Clustering and hierarchy of financial markets data: advantages of the DBHT

Nicoló Musmeci[1], Tomaso Aste[2,3,*], Tiziana Di Matteo[1]
1 Department of Mathematics, King's College London, The Strand, London, WC2R 2LS
2 Department of Computer Science, UCL, Gower Street, London, WC1E 6BT, UK
3 Systemic Risk Centre, London School of Economics and Political Sciences, London, WC2A2AE, UK
∗ **E-mail: t.aste@ucl.ac.uk**

Abstract

We present a set of analyses aiming at quantifying the amount of information filtered by different hierarchical clustering methods on correlations between stock returns. In particular we apply, for the first time to financial data, a novel hierarchical clustering approach, the Directed Bubble Hierarchical Tree (DBHT), and we compare it with other methods including the Linkage and k-medoids. In particular by taking the industrial sector classification of stocks as a benchmark partition we evaluate how the different methods retrieve this classification.

The results show that the Directed Bubble Hierarchical Tree outperforms the other methods, being able to retrieve more information with fewer clusters. Moreover, we show that the economic information is hidden at different levels of the hierarchical structures depending on the clustering method. The dynamical analysis also reveals that the different methods show different degrees of sensitivity to financial events, like crises. These results can be of interest for all the applications of clustering methods to portfolio optimization and risk hedging.

Introduction

Correlation-based networks have been extensively used in Econophysics as tools to filter, visualise and analyse financial market data [1], [2], [3], [4], [5], [6], [7], [8]. Since the seminal work of Mantegna on the Minimum Spanning Tree (MST) [1] they have provided insights into several aspects of financial markets including financial crises [9], [10], [11], [12], [13], [14], [15].

The MST is strictly related [16] to a hierarchical clustering algorithm, namely the Single Linkage (SL) [17]. Starting from a set of elements (e.g., stocks) and a related distance matrix (e.g., a convenient transformation of the stocks correlation matrix [1]), the SL performs an agglomerative algorithm that ends up with a tree (dendrogram) that arranges the elements in a hierarchical structure [16]. The filtering procedure linked to MST and SL has been succesfully applied to improve portfolio optimization [10]. Another hierarchical clustering method, the Average Linkage (AL), has been shown to be associated to a slightly different version of spanning tree [18], that the authors called the Average Linkage Minimum Spanning Tree. Another variant of Linkage methods, not associated to a spanning tree representation, is the Complete Linkage (CL) [17].

The MST is the first but not the only correlation-based filtered network studied in literature. In particular the Planar Maximally Filtered Graph (PMFG) is a further step from the MST, that is able to retain a higher amount of information [3], [4], [19], having less strict topology constraint allowing to keep a larger number of links. The PMFG has been proven to have interesting pratical applications, in particular in the field of investment strategies to hedge risk [20].

Since the MST has associated a clustering method, and the PMFG is a generalization of the MST, it could be raised the question whether the PMFG provides a clustering method that exploits this higher amount of information. In a recent work [21] it has been shown that this is the case: the Directed Bubble Hierarchical Tree (DBHT) is a novel hierarchical clustering method that takes advantage of the

topology of the PMFG in order to yield a clustering partition and an associated hierarchy. (For the DBHT algorithm refer to supplementary material of [21] or, for a slightly modified version, to [22].) The approach is completely different from the agglomerative one adopted in the Linkage methods: the idea of DBHT is to use the hierarchy hidden in the topology of a PMFG, due to its property of being made of three-cliques [21], [23]. In [21] the DBHT hierarchical clustering has been applied to synthetic and biological data, showing that it can outperform many other clustering methods.

In this paper we present the first application of DBHT to financial data. To this purpose we have analysed the correlations among returns of 342 US stocks prices, across a period of 15 years (1997-2012). We have studied the structure of the clustering and we have compared the results with other clustering methods, as the Linkage ones and the k-medoids [24] (a partitioning method strictly related to the k-means [25]). The perspective of our study focuses not only on the clusterings, but on the entire hierarchies associated to those clusterings, covering all the different levels of the hierarchical structures. We have also studied the dynamical evolution of these structures, describing how the hierarchies change with time. The dynamical perspective is crucial for applications, in particular for what concerns hedging risk and portfolio optimization. For this reason we have given a particular attention to the effects of the financial crises on the hierarchical structures, highlighting differences among the clustering methods.

In order to compare quantitatively the amount of information retrieved by the different hierarchical clustering methods, we have taken the Industrial Classification Benchmark (ICB) as a benchmark community partition for the stocks and then we have compared it with the output of each clustering method. The idea is to use the degree of similarity between the ICB and the clustering as a proxy for the amount of information filtered by the methods. The degree of similarity is measured by using tools as the Adjusted Rand Index [26] and the Hypergeometric hypothesis test [27]. This is not the first work comparing correlation-based clusterings and industrial sector classification; however to our knowledge the comparison has been performed only qualitatively so far [28], [29]. In Ref. [30] the authors measured quantitatively the amount of filtered information without looking at the industrial sector classification by assuming a multivariate Gaussian distribution for the stocks returns [16]. Our approach is instead model-free. This is a relevant improvement since the multivariate Gaussian models are known not to be a suitable description for the returns [31], [32], and more in general the correlation-based networks obtained from real data have been found to be incompatible with some widespread models for asset returns [33].

The paper is organized as follows. In Section we present the dataset and some preliminary empirical analyses on it. In Section we perform a set of analyses on correlations and clusterings calculated taking the whole 15 years period as time window, hence ending up with only one hierarchical structure of dependences for each method. In particular we compare the clustering compositions in terms of ICB supersectors for different clustering methods and we measure the similarity between clusterings and ICB partition. In Section we perform instead analyses using a dynamical approach, with moving time windows. In Section we discuss the results and the future directions of our work.

Results

Dataset and preliminary analyses

The correlation structure that we have studied concerns $N = 342$ stocks from the New York Stock Exchange (NYSE). A complete description of the dataset is in Supplementary Informations (SI). We have analysed the closing daily prices $P_i(t)$ with $i = 1, ..., N$, during the time between 1 January 1997 to 31 December 2012 (4026 trading days). From the prices, we have calculated the daily log-returns [31], [32]:

$$r_i(t) \equiv log(P_i(t)) - log(P_i(t-1)). \quad (1)$$

From the set of N log-return time series over a time window T we have then calculated the $N \times N$ correlation matrix $\rho(T)$, whose elements are given by the Pearson estimator [34]:

$$\rho_{ij}(T) = \frac{\langle r_i(t) r_j(t) \rangle_T}{\sqrt{[\langle r_i^2(t) \rangle_T - \langle r_i(t) \rangle_T^2][\langle r_j^2(t) \rangle_T - \langle r_j(t) \rangle_T^2]}}, \quad (2)$$

where $\langle ... \rangle_T$ represents the average over the considered time window T. The clustering analysis is then carried out on the distance matrix D, with elements $D_{ij}(T) = \sqrt{2(1 - \rho_{ij}(T))}$ [1]. A weighted version of the Pearson estimator in Eq. 2, where terms in the average are multiplied by a weight $w_t = w_0 \exp(\frac{t - t_{end}}{\theta})$ according to their temporal distance from the last trading day (t_{end}) in the time window, has been used for the analysis on moving windows. This exponential smoothing scheme [35] allows to mitigate excessive sensitiveness to outliers in remote observations. The parameter θ has been set to $\theta = T/3$ according to criteria previously established [35].

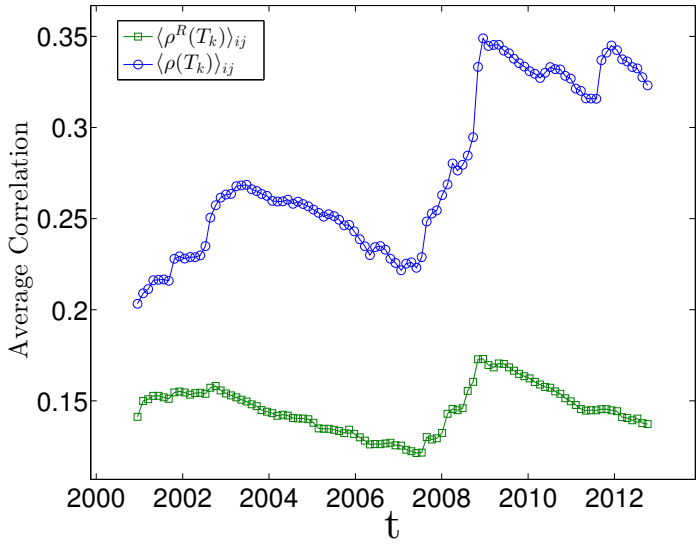

Figure 1. Demonstration that the average correlation evolves during time with large changes during periods of market instability. The figure reports the average correlation for each time window T_k with $k = 1, ..., n$, for both non- detrended (blue circles) and detrended log-returns (green squares). The average correlation is highly reduced by detrending the market mode.

By using this moving time window approach we have performed a set of preliminary studies on the average correlation of our set of stocks, looking in particular at the 2007-2008 financial crisis. To this aim we have considered a set of $n = 100$ overlapped time windows T_k ($k = 1, ..., n$) of length $L = 1000$ trading days (four years) for correlation matrices $\rho(T_k)$. In particular we have calculated the average correlation $\langle \rho(T_k) \rangle_{ij}$, that we show in Fig. 1 (blue circles). Analysing different L and n ($L = 750, 1250$; $n = 50, 80, 150$) we have verified that these results are robust.

To check the effect of the overall market mode on the results we have looked also at detrended log-returns, i.e. log-returns subtracted of the average return over all the stocks, and we have calculated the Pearson's correlations on it. In order to obtain these detrended log-returns we have considered a single factor model for each stock i [36]:

$$r_i(t) = \alpha_i + \beta_i I(t) + c_i(t), \quad (3)$$

where the common market factor $I(t)$ is the market average returns, $I(t) = \frac{1}{N}\sum_i r_i(t)$. The residuals, $c_i(t)$, are the log-returns detrended by the market mode. After estimating the coefficients α_i and β_i, the residuals $c_i(t)$ can be calculated and used to evaluate the new correlation matrix [36]. We call this matrix, estimated in the time window T_k, $\rho^R(T_k)$. We refer to the analyses based on this kind of correlation matrix as the "detrended case". These detrended correlation matrices are worth analyising since they have been found to provide a richer and more robust clustering [36] that can carry information not evident in the original correlation matrix [37].

In Fig. 1 it is shown the average correlation for these new correlation matrices, i.e. $\langle \rho^R(T_k) \rangle_{ij}$, compared to the average correlation for the original set of correlation matrices $\langle \rho(T_k) \rangle_{ij}$. As one can see, the subtraction of the market mode decreases by about 50% the average level of correlation, pointing out the important role of the market factor in the correlation structure. However, we can still observe the increase correspondent to the credit crunch. Moreover, and interestingly, the level of correlation reduces after a peak in 2009, unlike the non-detrended case. This fact suggests that, although the market mode plays an important role in terms of average amount of correlation, yet the peak of the credit crunch crisis seems not to be only a global market trend. Indeed it must involve, to some extent, the internal dynamics among stocks that remains after the subtraction.

Static analysis

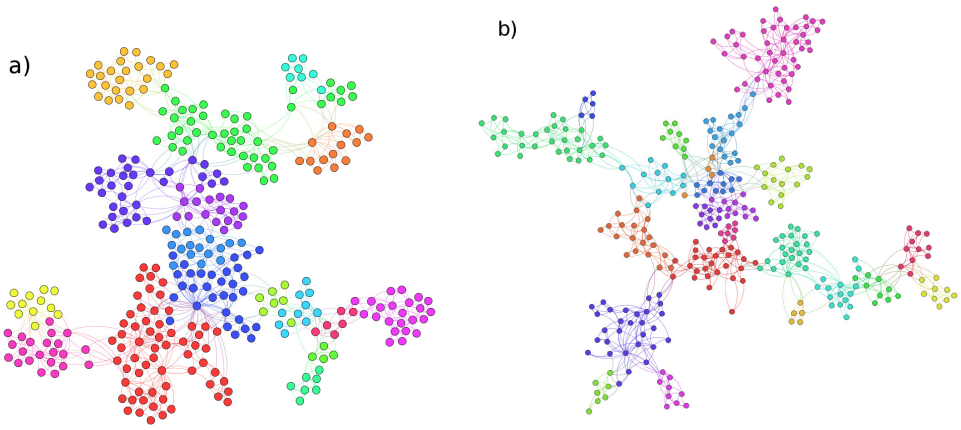

Figure 2. Visualization of the Planar Maximally Filtered Graph (PMFG) and DBHT clusters, for both non-detrended and detrended log-returns. a) PMFG calculated on the entire period 1997-2012, using non-detrended log-returns. Stocks of the same color belong to the same DBHT cluster. b) PMFG calculated on the same data as in a), but using detrended log-returns. Stocks of the same color belong to the same DBHT cluster.

