without this reference there is no way to determine where one branch ends and the other begins. In our case, the reference is the central node. There are some branches in the tree, in which most of the stocks belong to just one sector, indicating that the branch is fairly homogeneous with respect to business sectors. This finding is in accordance with those of Mantegna [1], although there are branches that are fairly heterogeneous, such as the one extending directly downwards from the central vertex, see Fig. 5.

Since the grouping of stocks is not perfect at the branch level, we define a smaller subset whose members are more homogeneous as measured by the uniformity of their sector classifications. The term *cluster* is defined, broadly speaking, as a subset of a branch. Let us now examine some of the clusters that have been formed in the sample tree. We use the terms *complete* and *incomplete* to describe, in rather strict terms, the success of clustering. A complete cluster contains all the companies of the studied set belonging to the corresponding business sector, so that none are left outside the cluster. In practice, however, clusters are mostly incomplete, containing most, but not all, of the companies of the given business sector, and the rest are to be found somewhere else in the tree. Only the Energy cluster was found complete, but many others come very close, typically missing just one or two members of the cluster.

Building upon the normalized tree length concept, we can characterize the strength of clusters in a similar manner, as they are simply subsets of the tree. These clusters, whether complete or incomplete, are characterized by the *normalized cluster length*, defined for a cluster c as follows:

$$L_c(t) = \frac{1}{N_c} \sum_{d_{ij}^t \in c} d_{ij}^t , \tag{8}$$

where N_c is the number of stocks in the cluster. This can be compared with the normalized tree length, which for the sample tree in Fig. 5 at time t^* is $L(t^*) \approx 1.05$. A full account of the results is to be found in Ref. [16], but as a short summary of results we state the following. The Energy companies form the most tightly packed cluster resulting in $L_{\text{Energy}}(t^*) \approx 0.92$, followed by the Health-care cluster with $L_{\text{Health care}}(t^*) \approx 0.98$. For the Utilities cluster we have $L_{\text{Utilities}}(t^*) \approx 1.01$ and for the diverse Basic Materials cluster $L_{\text{Basic materials}}(t^*) \approx 1.03$. Even though the Technology cluster has the fewest number of members, its mean distance is the highest of the examined groups of clusters being $L_{\text{Technology}}(t^*) \approx 1.07$. Thus, most of the examined clusters seem to be more tightly packed than the tree on average.

One could find and examine several other clusters in the tree, but the ones that were identified are quite convincing. The minimum spanning tree, indeed, seems to provide a taxonomy that is well compatible with the sector classification provided by an outside institution, Forbes in this case. This is a strong vote for the use of the current methodology in stock market analysis. Some further analysis of the identified clusters is presented in Ref. [16].

There are, however, some observed deviations to the classification, which call for an explanation. For them the following points are raised.

(i) The seemingly random asset price fluctuations stem not only from standard economic factors, but also from psychological factors, introducing noise in the correlation matrix. Therefore, it is not reasonable to expect a one-to-one mapping between business sectors and MST clusters.

(ii) Business sector definitions are not unique, but vary by the organization issuing them. In this work, we used the classification system by Forbes [15], where the studied companies are divided into 12 business sectors and 51 industries. Forbes has its own classification principle, based on company dynamics rather than size alone. Alternatively, one could have used, say, the Global Industry Classification Standard (GICS), released on January 2, 2001, by Standard & Poor's [17]. Within this framework, companies are divided into 10 sectors, 23 industry groups, 59 industries, and 122 subindustries. Therefore, the classification system clearly makes a difference, and there are discrepancies even at the topmost level of business sectors amongst different systems.

(iii) Historical price time series is, by definition, old. Therefore, one should use contemporary definitions for business sectors, etc., as those most accurately characterize the company. Since these were not available to the authors, the current classification scheme by Forbes was used. The error caused by this approach varies for different companies.

(iv) In many classification systems, companies engaged in substantially different business activities are classified according to where the majority of revenues and profits comes from. For highly diversified companies, these classifications are more ambiguous and, therefore, less informative. As a consequence, classification of these types of companies should be viewed with some skepticism. This problem has its roots in the desire to categorize companies by a single label, and the approach fails where this division is unnatural.

(v) Some cluster outliers can be explained through the MST clustering mechanism, which is based on correlations between asset returns. Therefore, one would expect, for example, investment banks to be grouped with their investments rather than with other similar institutions. Through portfolio diversification, these banks distance themselves from the price fluctuations (risks) of a single-business sector. Consequently, it would be more surprising to find a totally homogeneous financial cluster than a fairly heterogeneous one currently observed.

