Eur. Phys. J. B **30**, 285–288 (2002)
DOI: 10.1140/epjb/e2002-00380-9

THE EUROPEAN PHYSICAL JOURNAL B

Dynamic asset trees and portfolio analysis

J.-P. Onnela[1], A. Chakraborti[1], K. Kaski[1], and J. Kertész[1,2]

[1] Laboratory of Computational Engineering, Helsinki University of Technology, PO Box 9203, 02015 HUT, Finland
[2] Department of Theoretical Physics, Budapest University of Technology and Economics, Budafoki út 8, 1111 Budapest, Hungary

Received 7 August 2002 / Received in final form 28 October 2002
Published online 19 December 2002 –

Abstract. The minimum spanning tree, based on the concept of ultrametricity, is constructed from the correlation matrix of stock returns and provides a meaningful economic taxonomy of the stock market. In order to study the dynamics of this asset tree we characterise it by its normalised length and by the mean occupation layer, as measured from an appropriately chosen centre called the 'central node'. We show how the tree evolves over time, and how it shrinks strongly, in particular, during a stock market crisis. We then demonstrate that the assets of the optimal Markowitz portfolio lie practically at all times on the outskirts of the tree. We also show that the normalised tree length and the investment diversification potential are very strongly correlated.

PACS. 89.65.-s Social systems – 89.75.-k Complex systems – 89.90.+n Other topics in areas of applied and interdisciplinary physics

Rapid Note

Portfolio optimisation is one of the basic tools of hedging in a risky and extremely complex financial environment. Many attempts have been made to solve this central problem, starting from the classical approach of Markowitz [1] to more sophisticated treatments, including spin glass type studies [2]. In all of these attempts, correlations between asset prices play a crucial role. A closely related problem is that of economic taxonomy. In a recent paper [3], Mantegna suggested studying the clustering of companies using the correlation matrix of asset returns, such that a simple transformation of the correlations into distances produces a connected graph. In the graph, the nodes are the companies and the distances between them are obtained from the correlation coefficients. The clusters of companies are identified by means of minimum spanning tree. It turned out that the hierarchical structure of the financial market could be identified in accordance with the results obtained by an independent clustering method, based on Potts super-paramagnetic transitions [4]. In another paper by Bonanno *et al.* [5], the time evolution of stock indices was studied and significant changes in the world economy were identified using appropriate time horizons and the minimum spanning tree clustering method. The hierarchical structure explored by the minimum spanning tree also seemed to give information about the influential power of the companies. The network of influence was recently investigated by means of a time-dependent correlation method [6]. Some other attempts have been made to understand the structure of correlation matrices in a highly random setting using the theory of random matrices [7]. In reference [8], the maximum likelihood approach to clustering of financial correlation data was applied and compared to other methods. Though there are differences in the observed cluster structure, both approaches provide a good basis for economic taxonomy. In this paper, we concentrate on the minimum spanning tree as a characteristic graph for the description of the correlations and call it an 'asset tree'. Although this asset tree can reveal a great deal about the taxonomy of the market at a given time, it only represents a static average of an evolving complex system. This evolution is a reflection of the changing power structure in the market and manifests the passing of different products and product generations, new technologies, management teams, alliances and partnerships, amongst many other things. This is why exploring the asset tree *dynamics* can provide new insights into the market. Here, by studying the time evolution of the asset tree we show that although the structure of the tree changes with time, the companies of the optimal Markowitz portfolio are always on its outer leaves. We also study the robustness of the 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.

We start our analysis by assuming that there are N assets with price $P_i(t)$ for asset i at time t. Then the logarithmic return of stock i is $r_i(t) = \ln P_i(t) - \ln P_i(t-1)$ which,

for a certain consecutive sequence of trading days, forms the return vector $\boldsymbol{r}_i$. In order to characterise the synchronous time evolution of stocks, we use the equal time correlation coefficients between stocks i and j defined as

$$\rho_{ij} = \frac{\langle \boldsymbol{r}_i \boldsymbol{r}_j \rangle - \langle \boldsymbol{r}_i \rangle \langle \boldsymbol{r}_j \rangle}{\sqrt{[\langle \boldsymbol{r}_i^2 \rangle - \langle \boldsymbol{r}_i \rangle^2][\langle \boldsymbol{r}_j^2 \rangle - \langle \boldsymbol{r}_j \rangle^2]}}, \qquad (1)$$

where $\langle ... \rangle$ indicates a time average over the trading days included in the return vectors. These correlation coefficients forming an $N \times N$ matrix with $-1 \leq \rho_{ij} \leq 1$, are then transformed into an $N \times N$ distance matrix with elements $d_{ij} = \sqrt{2(1-\rho_{ij})}$, such that $2 \geq d_{ij} \geq 0$. The d_{ij}s fulfil the requirements of distance, even those of ultrametricity [3]. The distance matrix is used to determine the minimum spanning tree (MST) of the distances, denoted by $\mathbf{T}$, which is a simply connected graph that connects all the N nodes of the graph with $N-1$ edges such that the sum of all edge weights, $\sum_{(i,j)\in\mathbf{T}} d_{ij}$, is minimum. 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.