DBHT clusters composition

In this section we present results of the PMFG and DBHT clustering method applied to the set of data described in the previous section (Section). In particular we have computed the PMFG and the

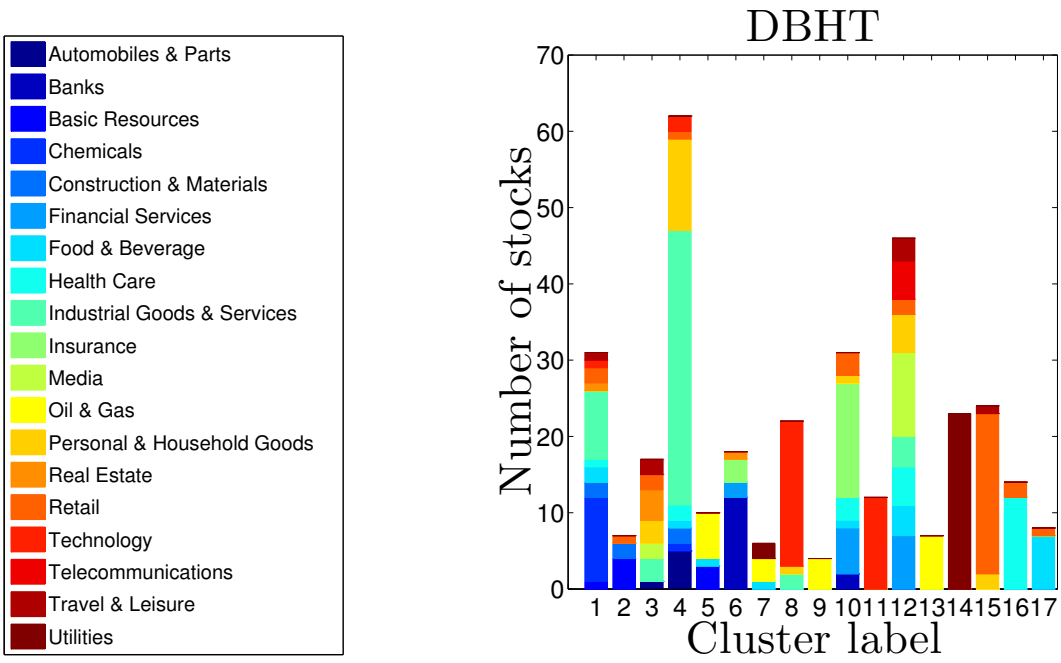

Figure 3. Number of stocks and composition of DBHT clusters in terms of ICB supersectors. The composition is shown by using different colours. See the Fig. S3 in Supplementary Information for the case detrended.

correspondent DBHT clustering in the time period from 1997 to 2012 and we plot it in Fig. 2 a) where we highlight, with the same color, stocks belonging to the same DBHT cluster. In the same figure (Fig. 2 b)) we plot the PMFG calculated by using the detrended log-returns (Eq. 3) as comparison. This PMFG looks more structured than the first one, with more homogeneous clusterings sizes.

We have then analysed the DBHT clustering structure in terms of industrial sectors. It is well known that the hierarchical structure of the stock return correlations shows a deep similarity with the industrial sectors categorization [1] [28] [37]. This fact supports the intuitive argument that stocks returns in the same industrial sector are affected mainly by the same flows of information. We can turn around the reasoning and claim that thus a desirable feature of a clustering method on stocks data is to retrieve, to some extent, the industrial sector classification. We will refer to the Industrial Classification Benchmark (ICB) [38]; this categorization divides the stocks in the market in 19 different supersectors, that in turn are gathered in 10 different Industries. For more details on the composition of our dataset in terms of ICB supersectors, refer to SI.

In Fig. 3 we report a graphical summary of the cluster compositions obtained applying the DBHT method to the whole time window 1997-2012 (that is, the clustering shown in Fig. 2 a)). The DBHT returns a number of clusters, N_{cl}, equal to 17: to each cluster is associated a bar, whose height represents the number of stocks in the cluster. Each bar is made of different colours, showing the composition of each cluster in terms of ICB supersectors. The legend on the left of the graph reports the corresponding industrial supersectors. Please note that the colours in Fig. 3 identify the ICB supersectors, and they have nothing to do with colours in Fig. 2, that identify DBHT clusters.

Cluster 4, the largest, is made of 62 stocks, accounting for about the 18% of the total number of

stocks; cluster 9, the smallest, contains 4 stocks. The average size of clusters is 20.1 stocks. As we can see, four clusters show a composition of stocks belonging to only one ICB supersector : cluster 9 and 13 (Oil & Gas), 11 (Technology) and 14 (Utilities). Similar cases are cluster 8, made of Technology stocks for more than 86%, cluster 15, within which 91% of stocks are from Retail, cluster 16 (75% of stocks from Health Care) and cluster 17 (87.5% of stocks from Food & Beverage). Moreover there are clusters that, although showing a mixed composition, are composed by supersectors strictly related: the number 6 is made of Banks, Financial Services and Insurance, all supersectors that the ICB gathers in the same industry (Financial) at the superior hierarchical step.

There are clusters that do not show an overexpression for a particular supersector or industry: this fact points out that the clustering is after all providing an information that cannot be reduced only to the industrial classification. In particular clusters 1, 3 and 12 have a heterogeneous composition, covering almost all the 19 supersectors and with no sector dominating the others. The cluster 4 is an intermediate case, since even though it overexpresses the Industrial Goods & Services (75%), it contains stocks belonging to 9 different supersectors. Interestingly the largest clusters (4, 12, 1 and 10) are all among these types of "mixed" clusters. We have performed the same study over the detrended log-returns, obtaining the PMFG shown in Fig. 2 b). The DBHT clusters are in this case 23 and show an higher overexpression of ICB supersectors, as expected since the market mode tends to hide the ICB structure [36]. In particular some supersectors that were mixed together in the non-detrended case are now overexpressed in distinct clusters: Chemicals, Insurance and Telecommunications. More details on the detrended case clustering are reported in Fig. 15 in SI.

Other clustering compositions

We here apply other clustering methods on the same data and compare results with DBHT clustering. The clustering methods considered are Single Linkage (SL), Average Linkage (AL), Complete Linkage (CL) and k-medoids. The latter is not a hierarchical clustering method, so it does not provide a dendrogram: however we analysed it to compare our results with a well established clustering method. The number of clusters, that unlike the DBHT, is a free-parameter for these methods, has been chosen equal to 17 in these cases, in order to compare the bar graphs with the Fig. 3 for DBHT. We plot in Fig. 4 a), b), c) and d) the clusters compositions obtained by using these four clustering methods, namely SL, AL, CL and k-medoids.

First of all we can observe that for each of them there is a strong heterogeneity in the size of clusters: SL and AL display two huge clusters of 323 and 322 stocks respectively (almost identical, having 318 stocks in common), with the other clusters made of one, two or three stocks. For both the algorithms this giant cluster contains stocks of all ICB sectors.

For the CL and the k-medoids the situation is quite different. For the CL, the giant cluster (cluster number 10) is much reduced in size (136 stocks), with also other three clusters (the number 12, 9 and 5) containing a relevant number of stocks (50, 33 and 25 respectively): the main supersectors that are overexpressed are Technology (cluster 12), Utilities (cluster 5), Retail (cluster 9), Oil & Gas (cluster 16) and Health Care (cluster 2). A very similar structure occurs with the k-medoids, but with the giant cluster splitting further in two large clusters (7 and 10). However the DBHT clustering is the one showing the largest degree of homogeneity in size and overexpression of ICB supersectors, at least for this number of clusters (see Fig. 3 for comparison).

These first comparisons are however made under a specific choice of the number of clusters (17), given by the DBHT. One could wonder what happens changing this parameter, i.e. moving along the hierarchical structure provided by each clustering method. Let us stress out that the DBHT method gives automatically the number of clusters that is instead an adjustable parameter for the other methods. However, DBHT can also be analysed for a varying number of clusters by thresholding over the clustering hierchical structure. In the following Sections and we discuss a set of quantitative analyses that explore all the hierarchical levels of the DBHT and the other clustering methods.

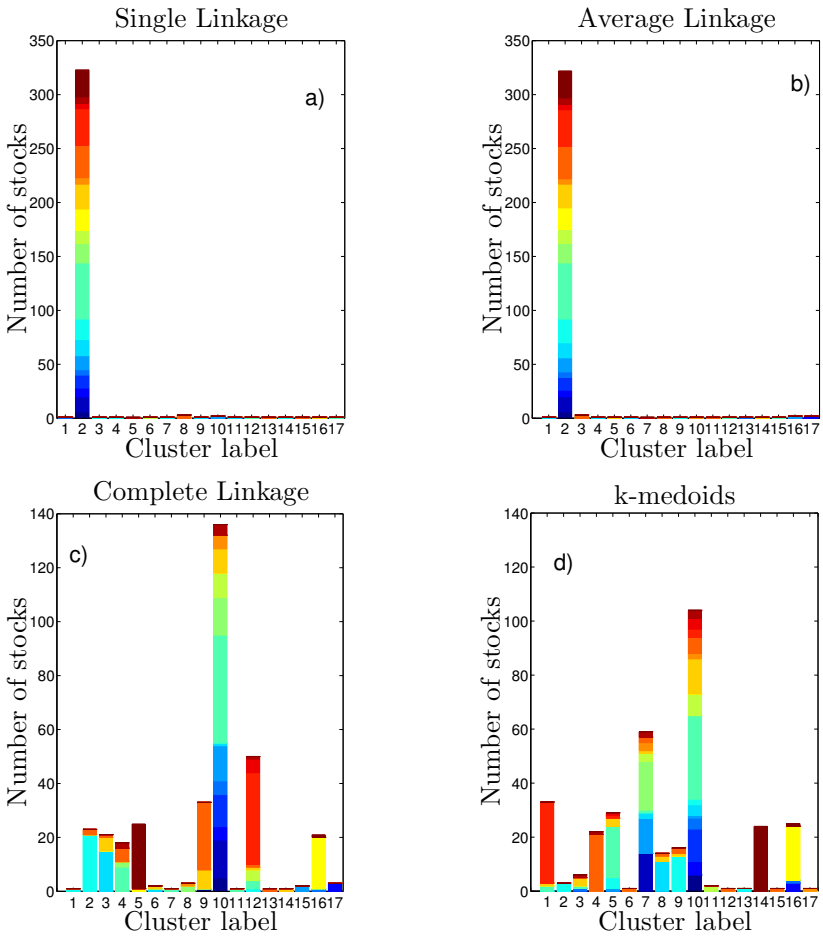

Figure 4. Composition of clustering in terms of ICB supersectors, for different clustering methods. The x-axis represents the single cluster labels, the y-axis the number of stocks in each cluster. Each colour corresponds to an ICB supersector (the legend is the same as in Fig. 3. The graphs a) show the results for SL clustering, b) for AL, c) for CL and d) for k-medoids. See Fig. 16 in SI for the case detrended.

Disparity in the clusters size

In Section we have seen that the SL and AL clustering methods show a giant cluster that contains more than 90% of the stocks, whereas DBHT, CL and k-medoids methods have a more homogenous distribution of cluster sizes.

Let us here check whether this difference in the structure depends on the choice of the number of clusters for the linkage methods, which might be penalising the Linkage methods with respect to DBHT. To do that, we vary the number of clusters N_{cl} for each clustering method by cutting the dendrograms at different levels. For the k-medoids, for which no dendrogram is present, N_{cl} is simply an input parameter of the algorithm. We then calculate the following measure of disparity:

$$y = \frac{\sigma_S}{\langle S \rangle}, \tag{4}$$

where σ_S is the standard deviation of clusters size and the normalization factor $\langle S \rangle$ is the average, given by:

$$\sigma_S = \sqrt{\frac{1}{N_{cl} - 1} \sum_a (S_a - \langle S \rangle)^2}, \tag{5}$$

and

$$\langle S \rangle = \frac{1}{N_{cl}} \sum_a S_a, \tag{6}$$

with S_a being the size of (number of stocks in) cluster a. In the limit of homogeneous arrangement of stocks among the clusters (i.e., each cluster has the same number of stocks), we obtain $\sigma_S = 0$ and then $y = 0$. The higher is the degree of heterogeneity in the distribution of sizes, the higher is σ_S and therefore y.

In Fig. 5 we show, for each clustering method, how the disparity measure varies with N_{cl}. The graph a) shows the non-detrended case, the graph b) the detrended case. As we can see the SL provides the higher disparity in both cases, regardless of N_{cl}, then the AL, CL and k-medoids follow. The DBHT values are below all of them, this means that the DBHT clustering provides a more structured community assignment at any level of the correlation hierarchy. Moreover, in the market mode case the SL and the AL show the highest values of disparity for N_{cl} in the interval 50-100. The CL and DBHT have instead a flatter pattern, with the highest values occurring for lower values of N_{cl}. Looking at the detrended case in Fig. 5 b), the removal of the market mode smooths also the pattern of the AL, whereas the SL is even sharper. Overall, subtracting the market mode makes the clusterings more homogeneous, suggesting that the largest clusters that emerged in the non-detrended case are associated to the market mode dynamics.

