Revealing Ownership Structure of Funds Using Minimum Spanning Tree

M. Rešovský, M. Gróf, D. Horváth and V. Gazda

Technical University of Košice, Faculty of Economics, Košice, Slovakia

The traditional method of analysing a time series of stocks and funds is to use simple Pearson correlations. However, experience shows that cross-correlations are not an accurate indicator of the mutual ownership relations. We show that the minimum spanning tree methodology, previously used to perform more comprehensive studies of asset returns correlations, can be used to deduce the underlying ownership structure with reasonable accuracy. We also show that adjusting the time series for a common trend of stocks and subsequent filtering of the short term variations of returns using the ARIMA model is a prerequisite for this application of the minimum spanning tree.

1. Introduction

Statistical analysis of multiple financial time series has undergone a complete transformation in the recent decades with the development of new methods, software and processing tools supporting advanced structural interpretation of data. Many authors have recently attempted to properly visualise relationships between assets traded in the international financial markets. Intuitively, the collection of assets or the corresponding data may be perceived as objects interconnected by communication and information exchange channels forming a graph or network characterised by specific topological irregularities, modules or hierarchies. Rosario N. Mantegna may be considered as one of the most representative authors dealing with the graph representation of financial markets. In his work Mantegna [1] analysed a network of stocks included in the Dow Jones Industrial Average and S&P500. The vertices of the graph represented individual stocks, with the weights of edges defined by the Pearson linear correlation coefficient transformed into Euclidean distances (if the correlation increases, the distance between edges becomes smaller). Clearly, the evaluation and consequent interpretation of the properties of these complete graphs is complicated at best. Therefore, Mantegna applied the concept of the *minimum spanning tree* (MST) already well-known in the graph theoretic community. This approach is based on the construction of a connected graph of n vertices using $n - 1$ edges minimising the sum of edge weights.

Subsequently, a period of natural diversification of methodological options followed. In this context three applications of Bonanno et al. [2], Onnela et al. [3], and Micciche et al. [4] might be mentioned. In [2], a one-factor model is used to filter out the common trend of the stock prices, while in [3] the temporal variability

of the graphs is studied using the rolling window approach. Finally, the work of Micciche et al. [4] puts the emphasis on the volatility of the stock prices. In addition to MST, several alternative methodologies can be mentioned. In order to display only the relevant relations between assets Huang et al. [5], Tse et al. [6], Namaki et al. [7] have used the threshold method, where the edge is added only when the correlation between the pair of assets exceeds a given threshold value. The most remarkable difference between the threshold and MST filtering consists in the formation of graph cycles (present only in the case of the threshold method).

The main problem of the MST algorithm is its tendency to select pairwise relations that do not coincide with the strongest correlations. The problem may be seen in the improper trade-off between the graph connectivity requirement and the selection of the $n - 1$ edges with the strongest weights. The disadvantages of both the MST and threshold method later encouraged the work on more advanced ways of graph representation of data (i.e. Tumminello et al. [8]). The algorithm applied in the mentioned work originates from the traditional MST, with the difference of allowing cycles but not self-crossing of edges. The method is more complex but less subjective compared to the threshold method. In addition, compared to the MST method, more relevant edges are included into the final graph including the circulation contours.

In our paper we are going deeper into understanding the structure consisting of funds and stocks owned by these funds. We compare the structure obtained by applying the minimum spanning tree method on the distances calculated from the correlation matrix of asset returns with the ownership structure obtained from the fund prospectus. The aim of this paper is to test the ability of the minimum spanning tree approach applied to the returns of assets to reveal the underlying ownership structure.

The data used in our analysis is described in the following section. The third section contains the methodology

---

corresponding author; e-mail: marcel.resovsky@tuke.sk

used. Selected graph and vertex properties are defined in Sect. 4. The results of the performed analysis are given in Sect. 5.

2. Data description

The data used was obtained from the Thomson Reuters database. It contains daily time series of the prices of eight open ended funds (further denoted as Funds) of the South Africa asset manager Stanlib and daily time series of prices of the stocks owned by these Funds (denoted as Stocks) from November 12, 2009 to June 12, 2012. The ownership structure was obtained from prospectus of every Fund in 3 month intervals between January 1, 2010 and June 30, 2012. A Stock was included into the data set, if it was in the portfolio of any of the Funds for at least one period. All together, the data contained 8 Funds and 58 Stocks. We excluded 30 days from the overall study period, for which the data were unavailable, for a total of 644 time series observations.