(vi) The risks imposed on the companies by the external environment vary in their degree of uniformity from one business sector to another. For example, companies in the Energy sector (price of their stocks) are prone to fluctuations in the world market price of oil, whereas it is difficult to think of one factor having equal influence on, say, companies in the Consumer/Noncyclical business sector. This uniformity of external risks influences the stock price of these companies, in coarse terms, leading to their more complete clustering than that of companies facing less uniform external risks. In conclusion, regarding all the above listed factors, the success of the applied method in identifying market taxonomy is remarkable.

V. SCALE-FREE STRUCTURE OF THE ASSET TREE

So far we have characterized the asset tree as an important subgraph of the fully connected graph derived from all the elements of the correlation matrix. Since the asset tree is expected to reflect some aspects of the market and its state, it is therefore of interest to learn more about its structure. During the last few years, much attention has been devoted to the degree distribution of graphs. It has become clear that the so called scale-free graphs, where this distribution obeys a power law, are very frequent in many fields, ranging from human relationships through cell metabolism to the Internet [18,19]. Scale-free trees have also been extensively studied (see, e.g., Ref. [20]). Recently, examples for scale-free networks in economy and finance have been found [6,21,22].

Vandewalle *et al.* [6] found scale-free behavior for the asset tree in a limited (one year, 1999) time window for 6358 stocks traded at the NYSE, NASDAQ, and AMEX. They proposed the distribution of the vertex degrees $f(n)$ to follow a power law behavior:

$$f(n) \sim n^{-\alpha}, \tag{9}$$

with the exponent $\alpha \approx 2.2$, where n is the vertex degree (or number of neighbors of a node). This exponent implies that the second moment of the distribution would diverge in the infinite market limit, or in other words, the second moment of the distribution is always dominated by the rare but extremely highly connected vertices.

Our aim here is to study the property of scale freeness in the light of asset tree dynamics. First, we conclude that the asset tree has, most of the time, scale-free properties with a rather robust exponent $\alpha \approx 2.1 \pm 0.1$ for normal topology (i.e., outside crash periods of "business as usual"), a result close to that given in Ref. [6]. For most of the time the distribution behaves in a universal manner, meaning that the exponent α is a constant within the error limits. However, when the behavior of the market is not business as usual (i.e., within crash periods), the exponent also changes, although the scale-free character of the tree is still maintained. For the Black Monday period, we have $\alpha \approx 1.8 \pm 0.1$. This result is in full agreement with the observation of the shrinking of the tree during market crashes, which is accompanied by an increase in the degree, thus explaining the lower value of the exponent α. The observation concerning the change in the value of the exponent for normal and crash period is exemplified in Fig. 6.

When fitting the data, in many cases we found one or two outliers, i.e., vertices whose degrees did not fit to the overall power law behavior since they were much too high. In all cases these stocks corresponded either to the highest con-

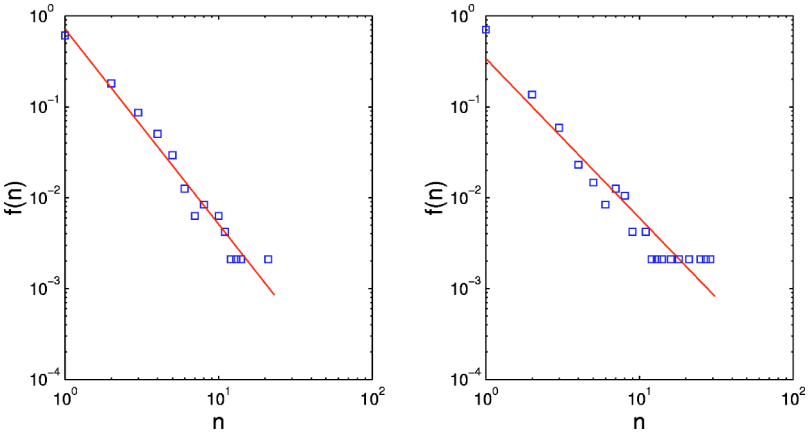

FIG. 6. Typical plots of vertex degree for normal (left) and crash topology (right), for which the exponents and goodness of fit are $\alpha \approx 2.15$, $R^2 \approx 0.96$ and $\alpha \approx 1.75$, $R^2 \approx 0.92$, respectively. The plot on the left is centered at 28.2.1994 and the right one at 1.5.1989, and for both $T = 1000$ days, i.e., 4 years.