We can conclude that, from the point of view of the disparity measure, the clustering methods provide quite different structures at any level of the dendrograms. The DBHT yields the most homogeneous clustering, whereas the SL displays high levels of disparity, even after the market mode subtraction.

Retrieving the industrial sectors

In this section we quantify the similarity between clustering and ICB, by varying the number of clusters N_{cl}. An industrial sector classification as the ICB is nothing but a partition in communities of the N stocks. Therefore, in order to provide a quantitative measure for the similarity between clustering and ICB, we can use the tools conceived to compare diffent clusterings on the same set of "objects" [39]. In particular we have used the popular Adjusted Rand Index ($\mathcal{R}_{adj}$) [26] that, given two clusterings on the same set of items, provides a numerical value equal to 1 for identical clusterings and to 0 for clusterings completely independent. The idea behind this measure is to calculate the number of pairs of objects that are in the same cluster in both clusterings, and then to compare this number with the one expected under the hypothesis of independent clustering (the formal definition is given in).

In Fig. 6 we show the results of the first set of analyses carried out by using the Adjusted Rand Index. We have applied the five clustering methods to the entire-period time window, and then we have varied the number of clusters N_{cl}. For each clustering obtained in this way, we have calculated the Adjusted Rand Index $\mathcal{R}_{adj}$ between that clustering and the ICB supersector partition. In this way we have explored the entire hierarchical structure.

Fig. 6 a) refers to the non-detrended case, Fig. 6 b) to the detrended case. The vertical dashed line in the graphs identifies the value $N_{cl} = 19$, that is the number of ICB supersectors. For all the methods we observe an increasing trend for low values of N_{cl}, a maximum and then a decreasing trend toward zero as

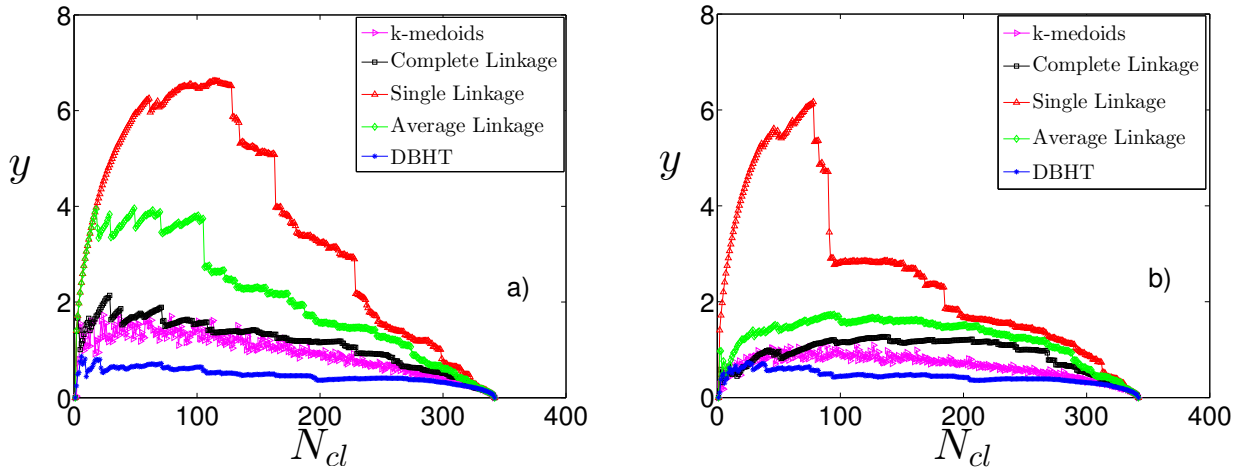

Figure 5. Demonstration that different clustering methods show different degrees of disparity in the clustering structure. The disparity measure y is shown for clusterings at different hierachical levels as function of N_{cl} in the dendrograms, for a) non-detrended log-returns and b) detrended log-returns.

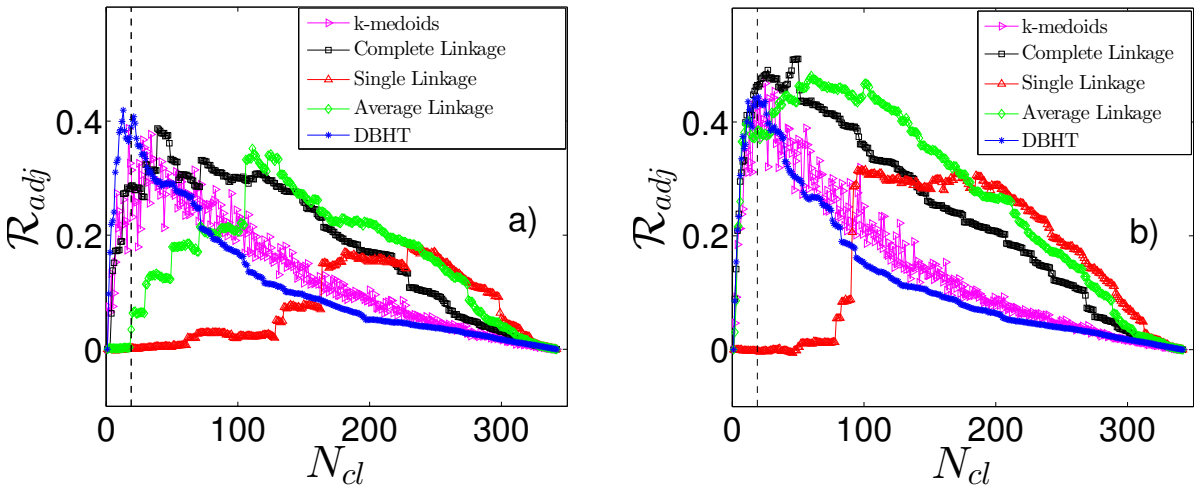

Figure 6. Demonstration that different clustering methods retrieve different amount of industrial sector information. The Adjusted Rand Index $\mathcal{R}_{adj}$ between clustering and ICB supersectors is shown for different number of clusters N_{cl}. In a) correlation are calculated on non-detrended log-returns, in b) are calculated on detrended log-returns. The vertical dashed line shows the value ($N_{cl} = 19$) correspondent to the actual number of ICB supersectors.

N_{cl} goes to 342. However the five methods show differences for what concerns the value of the maximum and its position. In the non-detrended case, we find that the highest values of Adjusted Rand Index, $\mathcal{R}^*_{adj}$, are reached by DBHT (0.419), k-medoids (0.387) and Complete Linkage (0.387). Interestingly these three values are quite close to each other, maybe indicating this level as the actual maximum similiarity between correlation clustering and ICB supersectors. However the numbers of clusters correspondent to these three maxima (N^*_{cl}) are different, respectively 13, 17 and 39. It is worth noticing that the maximum for the DBHT and k-medoids occurs very close to the "real" ICB supersectors N_{cl} value indicated by the dashed line. However, the k-medoids values are sensitively more fluctuating than the two hierachical methods. The Average Linkage and Single Linkage reach instead much lower $\mathcal{R}^*_{adj}$ (respectively 0.352 and 0.184) and much higher N^*_{cl} (respectively 111 and 229).

For what concerns the detrended case, we notice first of all that the maximum values of $\mathcal{R}_{adj}$ increase for all the methods. The natural explanation for this is that the market mode, driving all the stocks regardless of their industrial supersector, hides to some extent the ICB structure [36]. The CL shows now the highest degree of similarity (0.510), followed by the AL (0.48, showing the most remarkable increase with respect to the non-detrended case), the k-medoids (0.467) and DBHT (0.444). The SL is again the last one in the ranking (0.315). The ranking in N^*_{cl} is instead equal to the market mode case: lowest N^*_{cl} for DBHT (20), followed by k-medoids (25), Complete Linkage (50), Average Linkage (60) and Single Linkage (101). Let us stress that, although the DBHT has not the highest $\mathcal{R}_{adj}$ in this case, its maximum is the closest to the real N_{cl} of 19. In general the effect of the market mode subtraction on N^*_{cl} changes according to the clustering method: for the DBHT, CL and k-medoids the subtraction raises N^*_{cl}, whereas for AL and SL the same quantity reduced remarkably. The effect is more pronunced for the Linkage methods, less for DBHT and k-medoids.

Overall we can conclude that varying the number of clusters (N_{cl}) the DBHT outperforms the other clustering method at retrieving the ICB information. DBHT, k-medoids and CL have however quite close values of $\mathcal{R}^*_{adj}$, but reached at different N^*_{cl} values. The DBHT and k-medoids are able to retrieve the ICB information at a N_{cl} that is both lowest and closest to the actual number of ICB supersectors (19). After subtracting the market mode also the AL reaches the same level of DBHT, k-medoids and CL, but at too high N_{cl}.

Interestingly, the Average and Single Linkage methods have the clusterings with both the lowest values of $\mathcal{R}_{adj}$ and the highest disparity values y (Section): i.e., the higher the disparity y is, the less the clustering method is able to retrieve the industrial classification. This could be due to the presence of a large cluster when y is very high: this might indicate a strong sensitivity to the market mode, that hides the intrasector correlations merging many stocks in a single cluster.

Industries overexpression

The Adjusted Rand Index provides an overall measure of similarity between the clustering and the ICB partition. In order to analyse how much each industrial sector is retrieved by the clustering we need a different approach that allows a direct comparison between clusters and industrial sector. We have therefore compared each cluster to each industrial sector, analysing the number of stocks that are within both of them. In this way we can measure for each industrial sector the level of overexpression by the clustering. Let us, in the following, describe the procedure we have applied to carry out such analysis.

In order to have a smaller number of ICB communities to compare with the clusters, let us consider here the ICB industries instead of the ICB supersectors. This means looking at the higher hierarchical level in the ICB classification, that gathers the stocks in 10 different industries. Let us remark that we have performed these analyses also on ICB supersectors, and the results are comparable (see for the results). We say that a cluster overexpresses a specific ICB industry if the percentage of stocks in common between them is sensitively higher than what expected from a random overlapping of communities. It is possible to give a precise meaning to "sensitively higher" by using a statistical one-tail hypothesis test, where the null hypothesis is the Hypergeometric distribution [40]. This distribution is a generalization

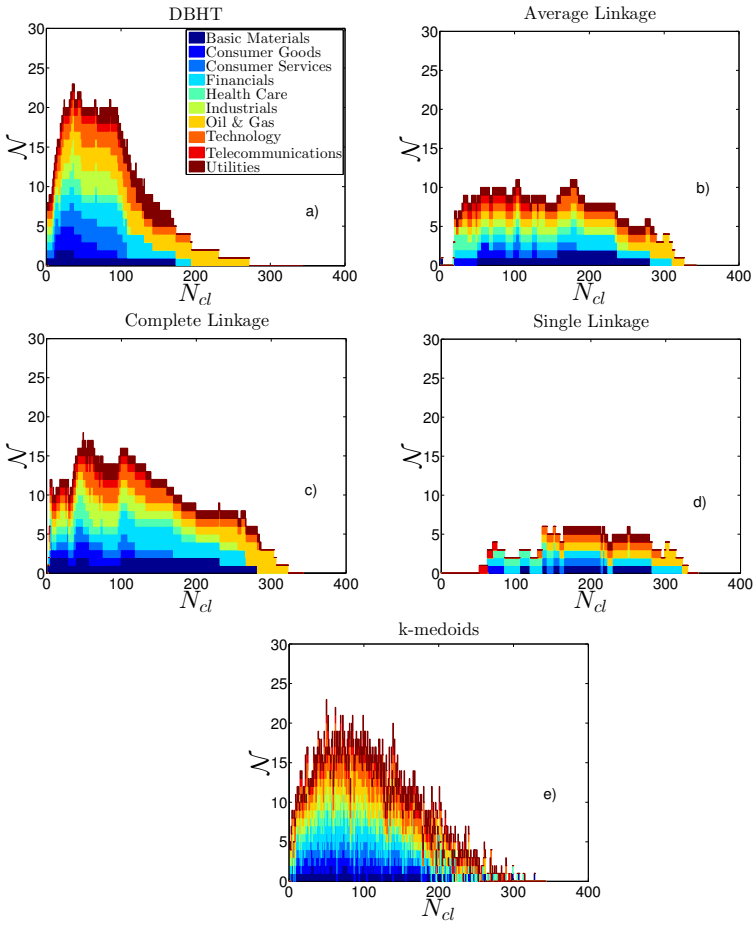

Figure 7. Amount of ICB information retrieved by the clustering methods, in terms of ICB industries overexpressed by each cluster. Each bar graph shows, varying the number of clusters N_{cl}, how many times ($\mathcal{N}$) an ICB industry is overexpressed by a cluster according to the Hypergeometric hypothesis test (i.e., number of tests being rejected). Each colour shows the number of overexpressions for each ICB industry. In graphs a)-e) the results for DBHT, AL, CL, SL and k-medoids clustering are shown respectively. The correlations are calculated on non-detrended log-returns. See Fig. 17, 18 and 19 in SI for the detrended cases and with ICB supersectors.

of the Binomial distribution and describes the probability $P(k)$ that the number of objects in common between two overlapping and independent communities is equal to k, taking into account the size of the two communities and the overall number of objects. More details are in .