For the purposes of our analysis we have chosen the following Funds from the portfolio of the investment manager Stanlib (Fund labels as used in the following text are in parentheses): Resources Fund (F1), Financials Fund (F2), Growth Fund (F3), Value Fund (F4), Alsi Fund (F5), SA Equity Fund (F6), Institutional Fund (F7), Industrial Fund (F8).

The primary aim when selecting data was to find Funds with overlapping portfolios and with complete portfolio composition information. In our case, almost 40% of Stocks are not uniquely owned and are included in portfolios of more than one Fund. Our intention is to find an appropriate tool to reveal the ownership structure, thus the proposed methodology should be applied on data with a more complicated structure (overlapping portfolios in the case of Funds).

3. Methodology

In this section, we describe our specific adaptation of the methodology of Mantegna [1] and Namaki et al. [7]. We started by defining the logarithmic returns of 66 assets

$$r_i^t = \ln P_i^t - \ln P_i^{t-1}, \qquad i = 1, 2, \ldots, 66, \qquad (3.1)$$

where indices $i = 1, 2, \ldots, 8$ correspond to the Funds, and the remaining indices belong to the 58 Stocks. In order to extract the common trend of the Stocks, we applied a one factor regression model (we take into consideration only the common trend of Stocks, since the prices of Funds are calculated based upon the prices of Stocks in their portfolios)

$$r_i^t = \alpha_i + \beta_i M^t + \epsilon_i^t, \qquad (3.2)$$

where M^t is the so called market factor, α_i, β_i are regression coefficients. The M^t dynamics can be approximated by the suitably selected market index. Due to empirical arguments given later, we decided to follow the methodology of Namaki et al. [7] who calculated M^t using the eigenvector corresponding to the largest eigenvalue of the correlation matrix of r_i^t. In our case, the eigenvector $(u_9, u_{10}, \ldots, u_{66})$ including 58 components has been used

to approximate

$$M^t = \sum_{i=9}^{66} u_i r_i^t. \qquad (3.3)$$

The residuals ϵ_i^t were then further adjusted for the short-term memory by the autoregressive-integrated-moving-average (ARIMA) model identified using the stepwise algorithm proposed by Hyndman and Khandakar [9]. Formally, the filtering leads to the modified residuals (used as filtered returns in the reminder of our work) $\bar{\epsilon}_i^t$. Afterwards, we applied the methodology of Mantegna [1]. Using the filtered returns we computed the correlation coefficient between each pair of assets

$$\rho_{ij} = \frac{\langle \bar{\epsilon}_i^t \bar{\epsilon}_j^t \rangle - \langle \bar{\epsilon}_i^t \rangle \langle \bar{\epsilon}_j^t \rangle}{\sqrt{(\langle \bar{\epsilon}_i^{t2} \rangle - \langle \bar{\epsilon}_i^t \rangle^2)(\langle \bar{\epsilon}_j^{t2} \rangle - \langle \bar{\epsilon}_j^t \rangle^2)}}, \qquad (3.4)$$

where $\langle \ldots \rangle$ stands for temporal averaging. Subsequently, the correlation coefficients were transformed into distances between the assets according to the well known relation

$$d_{ij} = \sqrt{2(1 - \rho_{ij})}, \qquad (3.5)$$

which can be used to construct the MST of the Stocks (denoted as MST(S)) and the MST including both Stocks and Funds (denoted as MST(F)).

Using the aforementioned mostly technically motivated steps, the focus of our research is now the identification of the underlying ownership structure of Funds and Stocks. We assume that the MST applied to the asset returns is a good candidate for the identification of the ownership relations. Since the common trend was calculated using all the Stocks combined and the best fit ARIMA model had the same parameter specification for each asset when adjusting for the short-term memory, we expect the information concerning the ownership relations to be preserved.

The aforementioned filtering procedure is justified by the results that can be seen in Figs. 1–3. Figure 1 depicts the reduction of the pair correlations of asset returns after applying the filtering procedure. However, as can be seen, the level of correlation between Funds is still high. This can be explained by the aforementioned overlapping of Fund portfolios. Since the prices of Funds are calculated from the prices of Stocks in their portfolios, the returns of Funds with overlapping portfolios are more correlated and vice versa. For example, the returns of Fund F7 are not as strongly correlated with the returns of other Funds, unlike the remaining Funds, because Stocks from the portfolio of Fund F7 are exclusively owned by this Fund. In like manner, returns of any two Funds with overlapping portfolios are more correlated than returns of Fund and returns of any Stock from its own portfolio.