The dataset we have used in this study consists of daily closure prices for 116 stocks of the S&P 500 index [9], obtained from the Yahoo website [10]. The time period of this data extends from the beginning of 1982 to the end of 2000 including a total of 4787 price quotes per stock, after the removal of a few days due to incomplete data. We divide this data into M *windows* of width T corresponding to the number of daily returns included in the window. Different windows, marked with time variable $t = 1, 2, ..., M$, overlap with each other, the extent of which is dictated by the window step length parameter δT, describing the displacement between two consecutive windows, measured also by the number of 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. In our studies, T was typically set at between 500 and 1500 trading days, *i.e.* 2 and 6 years, and δT at one month, including about 21 trading days. This is in accordance with the suggestions of the Basel committee [11]. A typical asset tree, based on the above described data, is shown in Figure 1, where it is evident that companies become clustered by business sectors.

In order to study the temporal state of the market, we define the *normalised tree length* as

$$L(t) = \frac{1}{N-1} \sum_{d_{ij}\in\mathbf{T}^t} d_{ij}, \qquad (2)$$

where t denotes the time at which the tree is constructed, and $N-1$ is the number of edges present in the MST. Figures 2a and b show how the normalised tree length L and the mean correlation coefficient, defined as $\bar{\rho} = \frac{1}{N(N-1)/2} \sum \rho_{ij}$, where we consider only the non-diagonal and independent ρ_{ij}, evolve with time. The two curves look like mirror images, which is corroborated by Pearson's linear correlation coefficient of -0.96, indicating that

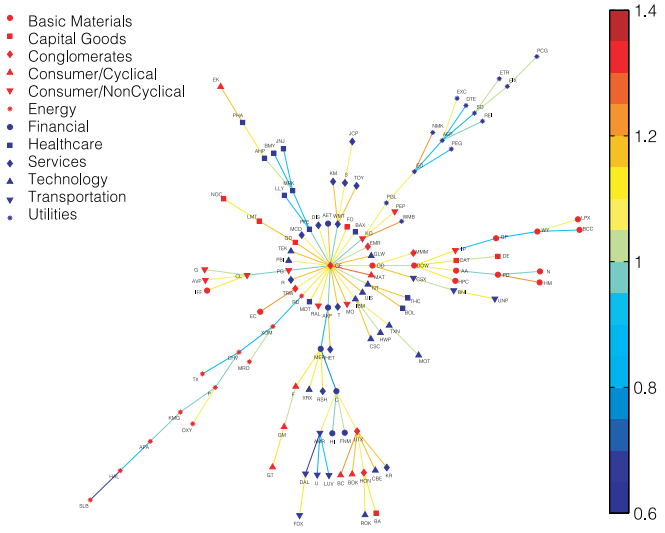

Fig. 1. A typical asset taxonomy graph (minimum spanning tree) connecting the examined 116 stocks of the S&P 500 index. The graph was produced using four-year window width and it is centred on January 1, 1998. Distance between the nodes is indicated by the colour of the edges, as given by the colour bar on the right. Business sectors are assigned according to Forbes, `http://www.forbes.com`. In this graph, General Electric (GE) was used as a central node and eight layers can be identified.

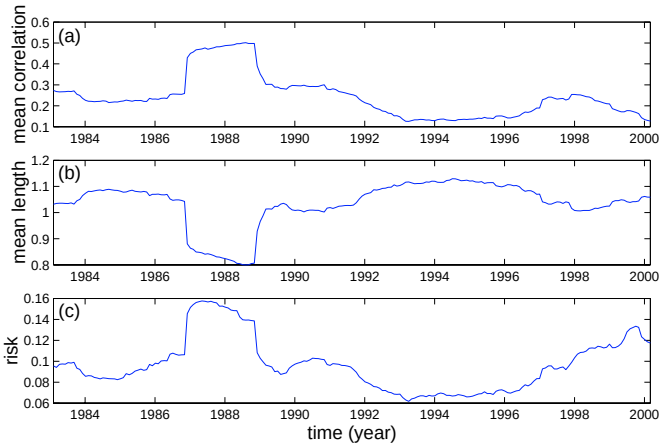

Fig. 2. Plots of (a) the mean correlation coefficient $\bar{\rho}$, (b) the normalised tree length L and (c) the risk of the minimum risk portfolio, as functions of time. The risk is determined with weight limits of zero lower bound (no short-selling) and unit upper bound (any asset may constitute the entire portfolio). For all plots the window width is $T = 500$, *i.e.* two trading years.