Varying the number of clusters N_{cl} we have performed this one-tail hypothesis test for each clustering. We have chosen a significance level for the test equal to 0.01, together with the conservative Bonferroni correction [40] (see for more details). In Fig. 7 we show with bar graphs the results of these analyses for each one of the five clustering methods, by using non-detrended log-returns. The x-axis shows N_{cl}, whereas $\mathcal{N}$ on the y-axis represents the total number of hypergeometric tests that have been rejected for

that value of N_{cl} (i.e. how many times an ICB industry has been found to be overexpressed by a cluster). The colors on each bar show the number of overexpressions for each different industry.

As we can see, the compositions and trends with N_{cl} are quite different among different methods. The DBHT shows the highest number of clusters overexpressing ICB industries, followed by k-medoids, CL, AL and SL. Not surprisingly, AL and SL show the worst performance also for the Adjusted Rand Index analysis in Fig. 6 a). In that analysis however the DBHT, k-medoids and CL were showing very close Adjusted Rand Index values; the hypothesis test is therefore able to highlight better the differences between these methods. For what concerns the industry composition, we can see that the DBHT, k-medoids, CL and AL show a quite homogenous composition, with almost each ICB industry overexpressed. The SL instead shows a much less rich composition, with not more than 6 overexpressed industries simultaneously even at the maximum level of total overexpressions.

In terms of trend of the $\mathcal{N}$ curve, the DBHT has a quite peaked shape, quickly dropping to low values. The three Linkage methods are instead flatter and more spread along the N_{cl} axis, a feature that was evident also in the trend of $\mathcal{R}_{adj}$ in Fig. 6. The k-medoids seems to be a mix between these two shapes, showing however a much higher level of noise and instability in the ICB composition when N_{cl} changes.

Finally, it is worth noticing that there is a change in the composition at different values of N_{cl}, and that similar patterns can be found among the four hierarchical clustering methods (for the k-medoids no clear patterns can be found, because of the higher level of instability of the method). There are industries that tend to become overexpressed for low values of N_{cl} and then disappear at intermediate values: this is the case of Consumer Goods, Telecommunications and Health Care. Others are instead more persistent, appearing along all the x-axis: Utilities, Technology, Financial, Oil & Gas. The most persistent is the latter, that is still overexpressed when all the other industries are not expressed anymore. We can then conclude that not only the ICB partition is hidden at different levels in the dendrograms (Section) depending on the clustering method, but also different ICB industries are retrieved at different levels. This is probably due to the different degrees of correlation within different ICB industries.

By using returns detrended of the market mode we obtain quite similar results, apart from an overall increase in $\mathcal{N}$ for all the methods, consistently with what found in Section . The details are shown in Fig. 17 in SI.

Dynamical analysis

Here we present a dynamical analysis of the DBHT clustering in the 15 years ranging from 1 January 1997 to 31 December 2012. We have selected the set of overlapping time windows described in Section ($n = 100$ time windows of length $L = 1000$ trading days) and used a weighted version of the Pearson estimator (Eq. 2 in Section) in order to mitigate excessive sensitiveness to outliers in remote observations.

In Fig. 8 a) is shown the number of DBHT clusters obtained for each time window, both for non-detrended log-returns (red circles) and detrended log-returns (blue squares). For the first case the number of clusters ranges between 6 and 19, for the second case the range is 14-26. The dashed lines are the values correspondent to the clustering obtained using the entire period 1997-2012 as time window. A study of the statistical robustness of these clusterings have been carried out by means of a bootstrapping approach: it turns out that the numbers of clusters shown in Fig. 8 a) is robust against resampling of the log-returns time series. The results are reported in and in Fig. 12.

As observed previously, the number of clusters in the non-detrended case is sistematically lower than the detrended case. Moreover, an overall decreasing trend characterizes the non-detrended values and makes them go below the corresponding dashed line; this decreasing pattern is not present in the detrended case, that however stays below the correspondent dashed line the most of the times either. It is interesting also to analyse the evolution of the disparity y, introduced in Eq. 4, over the period. In Fig. 8 b) we show y for each time window, both for the non-detrended and detrended case. Again the dashed lines are the values for the all period clusterings. In the non-detrended case we see an overall increasing trend, especially after the 2006; an analysis of the sizes distribution show that largest cluster

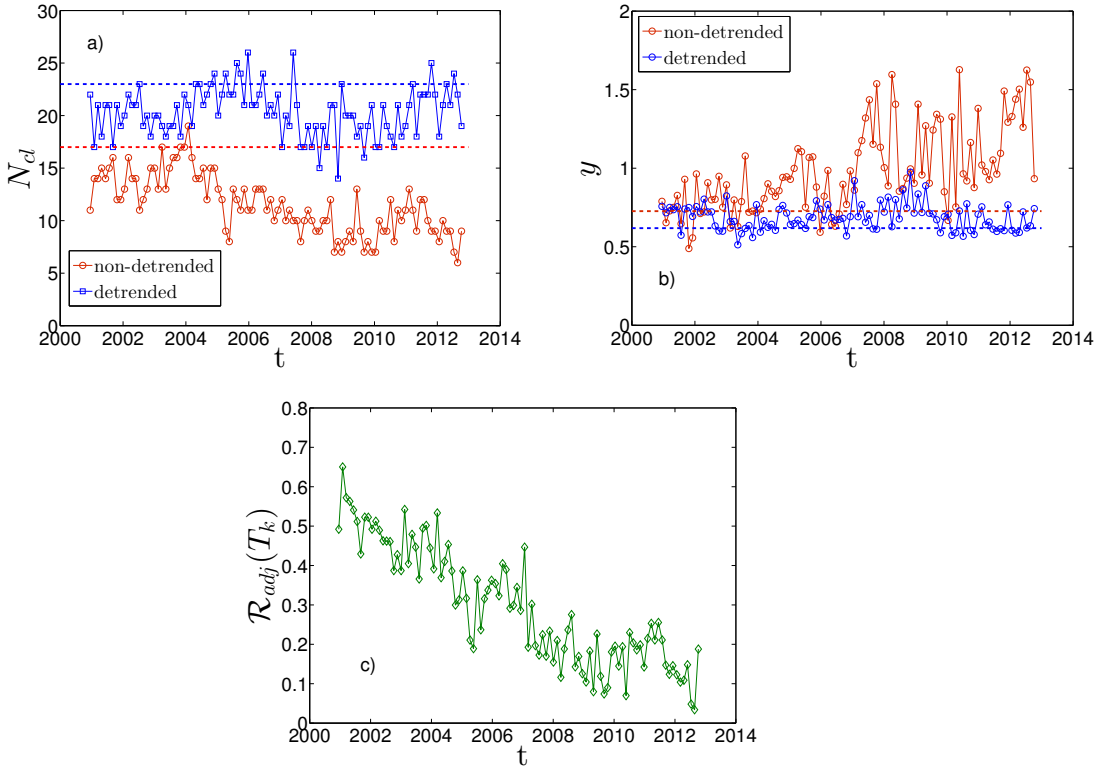

Figure 8. Dynamical evolution of the DBHT clustering. Each plot refers to 100 moving time windows of length 1000 trading days. Specifically, in graph a) we plot the number of DBHT clusters, N_{cl}, for both log-returns non-detrended (red circles) and detrended by the market mode (blue squares), whereas the two dashed horizontal lines are the N_{cl} values obtained by taking the largest time window of 4026 trading days. Overall the non-detrended case shows a decreasing trend. In graph b) it is shown the disparity measures, y, again for the two sets of DBHT clustering (red dots non-detrended, blue dots detrended), the dashed horizontal lines being the y values from the 4026 length time window. In the non-detrended case the 2007 marks a transition to higher and more volatile values of y. Finally in graph c) it is shown the Adjusted Rand Index, $\mathcal{R}_{adj}$, measured at each time window between the detrended and non-detrended clusterings. A steady decreasing trend is evident.

contains up to 240 stocks (70% of total number of stocks); moreover, from 2006 on we observe also a much higher fluctuation in the values. This behaviour is of interest since it concerns the overall influence of the market mode on the correlation structure, with higher y indicating a stronger influence of the market mode that tends to gather all stocks in one cluster. Indeed, in the detrended case we find that subtracting the market mode makes the increasing trend disappear. Overall the disparity values decrease and stay closer to the dashed line, without significant pattern apart from some fluctuations.

In order to better understand the relation between the DBHT clusterings obtained with detrended and non-detrended log-returns, we have also performed a dynamical Adjusted Rand Index analysis. Now we compare no longer the clustering and the ICB partition, but the two clusterings (non-detrended and detrended) at each time window. In Fig. 8 c) the Adjusted Rand Index between the two sets of DBHT clusters is shown. Interestingly, it appears a steady decreasing trend that drives the similarity

from relatively high values (about 0.7) to values close to zero, indicating complete uncorrelation between the two clusterings. We can therefore conclude that the influence of the market mode has increased remarkably over the last 15 years, making the detrended clustering structure more and more different from the non-detrended one. This observation would have not been possible without the clustering analysis, since from the preliminary dataset measures (see Fig.1 and Fig. 13 in SI for details) it is not evident any constant pattern either in the average return or in the average correlation.

Dynamically retrieving the industrial sectors

Let us here investigate the relation between industrial classification and clustering under a dynamic perspective. To this end we here perform the previous dynamical analysis by considering the set of 100 overlapping time windows T_k and calculating for each of them the Adjusted Rand Index $\mathcal{R}_{adj}(T_k)$ between clustering and ICB supersectors classification. Since $\mathcal{R}_{adj}(T_k)$ varies with the chosen threshold and N_{cl}, we select at every time the N_{cl} that maximizes $\mathcal{R}_{adj}(T_k)$; the numbers that we report are these maximum values and account therefore for the maximum ability of the clustering methods to retrieve the ICB.

In Figs. 9 a)-e) we show the results for each of the five clustering methods, using returns with market mode. Interestingly, all of them show a decreasing trend across the time. On average, the DBHT and CL display the highest similarity with industrial classfication, whereas the Single Linkage the lowest. This is consistent with what found in the static analyses. We have also highlighted in the graphs the major events that affected the stock market in the last 15 years. It can be observed that different clustering methods are affected in different ways by these events. Indeed, if the 2007-2008 credit crunch crisis and the following recession is evident in all the methods as a significant drop in the similarity, other events as 11/09/2001 or the 2002 stock market downturn appear only in the Single and Average Linkage plots. In particular the 2002 downturn drives a steep decrease in the similarity of SL and AL, that stay at low values until the end of 2005. For DBHT, Complete Linkage and k-medoids instead these events do not seem to affect the similarity in a noticeable way compared to the statistical fluctuations. This observation points out that the DBHT, CL and k-medoids are more robust than SL and AL against exogenous events in their ability to retrieve an economic information as the industrial classification. Nonetheless, there are differences also among DBHT, CL and k-medoids: in particular in the period following the 2008 crisis, DBHT and k-medoids show a peak that does not appear with CL. Moreover, for the k-medoids the drop in similarity seems to start more than one year before the 2007. All these features have non-trivial implications for both portfolio optimization and systemic risk evaluation. Implications that we plan to investigate in future works.

Fig. 9 f) shows the number of clusters N_{cl} that maximizes, in each time window T_k, the Adjusted Rand Index shown in the previous plots. As we can see, N_{cl} for SL is always the highest, followed by AL, CL, k-medoids and DBHT. This is consistent with what we found in the static analysis in Section : different clustering methods "hide" the industrial information at different levels of the hierarchy. SL and AL, that yield higher N_{cl} (i.e., lower levels in the hierarchy), are also the methods that show the lowest level of similarity with industrial classification and the highest degree of disparity (see Subsection).

In Figs. 10 a)-f) we show the same set of plots for the detrended case. The main differences with the non-detrended case are the following:

- the average similarity with the industrial classification rises for all methods; this confirms in the dynamical case what we found for the static case;
- the average N_{cl} is lower for all methods: the absence of market mode "moves" the industrial classification to higher levels of the hierarchy;
- the strong influence of the 11/09/2001 and 2002 downturn on the SL and AL pattern seems to disappear, whereas the 2007-2008 crisis is still evident in all the five methods. This could be

explained claiming that the former are global events in the market, whereas the latter exhibits also a "local" dynamics;

- the AL shows the most evident change in the dynamical behaviour, displaying a trend much more similar to the DBHT and CL one. Also in terms of N_{cl}, it shows values closer to DBHT, CL and k-medoids than SL. A similar observation was made in the static analysis in Section , where the AL turned out to perform like DBHT, CL and k-medoids once the market mode was removed.

