Available online at www.sciencedirect.com
ScienceDirect
Physica A 379 (2007) 635–644
PHYSICA A
www.elsevier.com/locate/physa

Hausdorff clustering of financial time series

Nicolas Basalto[a], Roberto Bellotti[b,c,d], Francesco De Carlo[b,c], Paolo Facchi[c,e], Ester Pantaleo[b,c], Saverio Pascazio[b,c,*]

[a]*Market Risk Management, Unicredito Italiano, Milano, Italy*
[b]*Dipartimento di Fisica, Università di Bari, Italy*
[c]*Istituto Nazionale di Fisica Nucleare, Sezione di Bari, Italy*
[d]*TIRES, Center of Innovative Technologies for Signal Detection and Processing, Bari, Italy*
[e]*Dipartimento di Matematica, Università di Bari, Italy*

Received 20 March 2006; received in revised form 20 December 2006
Available online 17 February 2007

Abstract

A clustering procedure is introduced based on the Hausdorff distance as a similarity measure between clusters of elements. The method is applied to the financial time series of the Dow Jones industrial average (DJIA) index to find companies that share a similar behavior. Comparisons are made with other linkage algorithms.
© 2007 Elsevier B.V. All rights reserved.

Keywords: Econophysics; Clustering; Hausdorff metric

1. Introduction

Clustering consists in grouping a set of objects in classes according to their degree of "similarity" [1]. This intuitive concept can be defined in a number of different ways, leading in general to different partitions. For this reason, it is clear that a clustering procedure can be profoundly influenced by the strategy adopted by the observer and his/her own ideas and preconceptions about the data set. In this article we will focus on a *linkage* algorithm that consists in merging, at each step, the two clusters with the smallest dissimilarity, starting from clusters made up of a single element and ending up in a single cluster collecting all data. Our objective will be to cluster the financial time series of the stocks belonging to the Dow Jones industrial average (DJIA) index.

From a mathematical point of view, given a set of objects $\mathscr{S} \equiv \{s\}$, an allocation function $m : \mathscr{S} \to \{1, 2, \ldots, k\}$, is defined so that $m(s)$ is the class label and k the total number of clusters (which we assume to be finite for simplicity). The aim of a clustering procedure is to select, among all possible allocation functions, the one performing the best partition of the set $\mathscr{S}$ into subsets $\mathscr{G}_\alpha \equiv \{s \in \mathscr{S} | m(s) = \alpha\}, (\alpha = 1, \ldots, k)$, relying on some measure of similarity. The operational meaning of similarity will be specified in the following and will be based on the Hausdorff distance, to be introduced in the next section.

*Corresponding author. Istituto Nazionale di Fisica Nucleare, Sezione di Bari, Italy. Fax: +39 80 544 2470.
E-mail address: saverio.pascazio@ba.infn.it (S. Pascazio).

0378-4371/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.physa.2007.01.011

Clustering algorithms can be classified in different ways according to the criteria used to implement them [2]. The so-called "hierarchical" methods yield nested partitions, represented by *dendrograms* [3], in which any cluster can be further divided in order to observe its underlying structure. Linkage algorithms, in particular, are hierarchical. Other non-hierarchical (or "partitional") methods are also possible [4–6], but will not be discussed here.

2. Hausdorff clustering

In order to cluster a given data set we will use a distance function introduced by Hausdorff. Given a metric space $(\mathscr{S}, \delta)$, with metric δ, the distance between a point $a \in \mathscr{S}$ and a subset $B \subseteq \mathscr{S}$ is naturally given by

$$\tilde{d}(a; B) = \inf_{b \in B} \delta(a, b) \tag{1}$$

(all subsets are henceforth considered to be non-empty and compact). Given a subset $A \subseteq \mathscr{S}$, let us define the function

$$\tilde{d}(A; B) = \sup_{a \in A} \tilde{d}(a; B) = \sup_{a \in A} \inf_{b \in B} \delta(a, b) \tag{2}$$

that measures the largest among all distances $\tilde{d}(a; B)$, with $a \in A$. This function is not symmetric, $\tilde{d}(A; B) \neq \tilde{d}(B; A)$, and therefore is not a *bona fide* distance. The Hausdorff distance [7] between two sets $A, B \subseteq \mathscr{S}$ is defined as the largest between the two numbers:

$$d_{\mathrm{H}}(A, B) = \max\{\tilde{d}(A; B), \tilde{d}(B; A)\}, \tag{3}$$

namely,

$$d_{\mathrm{H}}(A, B) = \max\left\{ \sup_{a \in A} \inf_{b \in B} \delta(a, b), \sup_{b \in B} \inf_{a \in A} \delta(a, b) \right\} \tag{4}$$

that is clearly symmetric.