nected node (i.e., the central vertex) or were nodes with very high degrees. This result suggests that it could be useful to handle these nodes with special care, thus providing further support to the concept of the central node. However, for the purpose of fitting the observed vertex degree data, such nodes were considered outliers. To give an overall measure of goodness of the fits, we calculated the R^2 *coefficient of determination*, which can be interpreted as the fraction of the total variation that is explained by the least-squares regression line. We obtained, on average, values of $R^2 \approx 0.86$ for the entire dataset with outliers included and $R^2 \approx 0.93$ with outliers excluded. Further, the fits for the normal market period were better than those obtained for the crash period as characterized by the average values of $R^2 \approx 0.89$ and $R^2 \approx 0.93$, respectively, with outliers excluded. In addition to the market period based dependence, the exponent α was also found to depend on the window width. We examined a range of values for the window width T between 2 and 8 yr and found, without excluding the outliers, the fitted exponent to depend linearly on T.

In conclusion, we have found the scaling exponent to depend on the market period, i.e., crash vs normal market circumstances and on the window width. These results also raise the question of whether it is reasonable to assume that different markets share the scaling exponent. In case they do not, one should be careful when pooling stocks together from different markets for the purpose of vertex degree analysis.

VI. ASSET TREE EVOLUTION

In order to investigate the robustness of asset tree topology, we define the *single-step survival ratio* of tree edges as the fraction of edges found common in two consecutive trees at times t and $t-1$ as

$$\sigma(t) = \frac{1}{N-1} |E(t) \cap E(t-1)|. \tag{10}$$

In this $E(t)$ refers to the set of edges of the tree at time t, $\cap$ is the intersection operator, and $|\cdots|$ gives the number of elements in the set. Under normal circumstances, the tree for two consecutive time steps should look very similar, at least for small values of window step length parameter δT. With this measure it is expected that while some of the differences can reflect real changes in the asset taxonomy, others may simply be due to noise. On letting $\delta T \to 0$, we find that $\sigma(t) \to 1$, indicating that the trees *are* stable in this limit [9].

A sample plot of single-step survival ratio for $T = 1000$ days and $\delta T \approx 20.8$ days is shown in Fig. 7. The following observations are made.

(i) A large majority of connections survives from one time window to the next.

(ii) The two prominent dips indicate a strong tree reconfiguration taking place, and they are window width T apart, positioned symmetrically around Black Monday, and thus imply topological reorganization of the tree during the market crash [10].

(iii) Single-step survival ratio $\sigma(t)$ increases as the window width T increases while δT is kept constant. Thus an increase in window width renders the trees more stable with respect to single-step survival of connections. We also find that the rate of change of the survival ratio decreases as the window width increases and, in the limit, as the window width is increased towards infinity $T \to \infty, \sigma(t) \to 1$ for all t. The survival ratio seems to decrease very rapidly once the window width is reduced below roughly 1 yr. As the window width is decreased further towards zero, in the limit as $T \to 0$, $\sigma(t) \to 0$ for all t.

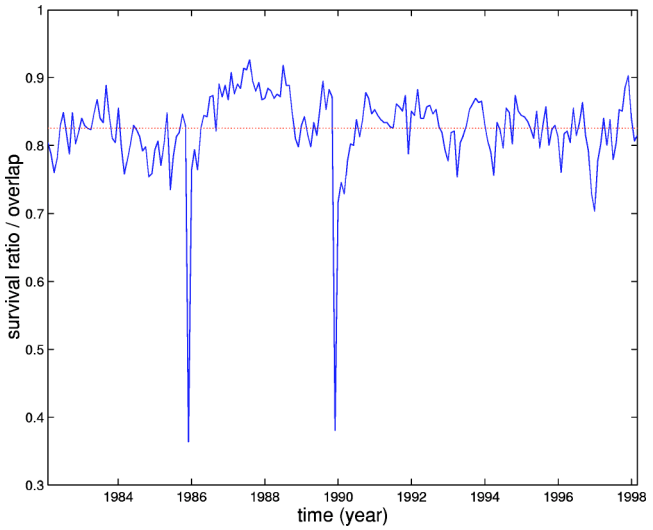

FIG. 7. Single-step survival ratio $\sigma(t)$ as a function of time. The average value is indicated by the horizontal line.