Alluvional diagram visualisation

To better highlight the changes in the clustering induced by the 2007-'08 credit crunch we have investigated the dynamics of each DBHT cluster over the period 1997-2012 and we have visualised such evolution by means of a so called alluvial diagram [41]. These diagrams are tools for visualising the evolution of communities in large networks, and have been introduced in Ref. [41]. The diagram in Fig. 11 shows the evolution of DBHT clusters from January 1997 to December 2012 and has been generated by using the Alluvial generator Flash applet created by M. Rosvall and C.T. Bergstrom [41], available online on http://www.mapequation.org/apps/AlluvialGenerator.html. Each colour represents a different ICB industry. Each vertical module represents the DBHT clustering at the time window specified on the bottom of the diagram (the very first vertical module being only the partition in different ICB industries). The number of clusters for each clustering method has been fixed to 8 at any time window, in order to have a diagram easier to visualize. It is important to note that the height of each cluster is proportional to the flux (number of stocks) entering into the cluster, and not to the number of stocks in that cluster. Each time window contains 1006 trading days; the third one covers the beginning of the 2007-2008 credit crunch.

As we can see, Utilities and Oil & Gas do not display any tendency to gather with other ICB industries, at most mixing together in the two central time windows. On the other hand, industries such as Financial, Industrial, Consumer Goods and Consumer Services are much spread among different clusters. Consumer Goods and Consumer Services in particular show the highest degree of diversification, being present in more than 5 clusters at each time. Technology is instead an intermediate case, that is well clustered on its own in the first two time windows, but then split in three different strands when the crisis happens, staying even more divided in the last time window.

We note that the 2007-2008 credit crunch drives clear changes in the clustering structure. In particular a cluster with high heterogeneous composition appears (the second from the bottom, third time window), with an almost equal percentage of Financial, Consumer Services and Consumer Goods. In this cluster also a small part of Technology industry converges, as we pointed out earlier. This new composite cluster is probably related to the high global trend driven by the credit crunch, that has changed the clustering structure that was instead quite stable over the previous 8 years. Interestingly, with the crisis another new cluster emerges, merging Industrial and Basic Materials. At the same time, the Utilities and Oil & Gas industries, as we said, stay separated from the others. This again highlights the peculiar nature of the 2007-2008 crisis, characterized by a driving factor that do not affect all the ICB industries. All these changes can be observed still in the last time window, indicating that the dynamic driven by the crisis is still affecting the market structure over all the 2012, and this change is perhaps not reversible.

Discussions

In this paper we have presented a set of static and dynamical analyses to quantify empirically the amount of information filtered from correlation matrices by different hierarchical clustering methods. By taking the Industrial Classification Benchmark (ICB) as a benchmark community partition we have been able to perform quantitative analyses on real data without any assumption on the returns distribution.

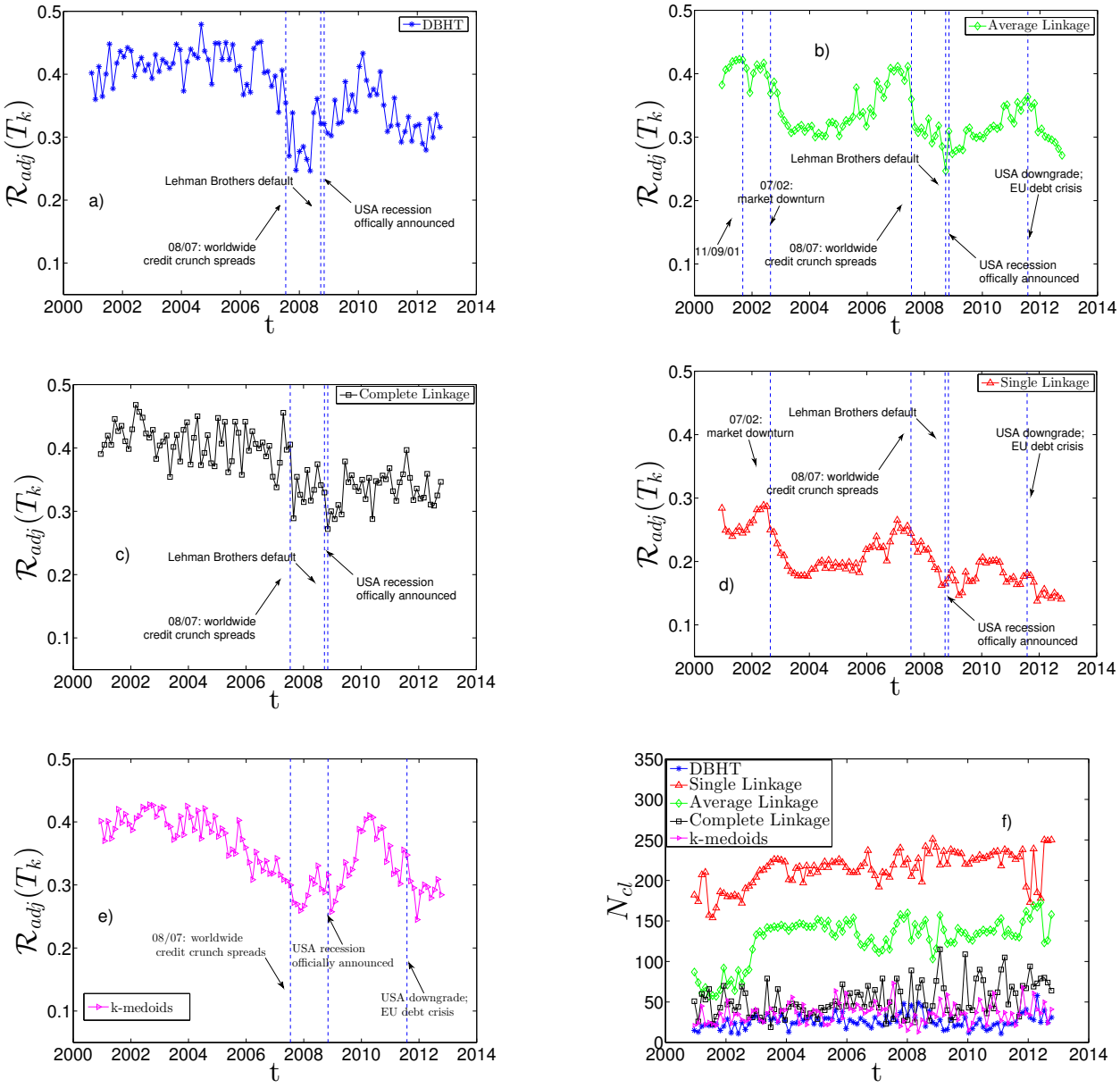

Figure 9. Dynamical evolution of the similarity between clustering and ICB. It is shown the Adjusted Rand Index, $\mathcal{R}_{adj}$, calculated at each time window T_k $(k = 1, ..., n)$ between clustering and ICB partition, for a) DBHT, b) AL, c) CL, d) SL and e) k-medoids method. A drop in the similarity occurs for all the methods during the 2007-2008 crisis. The AL and SL show decreases also during other financial events. At each time window the number of clusters, N_{cl}, has been chosen in order to maximize the $\mathcal{R}_{adj}$ itself: in f) we plot these N_{cl} values for each clustering method. It is evident as the maximum similarity clustering-ICB is reached at different hierarchical levels depending on the clustering method. The correlations are calculated on non-detrended log-returns.

In particular we have considered three variants of Linkage methods (Single, Average and Complete) and the k-medoids, and we have compared them with the Directed Bubble Hierarchical Tree (DBHT), a

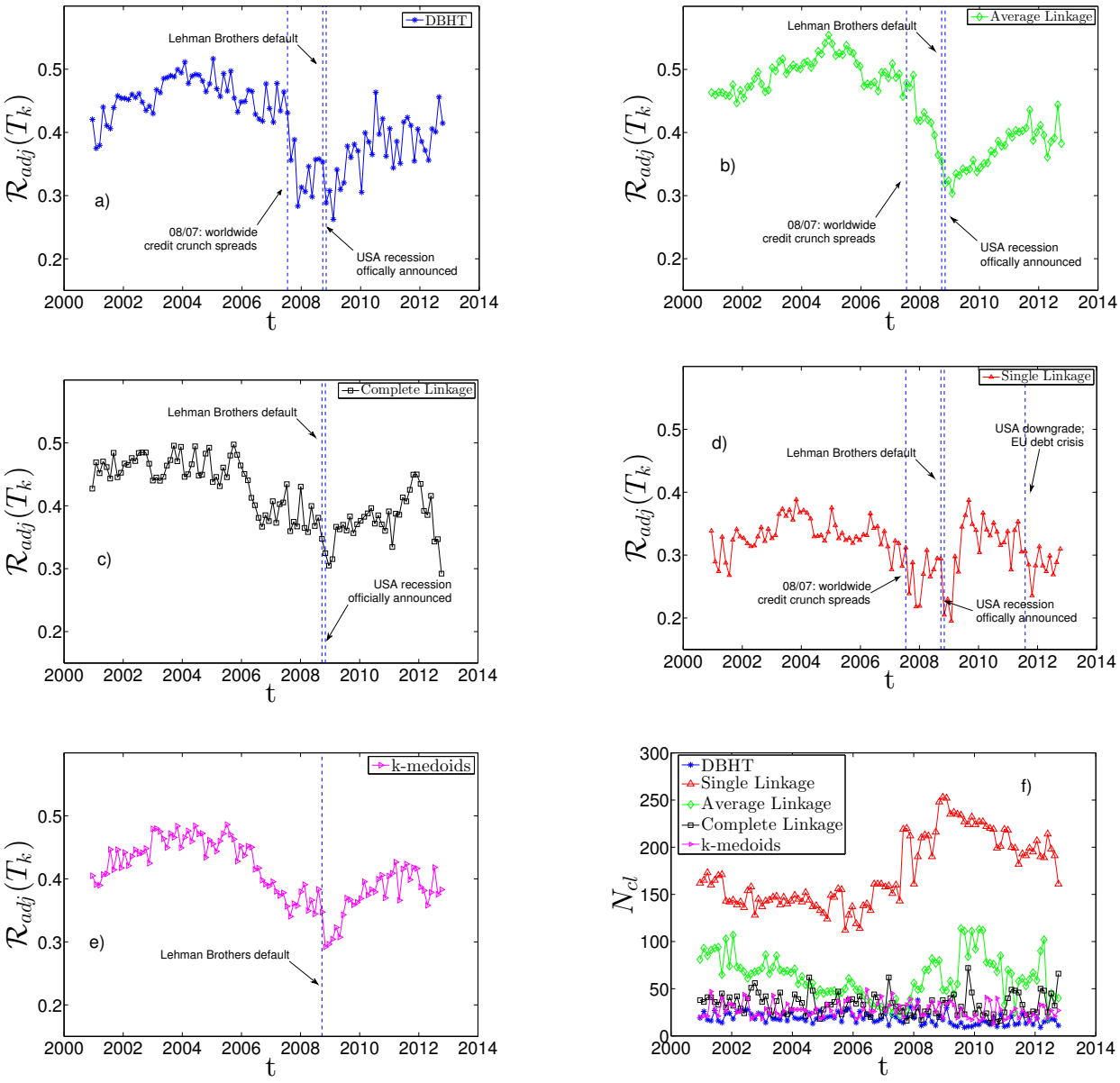

Figure 10. Dynamical evolution of the similarity between clustering and ICB, with detrended log-returns. a)-f): same graphs as in Fig. 9, but by using correlations on detrended log-returns.

novel clustering method applied here for the first time to financial data. Our analyses have shown that the DBHT is a more suitable clustering method than the Linkage and the k-medoids for financial data, being able to retrieve more information with fewer clusters.