In words, the Hausdorff distance between A and B is the smallest positive number r, such that every point of A is within distance r of some point of B, *and* every point of B is within distance r of some point of A. The geometrical meaning of the Hausdorff distance is best understood by looking at an example, such as that in Fig. 1. We emphasize that the Hausdorff metric relies on the metric δ on $\mathscr{S}$.

If the data set is finite and consists of N elements, all distances can be arranged in a $N \times N$ matrix δ_{ij} and Eq. (4) reads

$$d_{\mathrm{H}}(A, B) = \max\left\{ \max_{i \in A} \min_{j \in B} \delta_{ij}, \max_{j \in B} \min_{i \in A} \delta_{ij} \right\}, \tag{5}$$

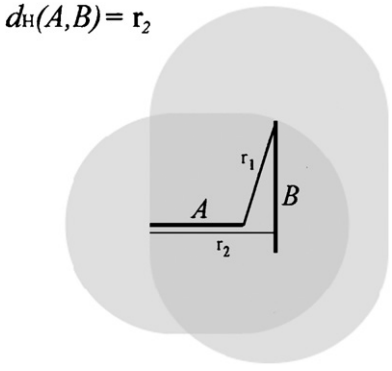


Fig. 1. Hausdorff distance between two sets A and B (black thick segments). $r_1 = \tilde{d}(B; A)$, $r_2 = \tilde{d}(A; B)$. The Hausdorff distance is equal to the larger radius r_2.

which is a very handy expression, as it amounts to finding the minimum distance in each row (column) of the distance matrix, then the maximum among the minima. The two numbers are finally compared and the largest one is the Hausdorff distance. This sorting algorithm is easily implemented in a computer.

We shall take the Hausdorff distance as the (inverse) similarity measure. This distance naturally translates in a linkage algorithm: at the first level each element is a cluster and the Hausdorff distance between any pair of points reads

$$d_{\mathrm{H}}(\{i\},\{j\}) = \delta_{ij} \tag{6}$$

and coincides with the underlying metric. The two elements of $\mathscr{S}$ at the shortest distance are then joined together in a single cluster. The Hausdorff distance matrix is recomputed, considering the two joined elements as a single set. This iterative process goes on until all points belong to a single final cluster.

3. Comparison with single and complete linkage

It is interesting to notice that the partitions obtained by the Haudorff linkage algorithm are intermediate between those obtained by the more commonly used "single" and "complete" linkage procedures: if A and B are two non empty compact subsets of $\mathscr{S}$, the single and complete linkage algorithms make use of the following similarity indexes

$$d_{\mathrm{S}}(A,B) = \inf_{a\in A, b\in B} \delta(a,b), \tag{7}$$

$$d_{\mathrm{C}}(A,B) = \sup_{a\in A, b\in B} \delta(a,b), \tag{8}$$

respectively, that are to be compared with (4). In terms of the distance matrix δ_{ij} of a finite data set, they are given by

$$d_{\mathrm{S}}(A,B) = \min_{i\in A, j\in B} \delta_{ij}, \quad d_{\mathrm{C}}(A,B) = \max_{i\in A, j\in B} \delta_{ij}, \tag{9}$$

instead of (5).

In order to compare these different algorithms, it is useful to recall the mathematical definition of distance. Given a set $\mathscr{X}$, a distance (or a metric) d is a non-negative application

$$d : \mathscr{X} \times \mathscr{X} \longrightarrow \mathbb{R}, \tag{10}$$

endowed with the following properties, valid $\forall x, y \in \mathscr{X}$:

$$d(x,y) = 0 \iff x = y, \tag{11}$$

$$d(x,y) = d(y,x), \tag{12}$$

$$d(x,y) \leqslant d(x,z) + d(y,z) \quad \forall z \in \mathscr{X}. \tag{13}$$

Incidentally, notice that symmetry (12), as well as non-negativity, are not independent assumptions, but easily follow from (11) and the triangular inequality (13).

It is not difficult to prove from the very definition (4) that the Hausdorff distance between compact and non-empty sets satisfies (11)–(13). On the other hand, (7) and (8) are *not* distances: the former does not satisfy the triangular inequality (13), while the latter does not fulfill the basic requirement (11), $d_{\mathrm{C}}(A,A) \neq 0$, for any compact set containing more than one point: in this sense, it performs a sort of coarse graining over the data set. The Hausdorff function, being a distance in a strict mathematical sense, enables us to rest on sound mathematical ground.