In Fig. 2 and Fig. 3 the MST for each Fund and Stocks owned by the given Fund is presented for both unfiltered and filtered returns. As can be seen, before the filtering the Funds were not located in the centre of their MST structure except for Fund F7. After the time series adjustment all of the Funds with the exception of Fund
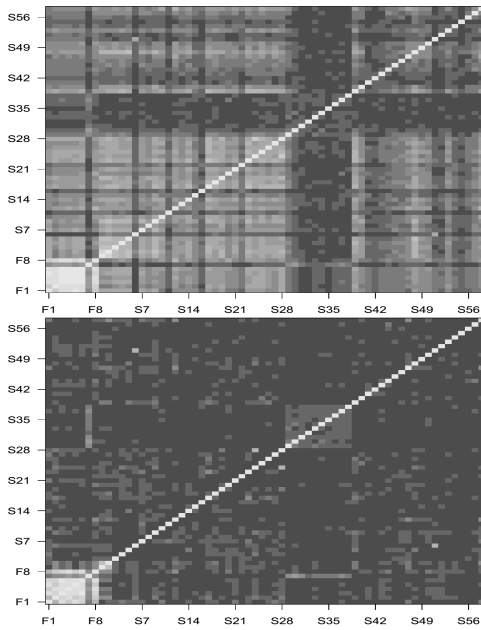
Fig. 1. The correlation matrix constructed for the logarithmic returns of assets for unfiltered (upper part) and filtered (lower part) returns. Stronger correlations are depicted by the lighter grey.
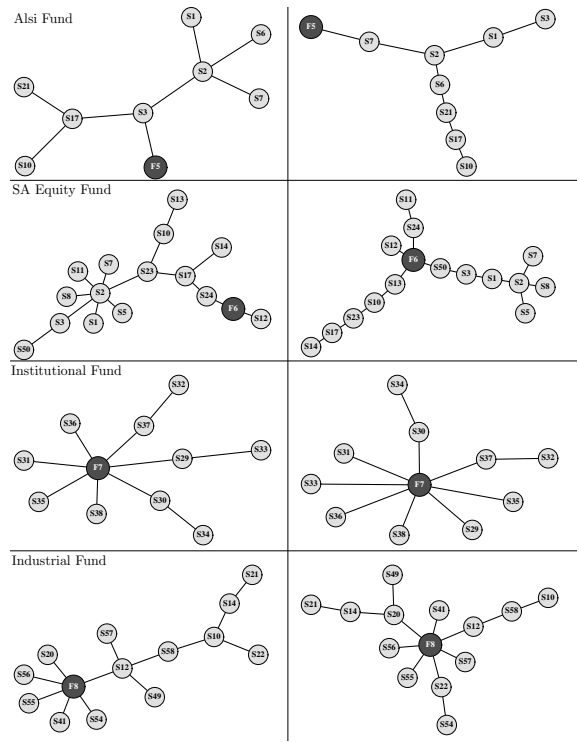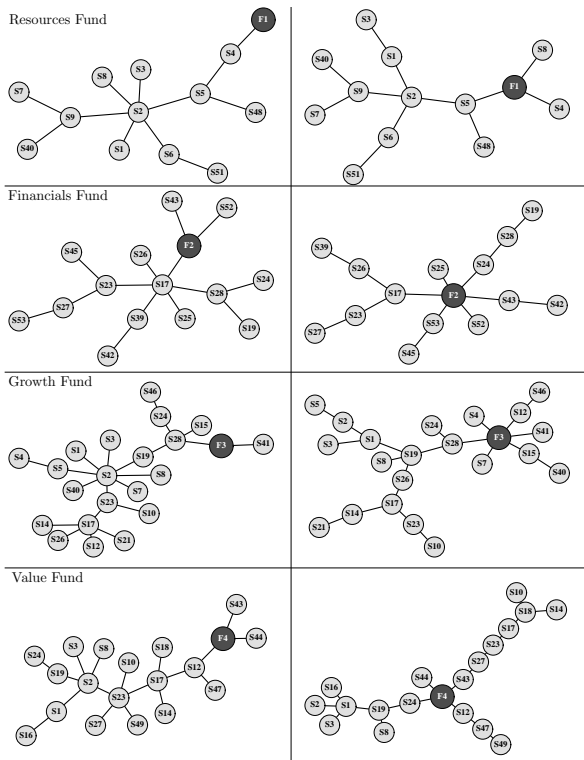

Fig. 2. The differences in the position of Funds (dark points) and Stocks (light points) observed within the MST when comparing unfiltered (left sequence of figures) and filtered (right sequence) returns. The figure illustrates the importance of filtering the logarithmic return series for the common trend and short memory to achieve a more reliable determination of the ownership structure of Stocks.