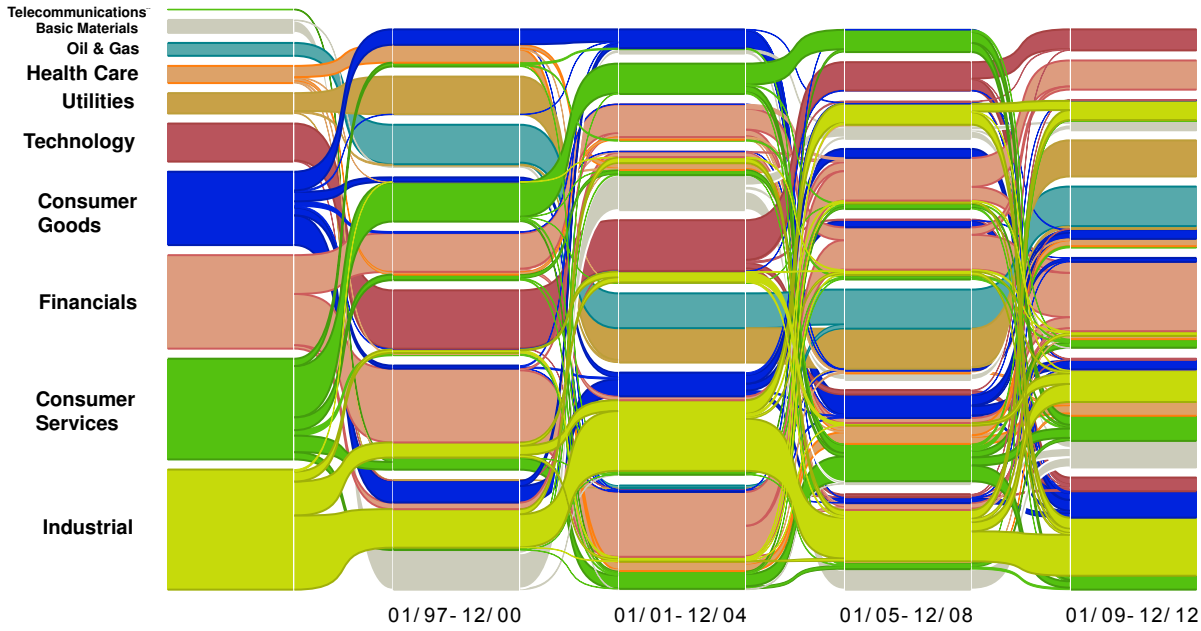

Figure 11. Alluvial diagram showing the evolution of DBHT clusters from January 1997 to December 2012. The first column shows the partition of the stocks according to their ICB industry. The other columns represent the clustering calculated over the periods reported on the bottom. Different colours highlight stocks belonging to different ICB industries.

The methods show remarkably different performances in retrieving the economic information encoded in the ICB, with big dissimilarities even among the Linkage methods. Moreover, the economic information appears to be hidden at different levels of the hierarchical structures depending on the clustering method. The DBHT and k-medoids methods show the best performances, but the latter seems to be affected by the noise much more than the DBHT and the Linkage methods. The DBHT turns out then to be a good mix between the advantages of the k-medoids and those of the Linkages. The dynamical analysis has also proved that the methods show different degrees of sensitivity to financial events, like crises. This is again a new result that could give insights into the dynamics of such events, as well as an indication on which clustering method is more robust for financial applications.

To check the robustness of the results we performed each analysis also on log-returns detrended by the market mode, by following a standard procedure in literature [36], [37]. Interestingly the effect of this subtraction is very dissimilar for different methods, with the weakest methods (Average and Single Linkage) improving remarkably their performance. However the main results are robust against the subtraction of the market mode.

In future works we plan to extend the present study to other datasets, covering different periods and different stock exchanges and considering other measures of dependences including non-linear dependences such as the Kendall's rank correlation [42] and the Mutual information [43]. Finally, since correlation-based networks and clustering methods have shown to be useful tools for portfolio optimization [20], [44], [45], we also plan to use these new insights into the hierarchical structures to improve further the current performances of portfolio optimization tools.

Methods

Measuring clustering similarity: Adjusted Rand Index

Following the notation of [39], let us call X the set of the N objects (stocks, in our case). Y is a partition into communities of X or simply a clustering, that is "a set $Y = \{Y_1, ..., Y_k\}$ of non-empty disjoint subsets of X such that their union equals X". Let us say we have also another different clustering Y': we call *contingency table* the matrix $M = \{m_{ij}\}$ where

$$m_{ij} \equiv |Y_i \cap {Y'}_j|, \tag{7}$$

i.e. the number of objects in the intersection of clusters Y_i and ${Y'}_j$. Let us call a the number of pairs of objects that are in the same cluster both in Y and in Y', and b the number of pairs that are in two different clusters both in Y and Y'. Then the Rand Index is defined as the sum of a and b, normalized by the total number of pairs in X:

$$\mathcal{R}(Y, Y') \equiv \frac{2(a+b)}{N(N-1)} = \sum_{i=1}^{k}\sum_{j=1}^{l}\binom{m_{ij}}{2}. \tag{8}$$

As null hypothesis associated to two independent clusterings we can assume a generalized hypergeometric distribution. The Adjusted Rand Index is defined as the difference between the measured Rand Index and its mean value under the null hypothesis, normalized by the maximum that this difference can reach:

$$\mathcal{R}_{adj}(Y, Y') \equiv \frac{\sum_{i=1}^{k}\sum_{j=1}^{l}\binom{m_{ij}}{2} - t_3}{\frac{1}{2}(t_1 + t_2) - t_3}, \tag{9}$$

where

$$t_1 = \sum_{i}^{k}\binom{|Y_i|}{2} \quad , \quad t_2 = \sum_{j}^{l}\binom{|{Y'}_j|}{2} \quad , \quad t_3 = \frac{2t_1 t_2}{N(N-1)}. \tag{10}$$

It turns out that $\mathcal{R}_{adj} \in [-1, 1]$, with 1 correspondent to the case of identical clusterings and 0 to two completely scorrelated clusterings. Negative values instead show anti-correlation between Y and Y' (that is, the number of pairs classified in the same way by Y and Y' is less even than what expected by a random overlapping between the two clusterings).

Hypergeometric test for cluster-industry overexpression

Let us call Y_i a cluster in our clustering and Y'_j a ICB industry. We want to verify whether Y_i overexpresses Y'_j. Say k is the number of stocks in common between Y'_j and Y_i, and $|Y_i|$, $|Y'_j|$ are the cardinalities of the cluster and the industry respectively; then the Hypergeometric distribution reads [40]:

$$P(X = k) \;=\; \frac{\binom{|Y'_j|}{k}\binom{N-|Y'_j|}{|Y_i|-k}}{\binom{N}{|Y_i|}}. \tag{11}$$

This is the null hypothesis for the test. If $P(X = k)$ so calculated is less than the significance level, then the test is rejected. The significance level of the tests performed is 1%. Since for each sector we perform N_{cl} hypothesis tests (one for each cluster), we use the conservative Bonferroni correction [40] used for multiple hypothesis tests, that reduces the significance level to $1/N_{cl}\%$.

If the test is not rejected, then we cannot reject the hypothesis that, in Y_i, the k stocks from the ICB industry Y'_j are picked up just by chance, without any preference for that industry. If instead the test is rejected, we conclude that the cluster Y_i overexpresses the industry Y'_j.

Bootstrapping test of robustness

The basic idea of the Bootstrapping technique is the following [46]: suppose, for a given time window of length L, we have N time series (one for each stock), each one having length L . We can fit this data in a $N \times L$ matrix, say X, and calculate the correlation matrix for it, say ρ , and a clustering using the DBHT, say Y. Now let us create a replica X' of the matrix X, such that each row of X' is drawn randomly among the rows of X, allowing multiple drawings of the same rows. From X' we can again calculate a correlation matrix ρ' and a clustering Y'. By repeating this procedure n_{boot} times, we end up with n_{boot} replica of clusterings, each one slightly different from the original one due to the differences between X and its replicas.

Robustness of DBHT: a bootstrapping analysis

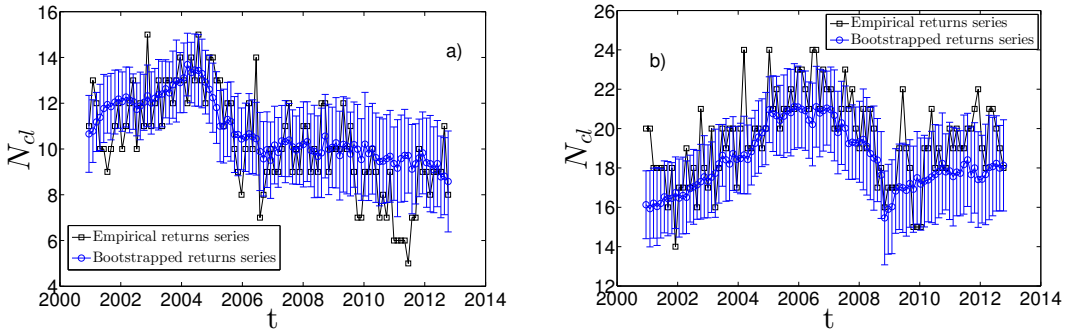

Figure 12. Test of robustness for the dynamical DBHT clustering. a) Number of clusters N_{cl} as a function of the time t: the black squares correspond to the DBHT clusterings obtained by using the empirical (non-detrended) log-returns, the blue dots are the average over the 100 N_{cl} given by the 100 replica correlation matrices (see text for further details). The bar errors in the blue dot plot is the standard deviation calculated among the same set of 100 N_{cl}. As one can see the empirical N_{cl} is quite robust against the bootstrapping test. b) Same plot as in a), but by using detrended log-returns.

In order to test the sensitiveness of the DBHT clustering to the statistical noise, inevitably present in every correlation estimate, we performed the Bootstrapping test to our dataset.

In Fig. 12 we show the result of a dynamical Bootstrapping, performed over all the 100 time windows that cover the entire period. We chose $n_{boot} = 100$. The blue points are the average number of clusters over the n_{boot} replicas Y', whereas the error bars are the standard deviations calculated over the same sample. The black squares are the empirical numbers of clusters yielded by the DBHT. The left-hand side plot (a)) is by using non-detrended log-returns, the right-hand side (b)) detrended log-returns.

The plot of empirical number of clusters is slightly different from what we have shown in Fig. 8 a) because for this bootstrapping analysis we did not use exponential smoothing for the correlations, but only bare correlations. The exponential smoothing, indeed, creates an asymmetry among the points in each time series that makes the bootstrapping test inapplicable.

From the plot we can observe that the method is statistically robust, with the most of empirical points within one standard deviation from the mean of replicas. More importantly, the mean of replicas follows the general trend of the empirical points; namely, the decreasing trend in the market mode case, and the drop after the 2007-2008 credit crunch in the detrended case.

Acknowledgments

The authors wish to thank Bloomberg for providing the data. TDM wishes to thank the COST Action TD1210 for partially supporting this work. TA acknowledges support of the UK Economic and Social Research Council (ESRC) in funding the Systemic Risk Centre (ES/K002309/1).

References

1. Mantegna RN (1999) Hierarchical structure in financial markets. Eur Phys J B 11: 193.
2. Onnela JP, Chakraborti A, Kaski K, Kertész J, Kanto A (2003) Asset trees and asset graphs in financial markets. Phys Scr T106: 48.
3. Aste T, Di Matteo T, Hyde ST (2005) Complex networks on hyperbolic surfaces. Physica A 346: 20.
4. Tumminello M, Aste T, Di Matteo T, Mantegna RN (2005) A tool for filtering information in complex systems. Proc Natl Acad Sci USA 102.
5. Di Matteo T, Aste T (2002) How does the eurodollars interest rate behave? J Theoret Appl Finance 5: 122-127.
6. Di Matteo T, Aste T, Mantegna RN (2004) An interest rate cluster analysis. Physica A 339: 181-188.
7. Di Matteo T, Aste T, Hyde ST, Ramsden S (2005) Interest rates hierarchical structure. Physica A 335: 21-33.
8. Bartolozzi M, Mellen C, Di Matteo T, Aste T (2007) Multi-scale correlations in different futures markets. Eur Phys J B 58: 207-220.
9. Onnela JP, Chakraborti A, Kaski K, Kertész J (2003) Dynamic asset trees and black monday. Physica A 324: 247-252.
10. Tola V, Lillo F, Gallegati M, Mantegna R (2008) Cluster analysis for portfolio optimization. J Econ Dyn Control 32: 235-258.
11. Fenn DJ, Porter MA, Mucha P, McDonald M, Williams S, et al. (2012) Dynamical clustering of exchange rates. Quantitative Finance 12: 1493.
12. Di Matteo T, Pozzi F, Aste T (2010) The use of dynamical networks to detect the hierarchical organization of financial market sectors. Eur Phys J B 73: 3-11.
13. Tumminello M, Di Matteo T, Aste T, Mantegna RN (2007) Correlation based networks of equity returns sampled at different time horizons. Eur Phys J B 55: 209-217.
14. Morales R, Di Matteo T, Aste T (2014) Dependency structure and scaling properties of financial time series are related. Sci Rep 4: 4589.

15. Aste T, Shaw W, Di Matteo T (2010) Correlation structure and dynamics in volatile markets. New J Phys 12: 085009.

16. Tumminello M, Lillo F, Mantegna RN (2010) Correlation, hierarchies, and networks in financial markets. J Econ Behav Organ 75: 40-58.

17. Anderberg MR Cluster Analysis for Applications. Academic Press.

18. Tumminello M, Coronnello C, Lillo F, Micciché S, Mantegna RN (2007) Spanning trees and bootstrap reliability estimation in correlation-based networks. Int J Bifurcat Chaos 17: 2319-2329.

19. Aste T. An algorithm to compute maximally filtered planar graphs. URL `http://www.mathworks.com/matlabcentral/fileexchange/38689-pmfg`.

20. Pozzi F, Di Matteo T, Aste T (2013) Spread of risk across financial markets: better to invest in the peripheries. Sci Rep 3: 1665.

21. Song WM, Di Matteo T, Aste T (2012) Hierarchical information clustering by means of topologically embedded graphs. PLoS ONE 7: e31929.