The Hausdorff distance is not commonly used in the context of clustering (see Ref. [2, Chapter 4, for a discussion]). It is a useful tool in the analysis of complex sets, with complicated (and even fractal-like) structures. It is in such a case that one expects that Hausdorff behaves better than the other methods, since it relies on rigorous mathematical concepts.

4. Application to financial data

We now apply the Hausdorff linkage algorithm to a topic of growing interest: the analysis of financial time series. In particular, we focus on the $N = 30$ shares composing the DJIA index, collecting the daily closure prices of its stocks for a period of 5 years (1998–2002). We chose this index for two reasons. First, because these data are easily accessible. The second, and more important reason is the "quality" (in the sense of reliability) of prices. The DJIA index, indeed, aggregates the shares of some of the more valuable and capitalized world corporations, so that their prices are highly contributed by market makers. This means that we always expect to find, even in the worst possible scenario, a financial intermediator (market maker) ready to quote both bid and offer prices for these assets. For this reason, these shares are very frequently traded. In financial terminology, they are said to be "liquid".

Fig. 2 displays the typical behavior of a stock value (IBM) for the investigated time period. The companies of the DJIA index are reported in Appendix A, together with the corresponding industries. We will look at the temporal series of the daily logarithm closure price differences

$$Y_i(t) \equiv \ln P_i(t) - \ln P_i(t-1), \tag{14}$$

where $P_i(t)$ is the closure price of the ith share at day t. Both P_i and Y_i are very irregular functions of time. In order to quantify the degree of similarity between two time series and use our linkage algorithm we adopt the following metric function, that quantifies the synchronicity in their time evolution [8–10]

$$d_{ij} = \sqrt{2(1 - c_{ij})}, \tag{15}$$

where c_{ij} are the correlation coefficients computed over the investigated time period:

$$c_{ij} = \frac{\langle Y_i Y_j \rangle - \langle Y_i \rangle \langle Y_j \rangle}{\sqrt{(\langle Y_i^2 \rangle - \langle Y_i \rangle^2)(\langle Y_j^2 \rangle - \langle Y_j \rangle^2)}} \tag{16}$$

and the brackets denote the average over the time interval of interest (one year in our case). Table 1 displays a part of the $N \times N$ matrix of the correlation coefficients (year 1998). It is worth stressing that almost all correlation coefficients are positive, with values not too close to 1, thus confirming that, in many cases, stocks belonging to the same market do not move independently from each other, but rather share a similar temporal behavior. The distance (15) is a proper metric in the "parent" space, ranging from 0 for perfectly correlated series ($c_{ij} = +1$) to 2 for anticorrelated stocks ($c_{ij} = -1$). The representative points lie therefore on an hypersphere.

The Hausdorff clustering algorithm used in this article inherits the concept of similarity from the distance d in (15) (that is defined at the level of couples of elements—i.e., time series) and lifts it to the level of sets. The elements in one set are more similar among themselves (in the sense that their Hausdorff distance is smaller) and are *globally* less similar to those belonging to other sets. In this way, the intuitive notion of similarity is mathematically extended to sets in a rigorous fashion.

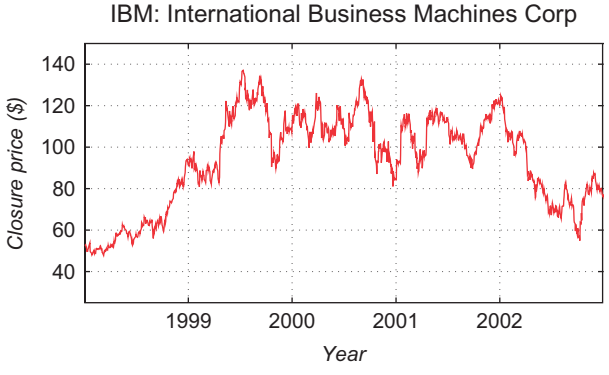


Fig. 2. Time evolution of the closure price of a stock value (IBM), for the period 1998–2002.

Table 1
A part of the matrix of the correlation coefficients c_{ij} (16) for the temporal series of the daily logarithm price differences of the stocks composing the DJIA index (year 1998)

	AA	AXP	BA	CAT	C
AA	1	0.37004	0.22458	0.3568	0.3508
AXP		1	0.35461	0.41916	0.61247
BA			1	0.32852	0.26917
CAT				1	0.33937
C					1

The acronyms (tickers) are explained in Appendix A.

5. Comparison of the methods and first results

We now compare the Hausdorff and single-linkage procedures, the latter being often proposed in the context of financial data analysis. A typical benchtest for clustering procedures is the construction of a tree.