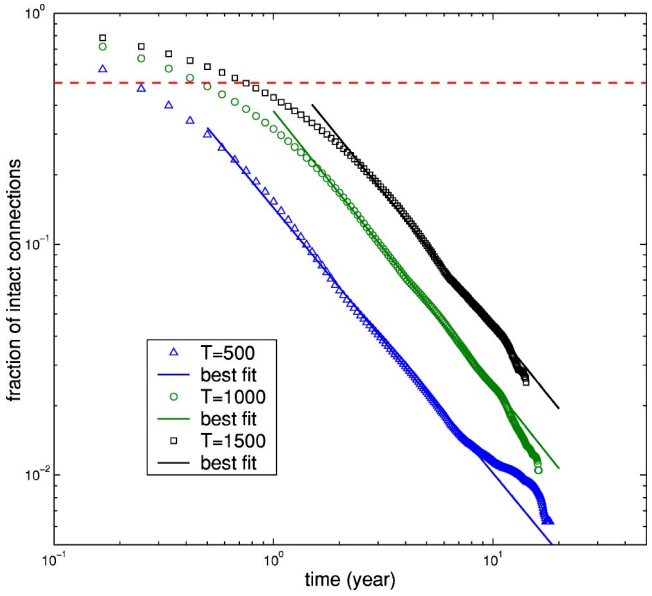

FIG. 8. Multistep survival ratio $\sigma(t,k)$ as a function of time for different parametric values of T (in days).

(iv) Variance of fluctuations around the mean is constant over time, except for the extreme events and the interim period, and it gets less as the window width increases.

In order to study the long term evolution of the trees, we introduce *the multistep survival ratio* at time t as

$$\sigma(t,k)=\frac{1}{N-1}\left|E(t)\cap E(t-1)\cdots E(t-k+1)\cap E(t-k)\right|, \tag{11}$$

where only those connections that have persisted for the whole time period without any interruptions are taken into account. According to this formula, when a bond between two companies breaks even once in k steps and then reappears, it is not counted as a survived connection. It is found that many connections in the asset trees evaporate quite rapidly in the early time horizon. However, this rate decreases significantly with time, and even after several years there are some connections that are left intact. This indicates that some companies remain closely bonded for times longer than a decade. The behavior of the multi-step survival ratio for three different values of window width (2, 4, and 6 yr) is shown in Fig. 8, together with the associated fits.

In this figure the horizontal axis can be divided into two regions. Within the first region, decaying of connections is faster than exponential, and takes place at different rates for different values of the window width. Later, within the second region, when most connections have decayed and only some 20%–30% remain (for the shown values of T), there is a crossover to power law behavior. The exponents obtained for the window widths of $T=500$, $T=1000$, and $T=1500$, in days, are -1.15, -1.19, and -1.17, respectively, and so remains the same within error margins. Thus, interestingly, the power law decay in the second region seems independent of the window width.

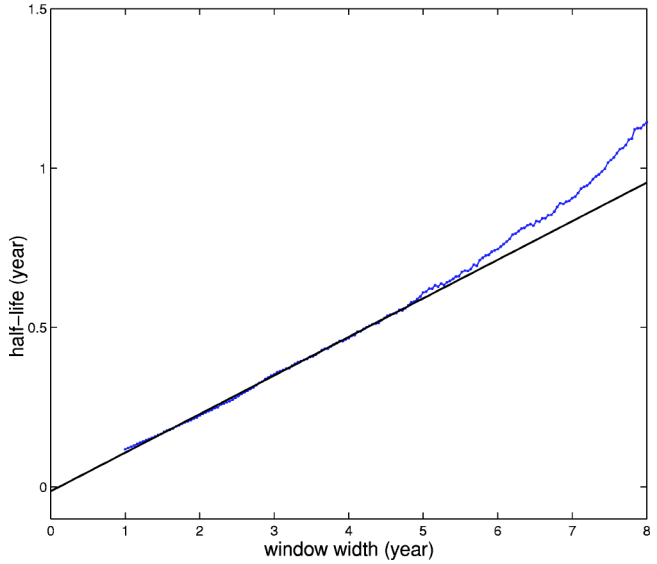

FIG. 9. Plot of half-life $t_{1/2}$ as a function of window width T.

We can also define a characteristic time, the so called half-life of the survival ratio $t_{1/2}$, or *tree half-life* for short, as the time interval in which half the number of initial connections have decayed, i.e., $\sigma(t,t_{1/2}/\delta T)=0.5$. The behavior of $t_{1/2}$ as a function of the window width is depicted in Fig. 9 and it is seen to follow a clean linear dependence for values of T being between 1 and 5 yr, after which it begins to grow faster than a linear function. For the linear region, the tree half-life exhibits $t_{1/2}\approx 0.12T$ dependence.