the minimum spanning tree is a strongly reduced representative of the whole correlation matrix, and bears the essential information about asset correlations. One would, indeed, expect the two measures to be anti-correlated in view of how the distances d_{ij} are constructed from correlation coefficients ρ_{ij}. However, the extent of this anti-correlation is different for different input variables and is lower if, say, daily transaction volumes are studied instead of daily closure prices [12]. 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 in Figure 2. The two sides of the ridge, in fact, converge as a result of extrapolating the window width $T \to 0$ [13]. In Figure 2a, the mean correlation of stocks is very high during the crash, which is due to market forces acting strongly on the stocks and pushing the market to behave in a unified way. The increased value of the mean correlation is in accordance with the observation by Drozdz *et al.*, who found that the maximum eigenvalue, which carries most of the correlations, is very large during market crashes [14]. Also Figure 2b supports this fact: $L(t)$ decreases indicating that the nodes on the graph are drawn closer together during the crash.

In order to characterise the spread of nodes on a graph, we introduce the quantity of *mean occupation layer* as

$$l(t) = \frac{1}{N} \sum_{i=1}^{N} \mathrm{lev}(v_i^t), \tag{3}$$

where $\mathrm{lev}(v_i)$ denotes the level of vertex v_i in relation to the *central node*, the level of which is taken to be zero. Although there is arbitrariness in the choice of the central node, 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 as a whole. Thus, the central node would be the company which is most strongly connected to its nearest neighbours in the tree. With this choice the sum of the correlation coefficients calculated for the incident edges would be maximum, and/or have the highest *vertex degree* (the number of edges which are incident with the vertex). It is also noted that one can have either a static (fixed at all times) or a dynamic (continuously updated) central node, without considerable effect on the results. In our studies, General Electric (GE) was chosen as the static central node, since for about 70% of the time windows it turned out to be the most connected node. We would like to emphasize that this central site concept, based on the local property of highest connectedness, should not be confused with the global property of *centre of mass* of the tree. The centre of mass is defined as the node v_i that produces the lowest value of the mean occupation layer $l(t, v_i)$, where v_i is used as the central node. We observed that roughly 80 percent of the time the central node coincides with the centre of mass, indicating that an essential feature is captured by identifying the central node [12]. In Figure 3, we have plotted the mean occupation layer $l(t)$ as a function of time for a static central vertex (GE). We find that $l(t)$ reaches a very low value on two occasions, which can be traced to the 1987 stock market crash.

Next, we apply the above discussed concepts and measures to portfolio analysis. Let us consider a minimum risk Markowitz portfolio $P(t)$ with the asset weights w_1, w_2,..., w_N. In the Markowitz portfolio optimisation scheme, financial assets are characterised by their average return and risk, both determined from historical price data, where risk is measured by the standard deviation of returns. The aim is to optimise the asset weights so that

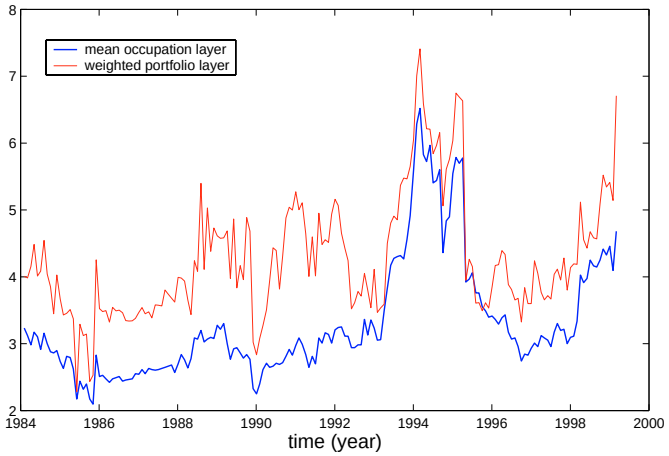

Fig. 3. Plots of mean occupation layer l and weighted portfolio layer l_P as functions of time. This plot is based on the window width $T = 1000$, *i.e.* four trading years.

the overall portfolio risk is minimised for a given portfolio return [15]. In the minimum spanning tree framework, the task is to determine how the assets are located with respect to the central node. Intuitively, we expect the weights to be distributed on the outskirts of the graph. In order to describe what happens, we define a single measure, the *weighted portfolio layer* as

$$l_P(t) = \sum_{i \in P} w_i \, \mathrm{lev}(v_i^t), \tag{4}$$

with the constraint $w_i \geq 0$ for all i, since we assume that there is no short-selling.