The dendrograms and trees generated by the distances relative to year 1998 are shown in Fig. 3. The trees are obtained from the dendrograms according to the following procedure. At each step (corresponding to a given distance—abscissa on the dendrogram) the linkage algorithm merges together two clusters (or two points or a point and a cluster). A link is then added into the tree, between the two stocks in each cluster that are at the shortest distance. We proceed until the last stock is linked, namely we stop at the leftmost level of the dendrogram. We used this procedure both for the single and the Hausdorff linkages, obtaining the trees in Fig. 3. Notice that the tree obtained by using the single-linkage algorithm coincides with the minimum spanning tree [8].

The two trees show similarities but also interesting differences. The dendrogram relative to the single-linkage algorithm in Fig. 3(a) suffers from the well-known "chaining effect", that is clearly visible and consists in successively linking single elements to the closest element of a set, irrespectively of the other elements making up the set itself. The dendrogram obtained by Hausdorff in Fig. 3(b) displays a richer structure, with inner subclusters that are subsequently merged into larger clusters. The chaining effect in Fig. 3(a) yields a less structured, "star-like" tree in Fig. 3(c).

One notices the formation of identical subtrees centered around GE in both cases: {AXP, GM, JPM, C, UTX, BA, HON, AA, EK}, {KO, XOM, SBC}, {MSFT, INTC, IBM} and {HD, WMT}. Another common feature is the subtree of four elements: {CAT, DD, IP, MMM}. The presence of a common large subtree is found also in the other years investigated. The remaining companies are sparsely linked to the biggest subtree and at different points in the two trees. In the Hausdorff tree, they are linked to companies sharing the same industrial area, such as KO and PG (Consumer non-Cyclical), MRK and JNJ (Healthcare) and SBC and T (Services). Although the single-linkage procedure also perceives these relationships, the link is "mediated" by the Conglomerate GE and is therefore less direct. In addition, we notice the presence of the technological cluster {IBM, INTC, MSFT} and the financial cluster {AXP, JPM, C} in both trees.

This preliminary analysis, besides clarifying how one can obtain a tree from the Hausdorff algorithm, displays also the "topology" of the links between the elements of the data set. The separation of the stocks in clusters also corresponds to a large extent to the relevant economic sectors. This observation will be the object of the following section, in which the Hausdorff algorithm is further analyzed.

6. Results and discussion

Figs. 4 and 5 show the results of our analysis based on the Hausdorff distance. Rather than showing the dendrograms, we prefer to give a pictorial representation of the evolution of the stocks: in Figs. 4 and 3(d) we show the Hausdorff trees, while in Fig. 5 we use bubbles to represent clusters and arrows to represent the movements of the stocks. Some innermost subclusters are indicated with a dashed bubble and full (dashed) arrows denote future (past) movements. A small "exploding" star represents a bubble/cluster that disappears.