Fig. 3. Continuation of Fig. 2. MSTs of unfiltered (left part) and filtered (right part) returns.

F5 moved closer to the graph centre in agreement with our previous expectations. This observation is further supported by the centrality measures calculated for Funds in their respective MSTs before and after filtering (see Table I).

TABLE I

Centrality measures for Funds in MSTs of unfiltered and filterd returns depicted in Fig. 2 and Fig. 3.

Fund	F1	F2	F3	F4	F5	F6	F7	F8
Degree centrality								
unfiltered returns	1.00	3.00	2.00	3.00	1.00	2.00	7.00	6.00
filtered returns	3.00	6.00	6.00	4.00	1.00	4.00	8.00	7.00
Closeness centrality								
unfiltered returns	0.27	0.44	0.26	0.28	0.40	0.25	0.77	0.46
filtered returns	0.38	0.56	0.35	0.33	0.29	0.34	0.83	0.62
Betweenness centrality								
unfiltered returns	0.00	25.00	20.00	33.00	0.00	14.00	42.00	50.00
filtered returns	21.00	76.00	117.00	108.00	0.00	73.00	43.00	68.00

4. Selected graph and vertex properties

In this section, the graph characteristics, which will be used to describe the most important properties of the MST structures, are defined*. Our focus is concentrated

---

*The given graph characteristics are specifically defined for a tree.

on the structure formed by Stocks exclusively as well as Stocks and Funds combined.

Diameter of a graph is the length of the longest path d_{ij} in the graph

$$d_{\mathrm{G}} = \max_{i,j} d_{ij}. \tag{4.1}$$

Closeness centrality of vertex v is the sum of the inverse lengths of paths d_{vi} connecting vertices $i \neq v$ to v. It is calculated as

$$C_v = \sum_{i \neq v} \frac{1}{d_{vi}}. \tag{4.2}$$

Betweenness centrality B_v of vertex v is the number of paths between vertices i and j that pass through vertex v in the graph.

Degree centrality D_v of vertex v is equal to the number of vertices adjacent to vertex v.

Normalised form of the closeness centrality can be calculated using

$$C_v^{\mathrm{norm}} = \sum_{i \neq v} \frac{n-1}{d_{vi}}. \tag{4.3}$$

Normalised form of the betweenness centrality

$$B_v^{\mathrm{norm}} = \frac{2B_v}{(n-1)(n-2)} \tag{4.4}$$

uses normalisation by the number of edges in a complete undirected graph with n vertices.

Normalised form of the degree centrality may be introduced as follows:

$$D_v^{\mathrm{norm}} = \frac{D_v}{n-1}. \tag{4.5}$$

Graph centralisation is a graph characteristic expressing the variability of a centrality measure in the given structure to the variability of the given centrality measure in the maximally centralised graph (star)

$$\mathbb{C}_{\mathrm{G}} = \sum_{i=1}^{n} (\mathbb{C}_* - \mathbb{C}_i) \bigg/ \sum_{j=1}^{n} (\mathbb{C}_*^{\mathrm{star\ graph}} - \mathbb{C}_j^{\mathrm{star\ graph}}), \tag{4.6}$$

where $\mathbb{C}_i$ is the centrality of the i-th vertex, $\mathbb{C}_*$ is the maximum value of vertex centrality in a given graph, and the denominator normalises the sum by the maximal sum, which can be obtained in a star with n vertices. The list of possible centrality measures includes: D_i, B_i, C_i. In the following, we considered only the case of $\mathbb{C}_i = D_i$.