Figure 3 shows the behaviour of the weighted minimum risk portfolio layer $l_P(t)$ together with the mean occupation layer $l(t)$. We find that the portfolio layer is higher than the mean layer almost all the time. The difference in layers depends to a certain extent on the window width: for $T = 500$ it is about 0.76 and for $T = 1000$ about 0.97. As the stocks of the minimum risk portfolio are found on the outskirts of the graph, 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 116 stocks using $T = 500$ as the window width. In Figure 2c, we plot the level of portfolio risk as a function of time, and find a striking similarity between the risk curve and the curves of the mean correlation coefficient $\bar{\rho}$ and normalised tree length L of Figures 2a and b. Pearson's linear and Spearman's rank-order correlation coefficients between risk and mean correlation coefficient $\bar{\rho}$ are 0.82 and 0.73, while those between risk and normalised tree length L are -0.90 and -0.88, respectively. Therefore, the latter result explains the diversification potential of the market better.

We believe these results 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 (see Fig. 1),

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, because the stocks included in the minimum risk portfolio are consistently located on the outskirts of the tree, the distance of the nodes from the root of the tree (*i.e.*, layer) must be 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. Characterisation 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. Therefore, dynamic asset trees provide an intuition-friendly approach to and facilitate *incorporation of subjective judgement* in the portfolio optimisation problem.

Finally, in order to investigate the robustness of the minimum spanning tree topology, we define the survival ratio of tree edges, *i.e.*, the fraction of edges found common in two consecutive graphs at time t and $t-1$, as

$$\sigma_t = \frac{1}{N-1}|E^t \cap E^{t-1}|.$$

In this E^t refers to the set of edges of the graph at time t, $\cap$ is the intersection operator and $|...|$ gives the number of elements in the set. Under normal circumstances, the graphs at two consecutive time windows t and $t+1$ (for small values of δT) should look very similar. Whereas some of the differences can reflect real changes in the asset taxonomy, others may simply be due to noise. We find that as $\delta T \to 0$, $\sigma_t \to 1$ [13], indicating that the graphs *are* stable in the limit and, hence, our portfolio analysis is justified.

In summary, we have studied the dynamics of asset trees and applied it to portfolio analysis. We have shown that the tree evolves over time and have found that the normalised tree length decreases and remains low during a crash, thus implying a particularly strong shrinking of the asset tree 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. As for 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 is about 1 level higher, or further away from the central node, than mean occupation layer for a window width of four years. The correlation between risk and mean correlation coefficient was found to be quite strong, though not as strong as the correlation between risk and normalised tree length. Thus, it can be concluded that the diversification potential of the market is very closely related to the behaviour of the normalised tree length.

J.-P.O. is grateful to European Science Foundation for the grant to visit Hungary, the Budapest University of Technology and Economics for the warm hospitality and L. Kullmann for stimulating discussions. This research was partially supported by the Academy of Finland, Research Centre for Computational Science and Engineering, project no. 44897 (Finnish Centre of Excellence Programme 2000-2005) and OTKA (T029985).

References

1. G. Kim, H.M. Markowitz, J. Portfolio Management **16**, 45 (1989)
2. S. Gallucio, J. -P. Bouchaud, M. Potters, Physica A **259**, 449 (1998); A. Gabor, I. Kondor, Physica A **274**, 222 (1999); L. Bongini *et al.*, Eur. Phys. J. B **27**, 263 (2002)
3. R.N. Mantegna, Eur. Phys. J. B **11**, 193 (1999)
4. L. Kullmann, J. Kertész, R.N. Mantegna, Physica A **287**, 412 (2000)
5. G. Bonanno, N. Vandewalle, R.N. Mantegna, Phys. Rev. E **62**, R7615 (2000)
6. L. Kullmann, J. Kertész, K. Kaski, Phys. Rev. E **66**, 026125 (2002)
7. L. Laloux *et al.*, Phys. Rev. Lett. **83**, 1467 (1999); V. Plerou *et al.*, preprint available at `cond-mat/9902283` (1999)
8. L. Giada, M. Marsili, preprint available at `cond-mat/0204202` (2002)
9. Standard and Poor's 500 index at `http://www.standardandpoors.com/`, referenced in June, 2002
10. Yahoo at `http://finance.yahoo.com` referenced in July, 2001
11. Basel Committee on Banking Supervision at `http://www.bis.org/bcbs/`, referenced in September, 2001
12. J.-P. Onnela, A. Chakraborti, K. Kaski, J. Kertész, in preparation (2002)
13. J.-P. Onnela, *Taxonomy of Financial Assets*, M. Sc. thesis, Helsinki University of Technology, Helsinki, Finland (2002)
14. S. Drozdz *et al.*, preprint available at `cond-mat/9911168` (1999)
15. Several software packages based on standard procedures are available. We used MATLAB with Financial Toolbox

Rapid Note