This can also be seen in Fig. 8, where the dashed horizontal line indicates the level at which half of the connections have decayed. For the studied values of the window width, tree half-life occurs within the first region of the multistep survival plot, where decaying was found to depend on the window width. Consequently, the dependence of half-life on window width T does not contradict the window width independent power law decaying of connections, as the two occur in different regions. In general, the number of stocks N, as well as the their type, is likely to affect the half-lives. Earlier, for a set of $N=116$ S&P 500 stocks, half-life was found to depend on the window width as $t_{1/2}\approx 0.20T$ [9]. A smaller tree, consisting primarily of important industry giants, would be expected to decay more slowly than the larger set of NYSE-traded stocks studied in this paper.

VII. PORTFOLIO ANALYSIS

Next, we apply the above discussed concepts and measures to the portfolio optimization problem, a basic problem of financial analysis. This is done in the hope that the asset tree could serve as another type of quantitative approach to and/or visualization aid of the highly interconnected market, thus acting as a tool supporting the decision making process. We consider a *general Markowitz portfolio* $\mathbf{P}(t)$ with the asset weights $w_1,w_2,\ldots,w_N$. In the classic Markowitz portfolio optimization scheme, financial assets are characterized by their average risk and return, where the risk associated with an asset is measured by the standard deviation of returns. The Markowitz optimization is usually carried out

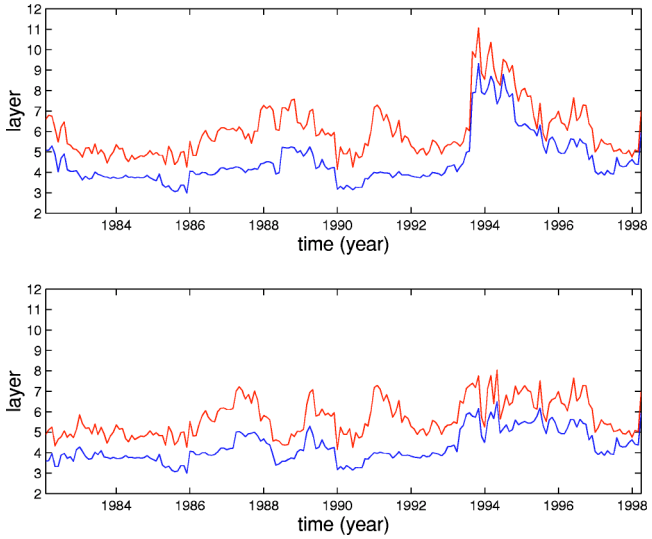

FIG. 10. Plot of the weighted minimum risk portfolio layer $l_{\mathbf{P}}(t,\theta=0)$ with no short selling (dotted) and mean occupation layer $l(t,v_c)$ (solid) against time. Top—static central vertex, bottom—dynamic central vertex according to the vertex degree criterion.

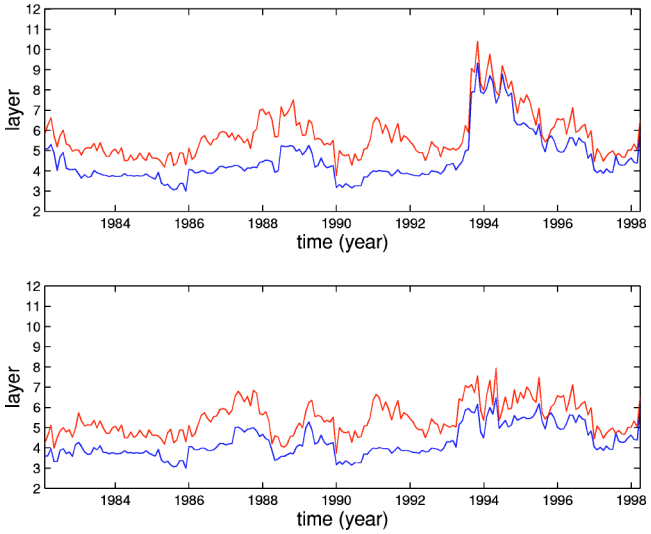

FIG. 11. Plot of the weighted minimum risk portfolio layer $l_{\mathbf{P}}(t,\theta=0)$ with short selling allowed (dotted) and mean occupation layer $l(t,v_c)$ (solid) against time. Top—static central vertex, bottom—dynamic central vertex according to the vertex degree criterion.

by using historical data. The aim is to optimize the asset weights so that the overall portfolio risk is minimized for a given portfolio return $r_{\mathbf{P}}$ [23]. In the dynamic asset tree framework, however, the task is to determine how the assets are located with respect to the central vertex.