5. Results

In this section, the application of the methodology is presented. Firstly, we construct the MST using the filtered returns of Stocks exclusively (see Fig. 4). Excluding the Funds from this graph means the ownership structure is not present, and the Stocks should have no commonalities as a result. The resulting graph structure is denoted as MST(S). Subsequently, we applied the methodology to the mixture of Stocks and Funds, denoted as MST(F), effectively introducing ownership relations into the graph. Even the visual inspection of these two MST graphs reveals crucial topological differences between them. As we expected, there is no Stock playing a central role

in the whole graph in the MST(S). Instead, there are only partially central Stocks that are connected with two or three other Stocks on average. On the contrary, the MST(F) is more centralised (graph diameter of 13 compared to graph diameter of 23 for MST(S)), with the central position of Funds. We attribute the fact the Funds are congested in the centre of the graph to the ownership structure introduced into the graph. The fact the Funds have direct connections among them is the result of the high correlations among their returns, as previously discussed.
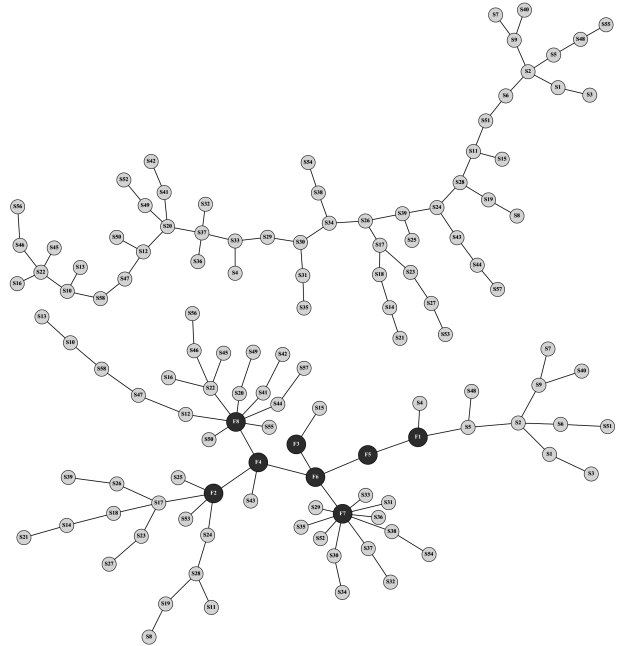

Fig. 4. MST constructed using all studied assets. Upper part depicts the MST(S) (exclusively Stock structure). The lower part of the figure depicts the MST(F) with the Funds (dark points) incorporated into the structure.

The influence of the ownership structure is well determinable in the case of Fund F7 and Fund F8. These Funds are represented by the vertices with a high degree centrality in the MST, with the edges connecting them to the Stocks predominantly corresponding to their portfolios. However, in general, the Stocks are not uniquely linked to their corresponding Funds. This is most likely due to the high overlap of the portfolios between individual Funds. Nevertheless, in the case of MST(F), the edges between Funds and Stocks mostly coincide with the ownership relations (21 Stocks in MST(F) are directly linked to Funds that include this Stocks in their portfolios, with 35 Stocks with a geodesic distance of 2 or less). The differences between MST(S) and MST(F) can be properly validated by the MST characteristics defined in Sect. 4. The centralisation of MST(S) (see Eq.(4.6)) is only 0.037 but it is 0.127 for MST(F).

The distributions of different graph centrality measures are depicted in Fig. 5. We see that distribution
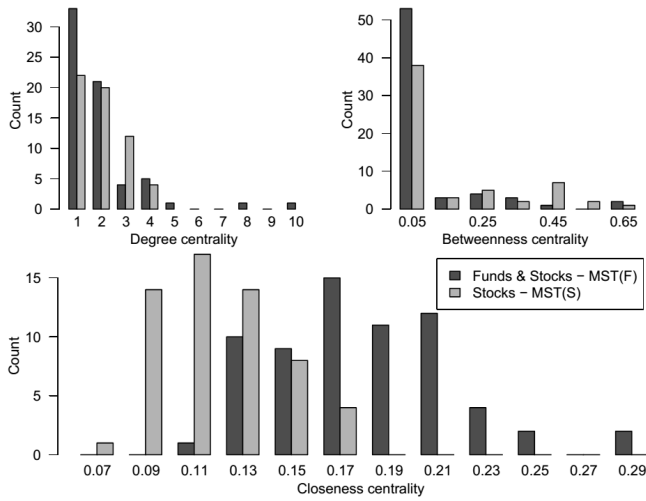
Fig. 5. The distributions of the centrality measures (see Sect. 4) calculated for MST(F) (black) and MST(S) (grey). The closeness and betweenness centrality measures are normalised according to Eq. (4.3) and Eq. (4.4).