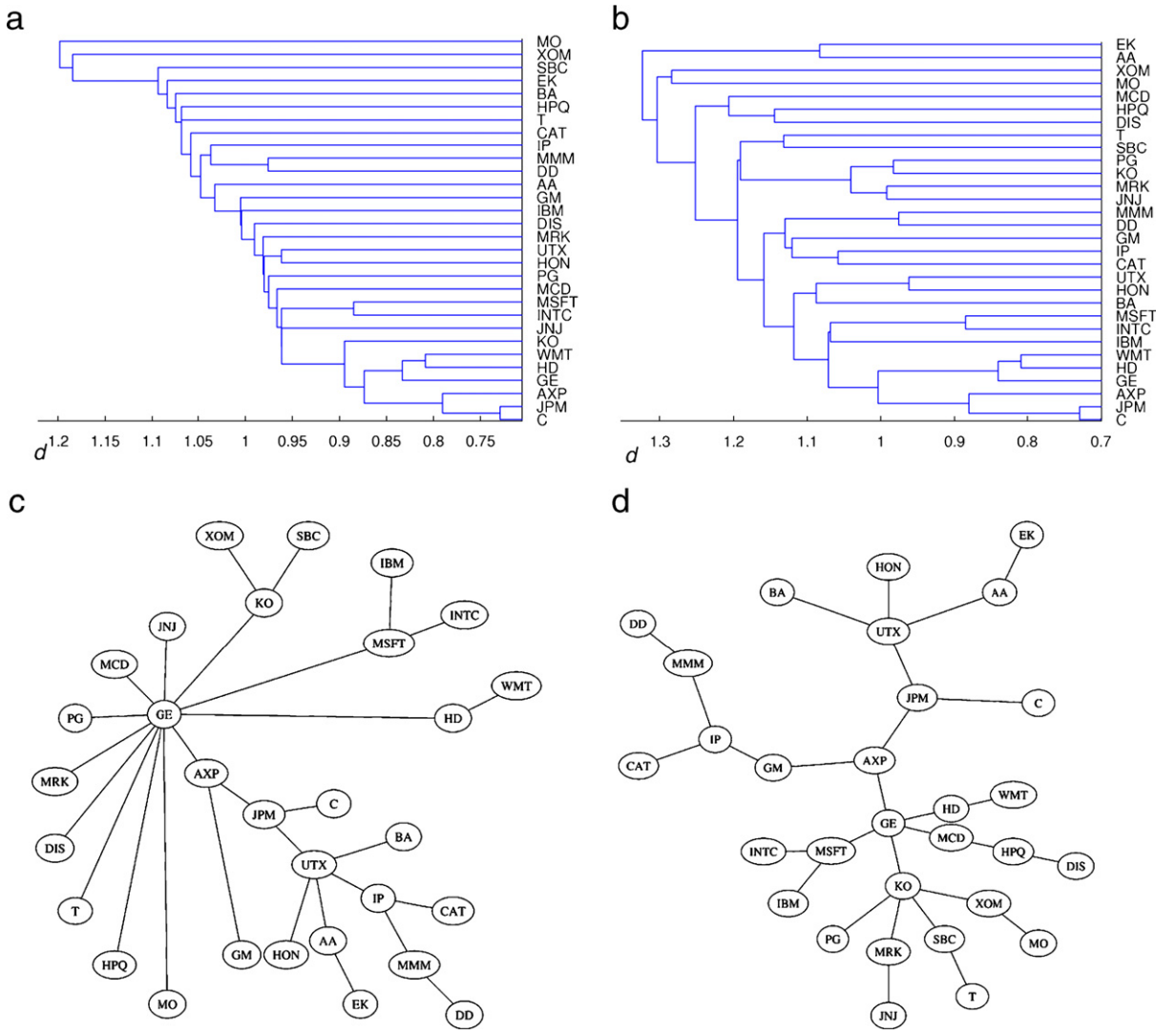


Fig. 3. Comparison between dendrograms and trees. Dendrograms emerging from the single linkage (a) and the Hausdorff algorithm (b). In (a), the single-linkage procedure displays the well-known chaining effect. (c) Minimum spanning tree based on the single linkage and (d) tree emerging from the Hausdorff linkage. The chaining effect in (a) is reflected in a less structured, ''star-like'' tree in (c). The acronyms (tickers) are explained in Appendix A. All figures refer to year 1998.

The clusters shown in Fig. 5 have been obtained by ''cutting'' the corresponding dendrograms at $n = 9$ clusters.

It is very interesting and challenging to try and analyze, from a mere economic viewpoint, some of the movements in the graphs, in order to catch some *a posteriori* hints about the dynamics of the stocks. Both in Figs. 4 and 5 one clearly recognizes that some of the clusters correspond to homogeneous groups of companies belonging to the same industry: this is the case of the financial services firms {AXP, JPM C}, retail companies {HD, WMT}, companies dealing with basic materials {AA, IP, DD}, the technological core {IBM, INTC, MSFT, HPQ} and the health care firms {JNJ, MRK}. However, one obtains a somewhat more detailed structure in Fig. 5, that is obtained directly from the relative dendrogram, rather than in Fig. 4, where the elements are merged into a tree, neglecting some more detailed information contained in the underlying Euclidean distance. In addition, the representation of Fig. 5 is more suitable to display the temporal evolution of the cluster structure.