Let r_m and r_M denote the returns of the minimum and maximum return portfolios, respectively. The expected portfolio return varies between these two extremes, and can be expressed as $r_{\mathbf{P},\theta}=(1-\theta)r_m+\theta r_M$, where θ is a fraction between 0 and 1. Hence, when $\theta=0$, we have the minimum risk portfolio, and when $\theta=1$, we have the maximum return (maximum risk) portfolio. The higher the value of θ, the higher the expected portfolio return $r_{\mathbf{P},\theta}$ and, consequently, the higher the risk the investor is willing to absorb. We define a single measure, the *weighted portfolio layer* as

$$l_{\mathbf{P}}(t,\theta)=\sum_{i\in\mathbf{P}(t,\theta)} w_i\mathcal{L}(v_i^t), \tag{12}$$

where $\Sigma_{i=1}^{N} w_i=1$ and further, as a starting point, the constraint $w_i\geqslant 0$ for all i, which is equivalent to assuming that there is no short selling. The purpose of this constraint is to prevent negative values for $l_{\mathbf{P}}(t)$, which would not have a meaningful interpretation in our framework of trees with central vertex. This restriction will shortly be discussed further.

Figure 10 shows the behavior of the mean occupation layer $l(t)$ and the weighted minimum risk portfolio layer $l_{\mathbf{P}}(t,\theta=0)$. We find that the portfolio layer is higher than the mean layer at all times. The difference between the layers depends on the window width, here set at $T=1000$, and the type of central vertex used. The upper plot in Fig. 10 is produced using the static central vertex (GE), and the difference in layers is found to be 1.47. The lower one is produced by using a dynamic central vertex, selected with the vertex degree criterion, in which case the difference of 1.39 is found.

Above we assumed the no short-selling condition. However, it turns out that, in practice, the weighted portfolio layer never assumes negative values and the short-selling condition, in fact, is not necessary. Fig. 11 repeats the earlier plot, this time allowing for short selling. The weighted portfolio layer is now 99.5% of the time higher than the mean occupation layer and, with the same central vertex configuration as before, the difference between the two is 1.18 and 1.14 in the upper and lower plots, respectively. Thus we conclude that only minor differences are observed in the previous plots between banning and allowing short selling, although the difference between weighted portfolio layer and mean occupation layer is somewhat larger in the first case. Further, the difference in layers is also slightly larger for static than dynamic central vertex, although not by much.

As the stocks of the minimum risk portfolio are found on the outskirts of the tree, we expect larger trees (higher L) to have greater *diversification potential*, i.e., the scope of the stock market to eliminate specific risk of the minimum risk portfolio. In order to look at this, we calculated the mean-variance frontiers for the ensemble of 477 stocks using $T=1000$ as the window width. In Fig. 12, we plot the level of portfolio risk as a function of time, and find a similarity between the risk curve and the curves of the mean correlation coefficient $\overline{\rho}$ and normalized tree length L. Earlier, in Ref. [14], when the smaller dataset of 116 stocks—consisting of primarily important industry giants—was used, we found Pearson's linear correlation between the risk and the mean correlation coefficient $\overline{\rho}(t)$ to be 0.82, while that between the risk and the normalized tree length $L(t)$ was -0.90. Therefore, for that dataset, the normalized tree length was able to explain the diversification potential of the market

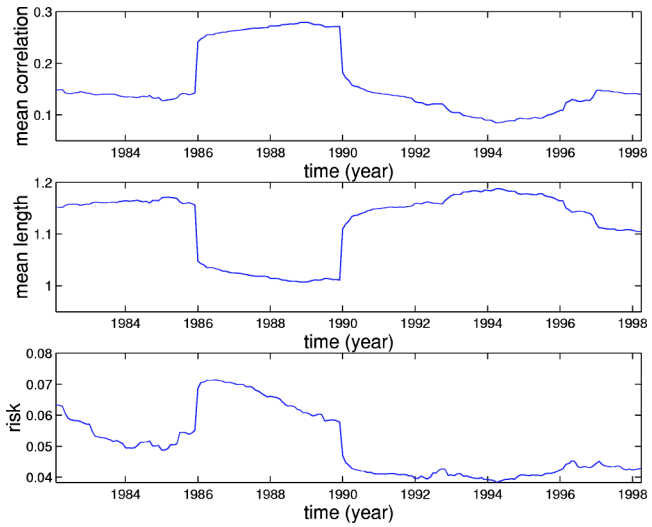

FIG. 12. Plots of (a) the mean correlation coefficient $\overline{\rho}(t)$, (b) the normalized tree length $L(t)$, and (c) the risk of the minimum risk portfolio, as functions of time.

better than the mean correlation coefficient. For the current set of 477 stocks, which includes also less influential companies, the Pearson's linear and Spearman's rank-order correlation coefficients between the risk and the mean correlation coefficient are 0.86 and 0.77, and those between the risk and the normalized tree length are -0.78 and -0.65, respectively.