of the closeness centrality for MST(F) is shifted to the higher values when compared to MST(S), which again indicates higher centralisation of MST(F). In addition, MST(F) exhibits a wider distribution of degree centrality, i.e. most of the vertices have small degree and betweenness centrality, which is caused mainly by the peripheral position of many Stocks that are linked only to the central Funds. On the other hand, there is a small group of Funds that achieve higher centrality scores compared to the group of Stocks (normalised betweenness centrality values over 0.6 and vertex degree centrality $D_v = 5, 6, 10$). A comparison of mean graph characteristics and graph diameter of MST(F) and MST(S) is given in Table II.

TABLE II

Comparison of graph characteristics for minimum spanning tree of Stocks and Funds combined (MST(F)) and Stocks exclusively (MSTS(S)).

	MST(F)	MST(S)
d_{G}	13	23
mean (C_v^{norm})	0.18	0.12
mean (B_v^{norm})	0.07	0.14

Finally, we decided to make a comparison with a more traditional method. The method is based on clustering aggregation and grouping of similar objects together. The illustrative results of the comparison for the selected Fund and its Stocks are depicted in Fig. 6. The important finding is that MST is much more efficient in revealing the ownership structure. On the other hand, we see that the clustering pushes Fund F7 away from its Stocks.
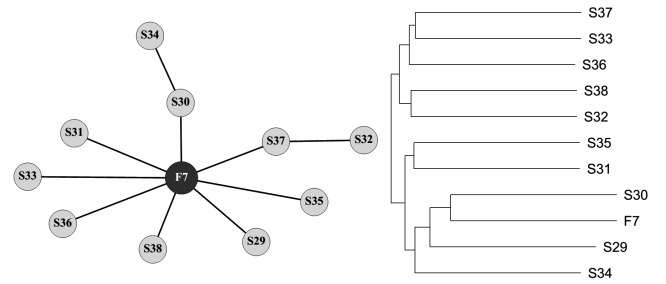

Fig. 6. The MST compared with a dendrogram (obtained for the Euclidean distance and complete linkage clustering) with a focus on the ownership relations. The results of Fund F7 and its Stocks.

6. Conclusions

In the paper, we concentrate on the analysis of financial data series, combining statistical and graph theoretical concepts. In particular, we discuss the application of the MST graph construction method. Even the first applications of MST to financial data by Mantegna [1] have shown that while a specific reduction of the cross-correlation matrix has some potential, to find more thorough interpretation, a comprehensive exploration of the MST approach into specific applications and different economic sectors is necessary. Our analysis shows that the MST is partially able to detect existing ownership relationships of Funds and Stocks. It should be noted that such relations are not straightforwardly deducible from the cross-correlations. We have shown that the MST method could be a promising tool in this regard. Our analysis also revealed that filtering of data for the common trend of Stocks and the short memory is an important prerequisite for the ownership structure estimation. This empirical findings suggest a rather close relationship between the long-term behaviour and ownership.

References

[1] R.N. Mantegna, *Europ. Phys. J. B* **11**, 193 (1999).
[2] G. Bonanno, G. Caldarelli, F. Lillo, R.N. Mantegna, *Phys. Rev. E* **68**, 046130 (2003).
[3] J.P. Onnela, A. Chakraborti, K. Kaski, J. Kertész, A. Kanto, *Phys. Scr.* **T106**, 48 (2003).
[4] S. Micciché, G. Bonanno, F. Lillo, R.N. Mantegna, *Physica A* **324**, 66 (2003).
[5] W.Q. Huang, X.T. Zhuang, S. Yao, *Physica A* **388**, 2956 (2009).
[6] Ch.K. Tse, J. Liu, F.C.M. Lau, *J. Empir. Fin.* **17**, 659 (2010).
[7] A. Namaki, A.H. Shirazi, R. Raei, G.R. Jafari, *Physica A* **390**, 3835 (2011).
[8] M. Tumminello, T. Aste, T. Di Matteo, R.N. Mantegna, *Proc. Natl. Acad. Sci.* **102**, 10421 (2005).
[9] R.J. Hyndman, Y. Khandakar, *J. Stat. Software* **27**, 1 (2008).