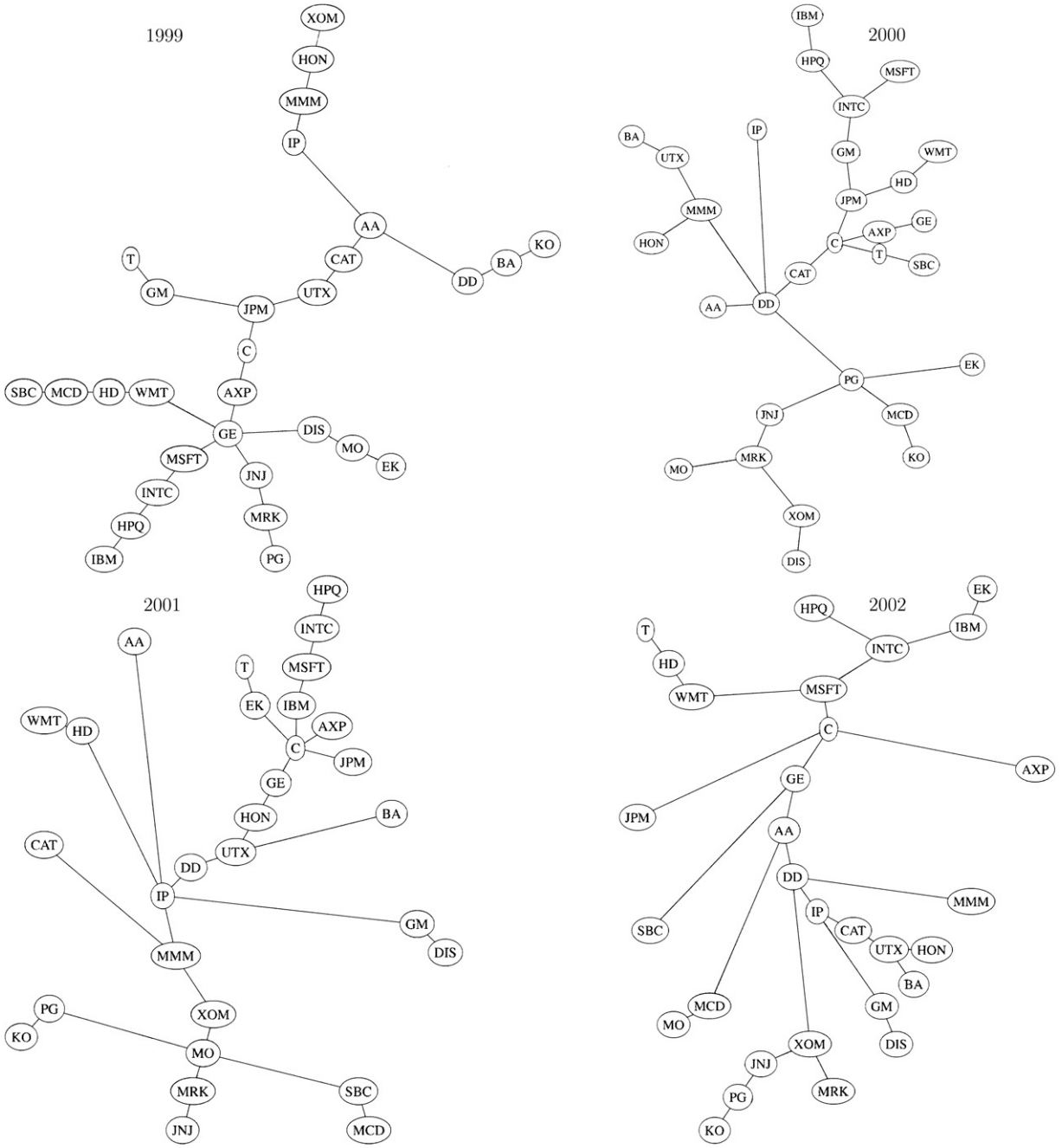
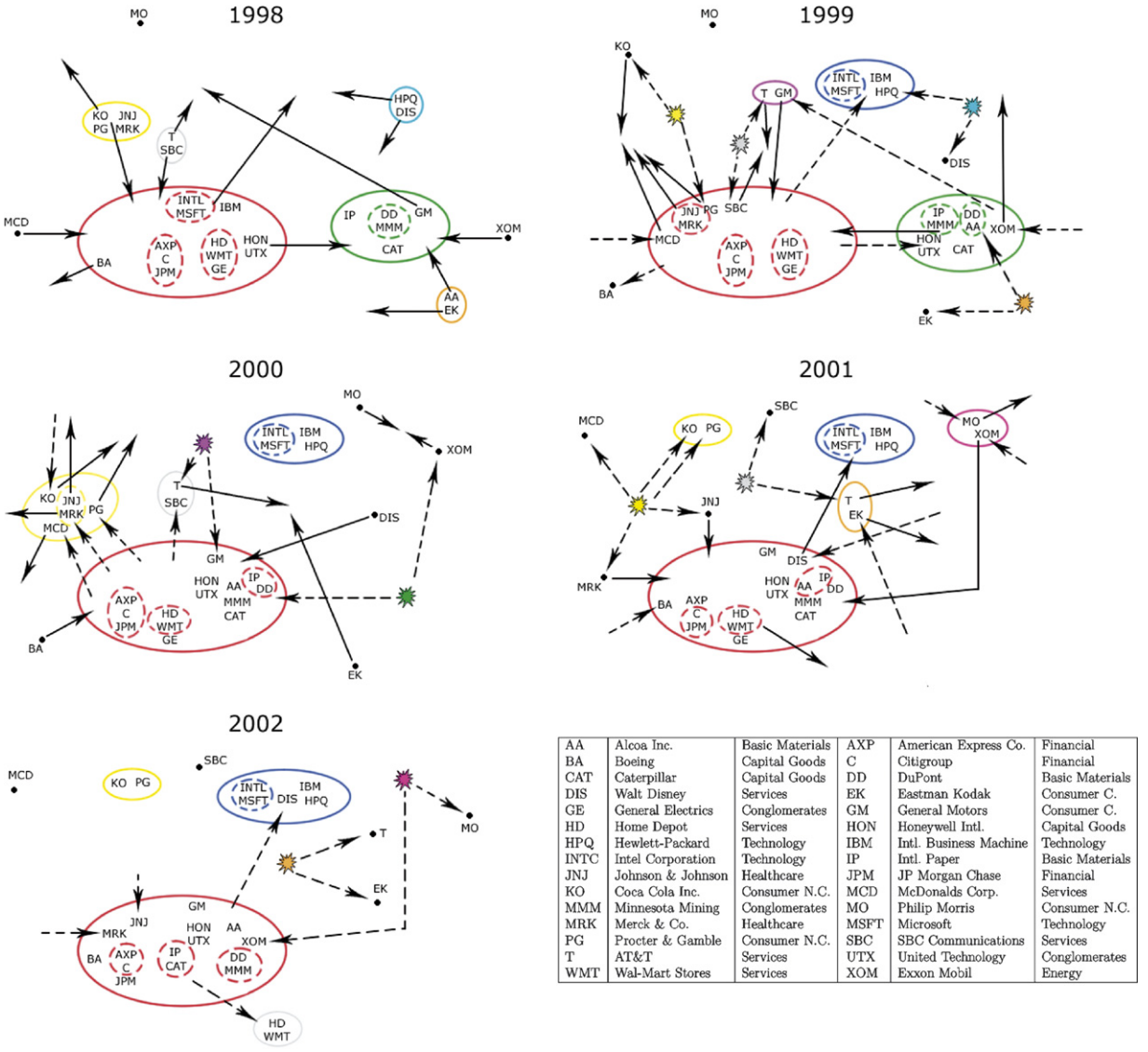