So far, we have only examined the location of stocks in the minimum risk portfolio, for which $\theta=0$. As we increase θ towards unity, portfolio risk as a function of time soon starts behaving very differently from the mean correlation coefficient and normalized tree length. Consequently, it is no longer useful in describing diversification potential of the market. However, another interesting result emerges: The average weighted portfolio layer $l_{\mathbf{p}}(t,\theta)$ decreases for increasing values of θ, as shown in Fig. 13. This means that out of all the possible Markowitz portfolios, the minimum risk portfolio stocks are located furthest away from the central vertex, and as we move towards portfolios with higher expected return, the stocks included in these portfolios are located closer to the central vertex. When static central node is used, the average values of the weighted portfolio layer $l_{\mathbf{p}}(t,\theta)$ for $\theta=0$, 1/4, 1/2, and 3/4 are 6.03, 5.70, 5.11, and 4.72, respectively. Similarly, for a dynamic central node, we obtain the values of 5.68, 5.34, 4.78, and 4.37. We have not included the weighted portfolio layer for $\theta=1$, as it is not very informative. This is due to the fact that the maximum return portfolio comprises only one asset (the maximum return asset in the current time window) and, therefore, $l_{\mathbf{p}}(t,\theta=1)$ fluctuates wildly as the maximum return asset changes over time.

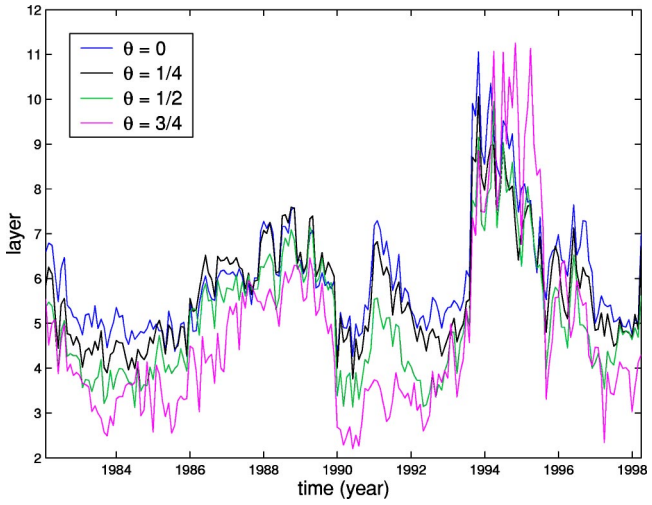

FIG. 13. (Color online) Plots of the weighted minimum risk portfolio layer $l_{\mathbf{p}}(t,\theta)$ for different values of θ.

We believe these results to have potential for practical application. Due to the clustering properties of the MST, as well as the overlap of tree clusters with business sectors as defined by a third party institution, it seems plausible that companies of the same cluster face similar risks, imposed by the external economic environment. These dynamic risks influence the stock prices of the companies, in coarse terms, leading to their clustering in the MST. In addition, the radial location of stocks depends on the chosen portfolio risk level, characterized by the value of θ. Stocks included in low-risk portfolios are consistently located further away from the central node than those included in high-risk portfolios. Consequently, the radial distance of a node, i.e., its occupation layer, is meaningful. Thus, it can be conjectured that the location of a company *within* the cluster reflects its position with regard to internal, or cluster specific, risk. Characterization of stocks by their branch, as well as their location within the branch, enables us to identify the degree of interchangeability of different stocks in the portfolio. For example, in most cases we could pick two stocks from different asset tree clusters, but from nearby layers, and interchange them in the portfolio without considerably altering the characteristics of the portfolio. Therefore, dynamic asset trees provide an intuition-friendly approach to and facilitate *incorporation of subjective judgment* in the portfolio optimization problem.

VIII. SUMMARY AND CONCLUSION

In summary, we have studied the distribution of correlation coefficients and its moments. We have also studied the dynamics of asset trees: the tree evolves over time and the normalized tree length decreases and remains low during a crash, thus implying the shrinking of the asset tree particularly strongly during a stock market crisis. We have also found that the mean occupation layer fluctuates as a function of time, and experiences a downfall at the time of market crisis due to topological changes in the asset tree. Further, our studies of the scale-free structure of the MST show that this graph is not only hierarchical in the sense of a tree but there are special, highly connected nodes and the hierarchical structure is built up from these. As for the portfolio analysis, it was found that the stocks included in the minimum risk portfolio tend to lie on the outskirts of the asset tree: on average the weighted portfolio layer can be almost one and a half levels higher, or further away from the central vertex,

than the mean occupation layer for window width of four years.