22. Aste T. An algorithm to compute dbht clustering. URL `http://www.mathworks.com/matlabcentral/fileexchange/46750-dbht`.

23. Song WM, Di Matteo T, Aste T (2011) Nested hierarchies in planar graphs. Discrete Appl Math 159: 2135.

24. Kaufman L, Rousseeuw PJ (1987) Clustering by means of medoids. Statistical Data Analysis Based on the L_1-Norm and Related Methods : 405-416.

25. MacQueen J (1967) Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability 1: 281-297.

26. Hubert L, Arabie P (1985) Comparing partitions. Journal of Classification 2: 193-218.

27. Tumminello M, Micciché S, Lillo F, Varho J, Piilo J, et al. (2011) Community characterization of heterogeneous complex systems. J Stat Mech P01019.

28. Coronnello C, Tumminello M, Lillo F, Micciché S, Mantegna RN (2011) Sector identification in a set of stock return time series traded at the london stock exchange. Acta Physica Polonica 36: 2653-2680.

29. Coronnello C, Tumminello M, Lillo F, Micciché S, Mantegna RN (2011) Sector identification in a set of stock return time series traded at the london stock exchange. Acta Physica Polonica 36: 2653-2680.

30. Tumminello M, Lillo F, Mantegna R (2007) Kullback-leibler distance as a measure of the information filtered from multivariate data. Phys Rev E Stat Nonlin Soft Matter Phys 76: 031123.

31. Cont R (2001) Empirical properties of asset returns: stylized facts and statistical issues. Quantitative Finance 1: 223-236.

32. Mantegna R, Stanley H (2000) An introduction to econophysics: correlations and complexity in finance. Cambridge University Press.

33. Bonanno G, Caldarelli G, Lillo F, Mantegna R (2003) Topology of correlation-based minimal spanning trees in real and model markets. Phys Rev E Stat Nonlin Soft Matter Phys 68: 046130.

34. Pearson K (1895) Notes on regression and inheritance in the case of two parents. Proceedings of the Royal Society of London 58: 240-242.

35. Pozzi F, Di Matteo T, Aste T (2012) Exponential smoothing weighted correlations. Eur Phys J B 85: 6.

36. Borghesi C, Marsili M, Miccichè S (2007) Emergence of time-horizon invariant correlation structure in financial returns by subtraction of the market mode. Phys Rev E Stat Nonlin Soft Matter Phys 76: 026104.

37. Ross G (2014) Dynamic multifactor clustering of financial networks. Phys Rev E Stat Nonlin Soft Matter Phys 89: 022809.

38. A guide to industry classification benchmark. from http://wwwicbenchmarkcom/ .

39. Wagner S, Wagner D (2007) Comparing clusterings - an overview. Technical Report, ITI Wagner, Faculty of Informatics, Universitt Karlsruhe (TH) .

40. Tumminello M, Miccichè S, Lillo F, Piilo J, Mantegna R (2011) Statistically validated networks in bipartite complex systems. PLoS ONE 6: e17994.doi:10.1371/journal.pone.0017994.

41. Rosvall M, Bergstrom C (2010) Mapping change in large networks. PLoS ONE 5: e8694. doi:10.1371/journal.pone.0008694.

42. Kendall MG (1938) A new measure of rank correlation. Biometrika 30: 81-93.

43. Shannon CE (1948) A mathematical theory of communication. Bell Syst Tech J 27: 379.

44. Nanda S, Mahanty B, Tiwari MK (2010) Clustering indian stock market data for porfolio management. Expert System with Applications 37: 8793-8798.

45. Pai GAV, Michel T (2009) Clustering indian stock market data for porfolio management. Evolutionary Optimization of Constrained K-means Clustered Assets for Diversification in Small Portfolios 13: 1030-1053.

46. Efron B (1979) Bootstrap methods: Another look at the jackknife. Ann Stat 7: 1-26.

Supplementary Materials

Dataset analysis

The set of stocks has been chosen in order to provide a significant sample of the different industrial sectors in the market. We have chosen the ICB industrial classification, that yields 19 different Supersectors, that in turns gather in 10 Industries: the percentage of stocks belonging to each ICB supersectors is reported in Fig. 13 .

In Fig. 14 two plots are shown that summarize the main features of this dataset. The graphs show the average price $\bar{P}(t) \equiv \frac{1}{N}\sum_i P_i(t)$ and the average log return of the prices, $\bar{r}(t) \equiv \frac{1}{N}\sum_i r_i(t)$, as a function of time. From these plots we can see that both the internet bubble bursting (2002) and the credit crunch (2007-08) are displayed by the market dynamics. In particular it is evident a steep increase in volatility for both periods, strongly autocorrelated in time: a well known feature of log-returns dynamics [31]. Such clusters of volatility can be observed also after the credit crunch, in 2010 and 2012.

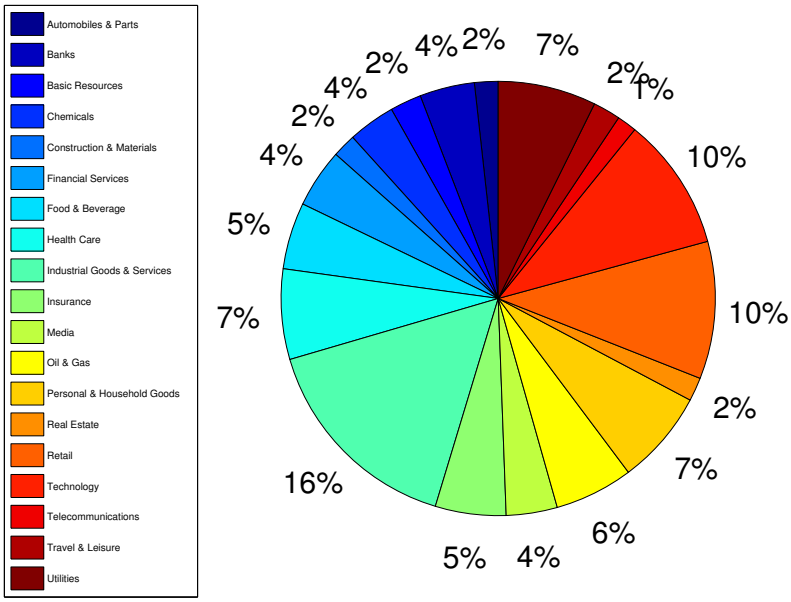

Figure 13. Pie chart showing the composition of the entire set of stocks in terms of ICB supersectors.

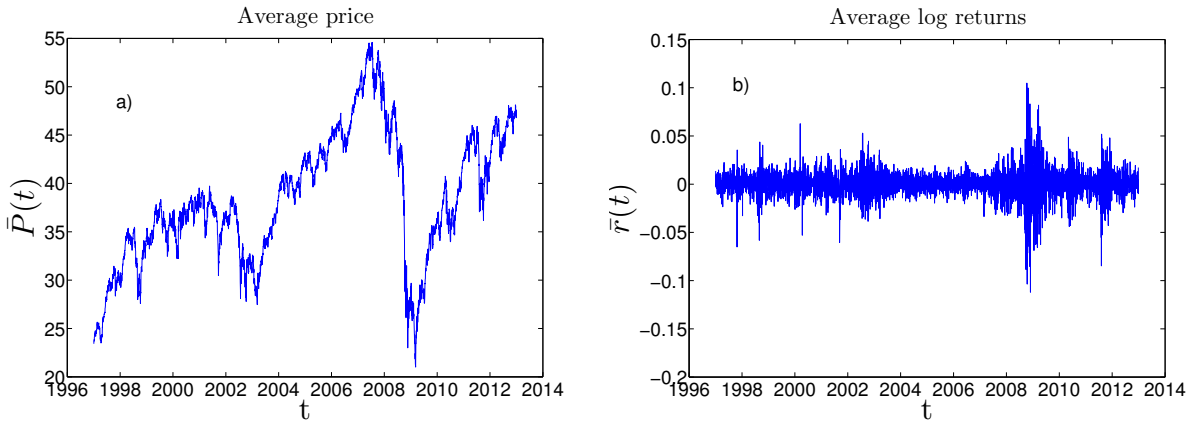

Figure 14. Average price and log-returns of the dataset, from January 1997 to December 2012. a) Average price $\bar{P}(t)$ of the 342 US stocks in the dataset; b) Average log-return $\bar{r}(t)$ of the same prices.

Clustering compositions: detrended case

DBHT clusters

In Fig. 15 we report a graphical summary of the clusters obtained applying the DBHT method to the whole time window of data (1997-2012), by using detrended log-returns.

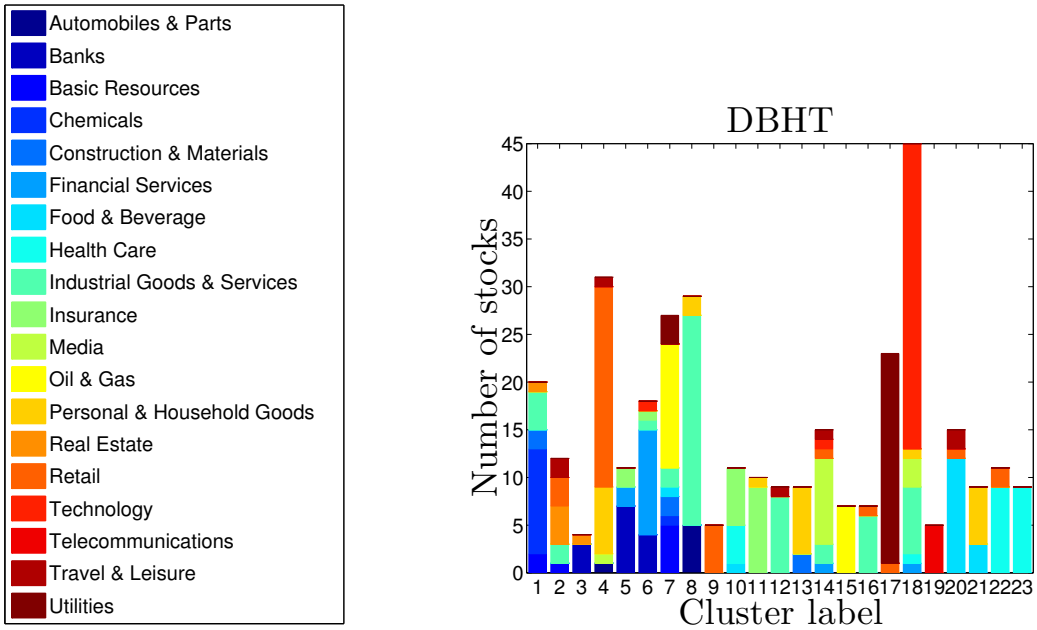

Figure 15. Composition of DBHT clusters in terms of ICB supersectors. The composition of each cluster in terms of ICB supersectors is shown by using different colours (legend on the left hand side). The clustering is obtained by using log-returns detrended by removing the market mode.

The number of clusters is 23: by detrending the market mode the cluster structure becomes richer than the non-detrended case (17 clusters). The largest cluster contains 45 stocks (13% of total), the smallest 4. The average size is now reduced to 14.8.

Comparing Fig. 3 in the paper with Fig. 15 we have observed that some supersectors that were mixed together in the non-detrended case are now overexpressed in distinct clusters: Chemicals (cluster 1), Insurance (cluster 11) and Telecommunications (cluster 19). Moreover, some supersectors that were overexpressed in the non-detrended case tend to be more spread over different clusters, still being overexpressed in these clusters: Utilities, Oil & Gas, the Financial industry.

Overall, and not surprisingly, we can conclude that by subtracting the market mode we get a richer, more structured clustering that shows a higher overexpression of ICB supersectors. However some of them, already overexpressed in the clustering that includes market mode, tend to be split among different clusters. This could point out that the detrended dynamics, although shows a clear industrial sector structure, is also driven by subsectorial correlations.

Linkage clusters

The effects of the market mode subtraction from the log-returns depend on the clustering method used. In Figs. 16 a), b) and c) we show the new clusterings obtained by using detrended log-returns and a number of clusters equal to 23 (so that the clusterings are comparable to that of DBHT in Fig. 15).

As we can see, for the SL the clustering does not change sensitively, being still present a giant cluster (318 stocks) with many small clusters of 1 or 2 stocks. The AL case shows instead a huge change: the size of the largest cluster shrinks to 58 stocks, and 6 different clusters of medium size (20-40 stocks) appear.