Fig. 4. Hausdorff trees for the years 1999–2002. The acronyms (tickers) are explained in Appendix A.

Let us henceforth focus on Fig. 5. One observes a large super-cluster made up of 10–15 stocks (financial, conglomerates, services, capital goods), containing some homogenous subclusters, which is more or less stable during the whole 5-year period investigated. Some interesting ''paths'' can be outlined, such as the migration from this cluster of the high-tech companies {IBM, INTC, MSFT} between 1998 and 1999. As is well known, 1999 is the year when the high-tech bubble started to grow up. At the end of these two years, they end up forming a separated cluster with HPQ, that remains stable for all the following period.



AA	Alcoa Inc.	Basic Materials	AXP	American Express Co.	Financial
BA	Boeing	Capital Goods	C	Citigroup	Financial
CAT	Caterpillar	Capital Goods	DD	DuPont	Basic Materials
DIS	Walt Disney	Services	EK	Eastman Kodak	Consumer C.
GE	General Electrics	Conglomerates	GM	General Motors	Consumer C.
HD	Home Depot	Services	HON	Honeywell Intl.	Capital Goods
HPQ	Hewlett-Packard	Technology	IBM	Intl. Business Machine	Technology
INTC	Intel Corporation	Technology	IP	Intl. Paper	Basic Materials
JNJ	Johnson & Johnson	Healthcare	JPM	JP Morgan Chase	Financial
KO	Coca Cola Inc.	Consumer N.C.	MCD	McDonalds Corp.	Services
MMM	Minnesota Mining	Conglomerates	MO	Philip Morris	Consumer N.C.
MRK	Merck & Co.	Healthcare	MSFT	Microsoft	Technology
PG	Procter & Gamble	Consumer N.C.	SBC	SBC Communications	Services
T	AT&T	Services	UTX	United Technology	Conglomerates
WMT	Wal-Mart Stores	Services	XOM	Exxon Mobil	Energy

Fig. 5. Clusters obtained by analyzing the daily logarithm closure price difference time series during 1998–2002. The innermost subclusters are indicated with a dashed bubble. Dashed arrows = past; full arrows = future. The position of the points representing the stocks is not directly related to the distance matrix (15) and has no effective "spatial" meaning: the pictorial representation simply reflects the aggregation of points and subclusters into larger clusters. Bottom right: acronyms (tickers) of the stocks and related industries (C = Cyclical; NC = Non-Cyclical; Intl = International).

We emphasize that these remarks are not an input of our analysis: our clustering algorithm is purely *mathematical*, and no genuinely "economical" information (e.g., on industrial homogeneity) was used at the outset. In this sense the position and movements of the stocks in the figures are implied from the market itself.