For many of the quantities we have studied, the behavior is significantly different for those data windows containing the dates around October 19, 1987 (Black Monday) from windows without them. We have studied the effects of this crash, more specifically in Ref. [10]. We should clarify that the period 1986–1990 which has shown a ''crashlike'' behavior is an artifact of the four-year window width used to analyze the data and except for the dates around October 19, 1987 this period 1986–1990 was ''normal.''

Correlation between the risk and the normalized tree length was found to be strong, though not as strong as the correlation between the risk and the mean correlation coefficient. Thus we conclude that the diversification potential of the market is very closely related also to the behavior of the normalized tree length. Finally, the asset tree can be viewed as a highly graphical tool, and even though it is strongly pruned, it still retains all the essential information of the market and can be used to add subjective judgment to the portfolio optimization problem.

ACKNOWLEDGMENTS

J.-P.O. is grateful to European Science Foundation for REACTOR grant to visit Hungary, the Budapest University of Technology and Economics for the warm hospitality, and the Graduate School in Computational Methods of Information Technology (ComMIT), Finland. The role of Harri Toivonen at the Department of Accounting, Helsinki School of Economics, is acknowledged for carrying out CRSP database extractions. We are also grateful to R. N. Mantegna for very useful discussions and suggestions. This research was partially supported by the Academy of Finland, Research Center for Computational Science and Engineering, Project No. 44897 (Finnish Center of Excellence Program 2000–2005) and OTKA (Grant No. T029985).

[1] R.N. Mantegna, Eur. Phys. J. B **11**, 193 (1999).
[2] L. Kullmann, J. Kertész, and R.N. Mantegna, Physica A **287**, 412 (2000).
[3] L. Giada and M. Marsili, Physica A **315**, 650 (2002).
[4] L. Laloux *et al.*, Phys. Rev. Lett. **83**, 1467 (1999); V. Plerou *et al.*, *ibid.* **83**, 1471 (1999).
[5] *The Economy as an Evolving Complex System II*, edited by W.B. Arthur, S.N. Durlauf, and D.A. Lane (Addison-Wesley, Reading, MA, 1997).
[6] N. Vandewalle, F. Brisbois, and X. Tordoir, Quant. Finance **1**, 372 (2001).
[7] H.M. Markowitz, J. Finance **7**, 77 (1952).
[8] S. Gallucio, J.-P. Bouchaud, and M. Potters, Physica A **259**, 449 (1998); A. Gabor and I. Kondor, *ibid.* **274**, 222 (1999); L. Bongini *et al.*, Eur. Phys. J. B **27**, 263 (2002).
[9] J.-P. Onnela, M. Sc. thesis, Helsinki University of Technology, Finland, 2002.
[10] J.-P. Onnela, A. Chakraborti, K. Kaski, and J. Kertész, Physica A **324**, 247 (2003).
[11] S. Drozdz *et al.*, Physica A **287**, 440 (2000).
[12] J.-P. Onnela, A. Chakraborti, K. Kaski, and J. Kertész, Phys. Scr. T **106**, 48 (2003).
[13] J.-P. Onnela, A. Chakraborti, K. Kaski, and J. Kertész (unpublished).
[14] J.-P. Onnela, A. Chakraborti, K. Kaski, and J. Kertész, Eur. Phys. J. B **30**, 285 (2002).
[15] Forbes at http://www.forbes.com/, referenced in March-April, 2002.
[16] Supplementary material is available at http://www.lce.hut.fi/~jonnela/
[17] Standard & Poor's 500 index at http://www.standardandpoors.com/, referenced in June, 2002.
[18] R. Albert and A.-L. Barabasi, Rev. Mod. Phys. **74**, 47 (2002).
[19] S.N. Dorogovtsev and J.F.F. Mendes, Adv. Phys. **51**, 1079 (2002).
[20] G. Szabó, M. Alava, and J. Kertész, Phys. Rev. E **66**, 026101 (2002).
[21] M. Marsili, Quant. Finance **2**, 297 (2002).
[22] I. Yang, H. Jeong, B. Kahng, and A.-L. Barabasi, e-print cond-mat/0301513; H.-J. Kim, Y. Lee, B. Kahng, and I. Kim, J. Phys. Soc. Jpn. **71**, 2133 (2002).
[23] Several software packages based on standard procedures are available. We used MATLAB with Financial Toolbox.