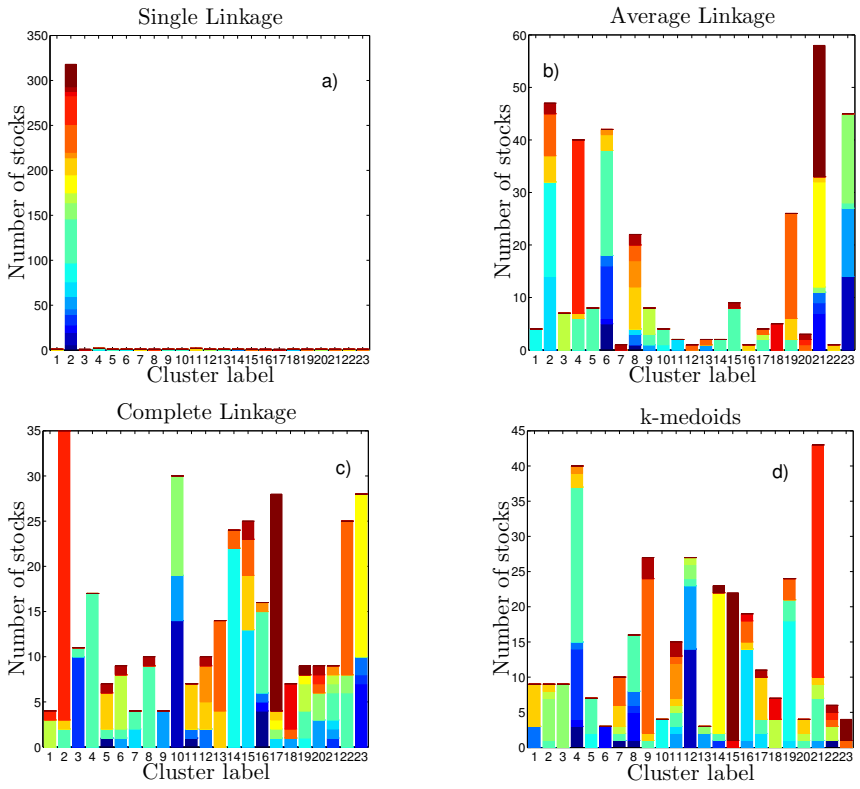

Figure 16. Composition of clustering in terms of ICB supersectors. The x-axis represents the single cluster labels, the y-axis the number of stocks in each cluster. Each colour corresponds to an ICB supersector (the legend is the same as in Fig. 15. The graphs show the results for a) SL clustering, b) for AL, c) for CL and d) for k-medoids. The clustering is obtained by using log-returns detrended by removing the market mode.

Moreover, these new clusters show a much higher overexpression of many industrial supersectors, such as Technology (cluster 4), Industrial Goods & Services (cluster 5 and 15), Media (cluster 3) and Financial related supersectors (cluster 23). However there are still 10 clusters whose size is at most 4 stocks.

A remarkable degree of supersectors overexpression can be found also in the CL case. Some of them, that were not overexpressed in the non-detrended case, appear now, such as Financial supersectors (cluster 10) and Industrial Goods & Services (clusters 4 and 8). The CL clustering shows moreover a higher homogeneity in the clusters size, with the largest cluster made of 35 stocks and only three clusters with less than 5 stocks.

Therefore we can conclude that the DBHT is better at highlighting the information contained in the ICB partition, at least for this particular choice of N_{cl}, showing at the same time a more homogeneous distribution of the cluster size. Among the Linkage methods, the CL performs better than AL and SL. The subtraction of the market mode makes all the clusterings methods (with the exception of SL) more homogenous in size and more able to retrieve the ICB partition. The SL clustering instead does not seem to be sensitive to this subtraction, and keeps not overexpressing any ICB supersector.

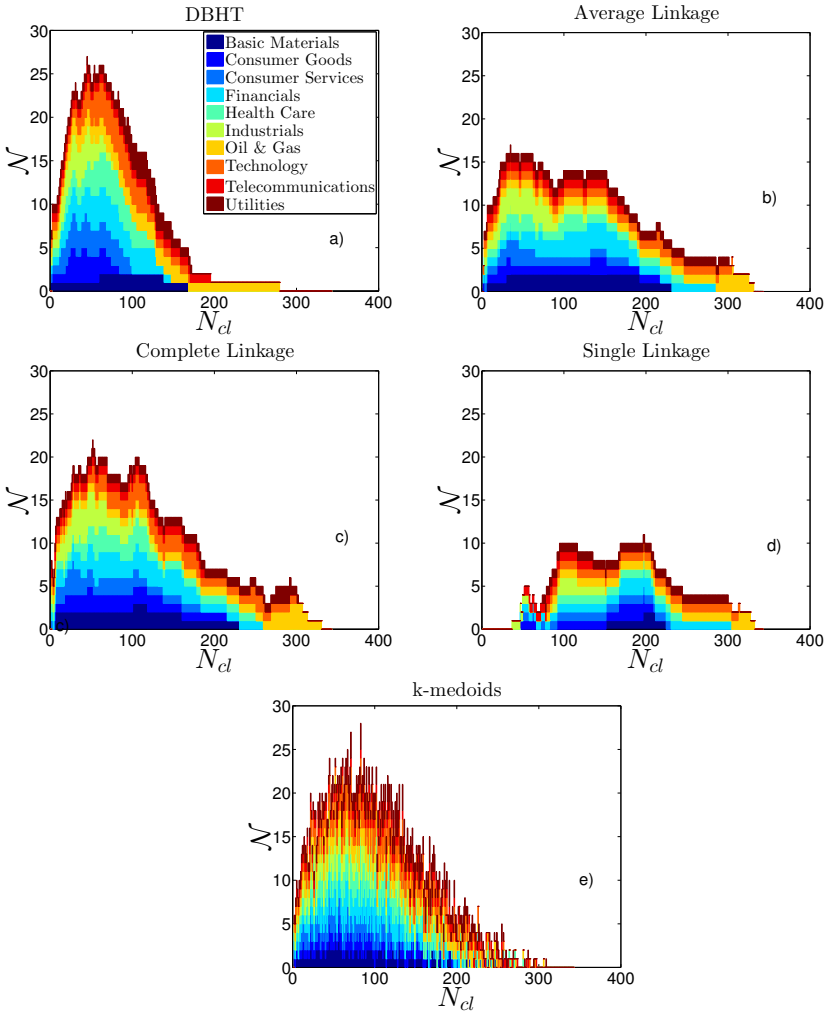

Figure 17. ICB industries overexpression at different levels of the hierarchies. Each bar graph shows, varying the number of clusters N_{cl}, how many times ($\mathcal{N}$) an ICB industry is overexpressed by a cluster, according to the Hypergeometric hypothesis test (i.e., number of tests being rejected). Each colour shows the number of overexpressions for each ICB industry. In graphs a)-e) the results for DBHT, AL, CL, SL and k-medoids clustering respectively are shown. The correlations are calculated on log-returns detrended by the market mode.

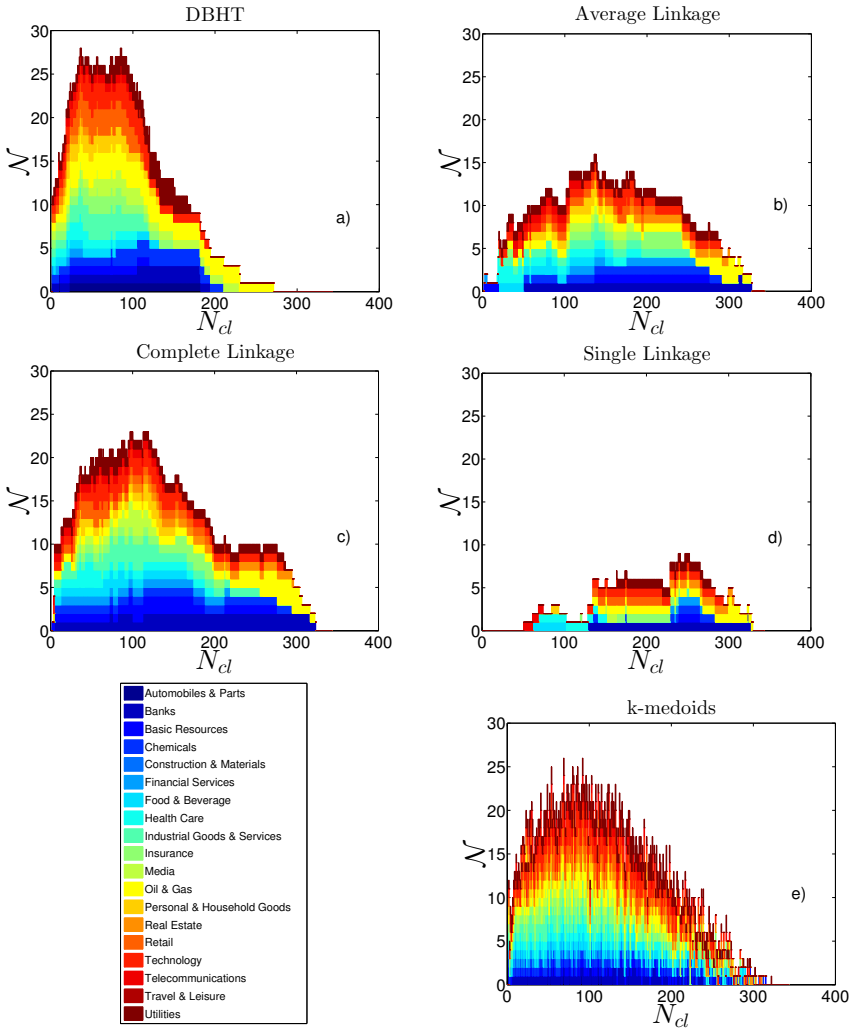

Figure 18. ICB supersectors overexpression at different levels of the hierarchies. Each bar graph shows, varying the number of clusters N_{cl}, how many times ($\mathcal{N}$) an ICB supersector is overexpressed by a cluster, according to the Hypergeometric hypothesis test (i.e., number of tests being rejected). Each colour shows the number of overexpressions for each ICB supersector. In graphs a)-e) the results for DBHT, AL, CL, SL and k-medoids clustering respectively are shown . The correlations are calculated on non-detrended log-returns.

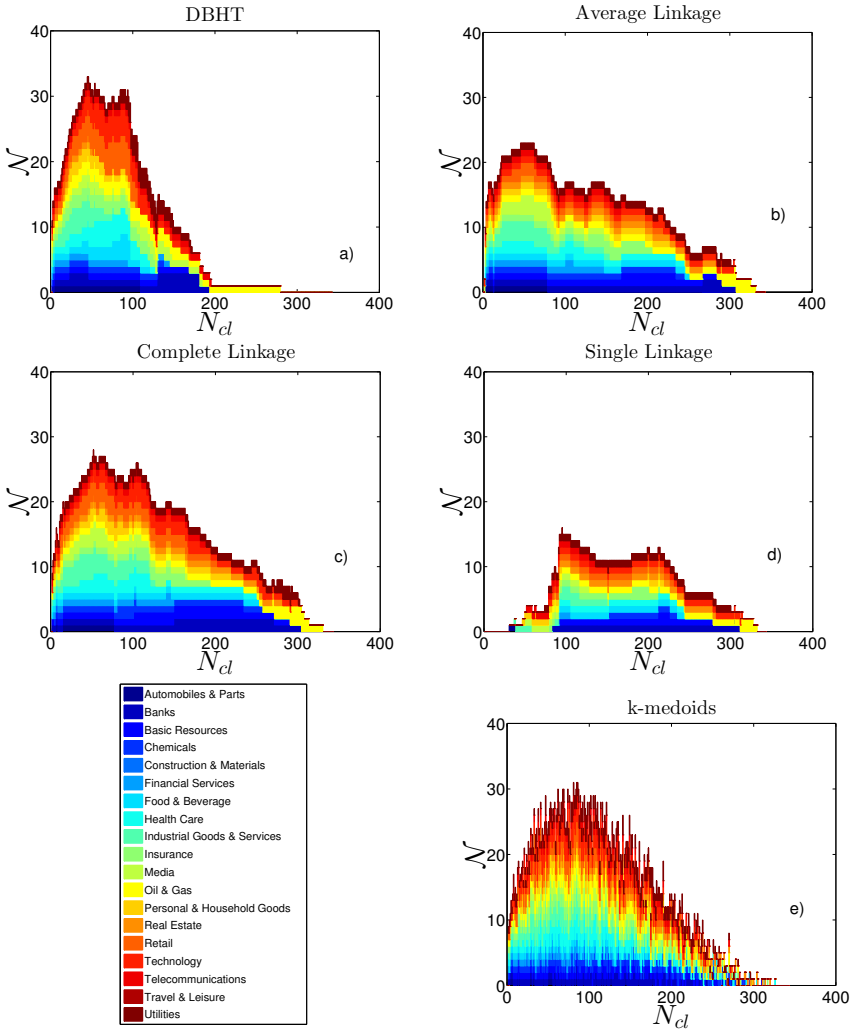

Figure 19. ICB supersectors overexpression at different levels of the hierarchies. Each bar graph shows, varying the number of clusters N_{cl}, how many times ($\mathcal{N}$) an ICB supersector is overexpressed by a cluster, according to the Hypergeometric hypothesis test (i.e., number of tests being rejected). Each colour shows the number of overexpressions for each ICB supersector. In graphs a)-e) the results for DBHT, AL, CL, SL and k-medoids clustering respectively are shown. The correlations are calculated on detrended log-returns.