The definition of the mutual positioning of companies can have an immediate pertinence in a matter of great interest for financial institutions: the portfolio optimization. In a few words (and without entering into complex matters), portfolio theory suggests that in order to minimize the risk involved in a financial investment, one should diversify among different assets by choosing those stocks whose price time evolutions are as diverse as possible (it is never safe to put all the eggs into a single basket). Moreover, this strategy must be continuously updated, by changing weights and components, in order to follow the market evolution. In the framework we presented, by investigating the shares' behavior and tracking the evolution of their mutual

interactions, a first, crude portfolio-optimization rule that emerges would be: choose stocks belonging to clusters that are as "distant" as possible from each other. Portfolio optimization being a complex matter, our findings should be compared and combined with other techniques [11–13], in order to define powerful strategies, hinging on diversified analytical tools, and analyze the dynamics and taxonomy of market correlations.

In conclusion, we have introduced a novel clustering procedure based on the Hausdorff distance between sets. This genuinely mathematical method was used to investigate the time evolution of the stocks belonging to the DJIA index. We found the resulting partitions through the 5-year period investigated to be significant from an economical viewpoint and suited to a meaningful *a posteriori* analysis and interpretation. We believe that this technique is able to extract relevant information from the raw market data and yield meaningful hints for the investigation of the mutual time evolution of the stocks. For the same reasons this procedure could be implemented as the first step towards an evolved portfolio selection and optimization procedure.

Acknowledgments

We thank Sabrina Diomede for a discussion and a pertinent remark.

Appendix A. Dow Jones stock market companies

AA	Alcoa Inc.—basic materials
AXP	American Express Co.—financial
BA	Boeing—capital goods
C	Citigroup—financial
CAT	Caterpillar—capital goods
DD	DuPont—basic materials
DIS	Walt Disney—services
EK	Eastman Kodak—consumer cyclical
GE	General Electrics—conglomerates
GM	General Motors—consumer cyclical
HD	Home Depot—services
HON	Honeywell International—capital goods
HPQ	Hewlett-Packard—technology
IBM	International Business Machine—technology
INTC	Intel Corporation—technology
IP	International Paper—basic materials
JNJ	Johnson & Johnson—healthcare
JPM	JP Morgan Chase—financial
KO	Coca Cola Inc.—consumer non-cyclical
MCD	McDonalds Corp.—services
MMM	Minnesota Mining—conglomerates
MO	Philip Morris—consumer non-cyclical
MRK	Merck & Co.—healthcare
MSFT	Microsoft—technology
PG	Procter & Gamble—consumer non-cyclical
SBC	SBC Communications—services
T	AT&T Gamble—services
UTX	United Technology—conglomerates
WMT	Wal—Mart Stores—services
XOM	Exxon Mobil—energy

References

[1] K. Fukunaga, Introduction to Statistical Pattern Recognition, Academic Press, San Diego, 1990.
[2] A.K. Jain, M.N. Murty, P.J. Flynn, ACM Comput. Surv. 31 (1999) 264.
[3] A.K. Jain, R.C. Dubes, Algorithms for Clustering Data, Prentice-Hall, New York, 1988.
[4] A. Gersho, R.M. Gray, Vector Quantization and Signal Processing, Kluwer Academic Publisher, Boston, 1992.
[5] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, Wiley, New York, 2002.
[6] T. Hofmann, J.M. Buhmann, Pairwise data clustering by deterministic annealing, IEEE Trans. Pattern Anal. Mach. Intell. 19 (1997) 1.
[7] F. Hausdorff, Grundzüge der Mengenlehre, von Veit, Leipzig, 1914 [Republished as Set Theory, fifth ed., Chelsea, New York, 2001].
[8] R.N. Mantegna, Eur. Phys. J. B 11 (1999) 193.
[9] R.N. Mantegna, H.E. Stanley, Introduction to Econophysics, Cambridge University Press, Cambridge, 2000.
[10] M. Bernaschi, L. Grilli, D. Vergni, Physica A 308 (2002) 381;
L. Grilli, Physica A 332 (2004) 441.
[11] E.J. Elton, M.J. Gruber, Modern Portfolio Theory and Investment Analysis, Wiley, New York, 1995.
[12] L. Laloux, P. Cizeau, J.-P. Bouchaud, M. Potters, Phys. Rev. Lett. 83 (1999) 1467;
V. Plerou, P. Gopikrishnan, B. Rosenow, L.A.N. Amaral, H.E. Stanley, Phys. Rev. Lett. 83 (1999) 1471.
[13] J.-P. Onnela, A. Chakraborti, K. Kaski, J. Kertesz, A. Kanto, Phys. Rev. E 68 (2003) 056110.