Journal of Complex Networks (2015) **3**, 303–318
doi:10.1093/comnet/cnu019
Advance Access publication on 4 June 2014

Visualizing the clustering of financial networks and profitability of stocks

C. M. CHEN[†]
Department of Physics, National Taiwan Normal University, Taipei 116, Taiwan
[†]Corresponding author. Email: cchen@phy.ntnu.edu.tw

AND

Y. F. CHANG
Department of Physics, National Taiwan Normal University, Taipei 116, Taiwan
College of Science, China Three Gorges University, YiChang, HuBei 443002, China

Edited by: Guido Caldarelli

[Received on 21 October 2013; accepted on 8 May 2014]

We propose an approach to visualize the clustering of financial networks and a long-term profitability of stocks using financial time series data, by combining several methods of quantitative analysis. For demonstration purposes, this method is applied to investigate the network of Dow Jones Industrial Average (DJIA). Based on the time series data of stock prices during 31 July 2007 to 18 July 2011, our classification method clusters the DJIA components into five groups according to their profitability and property. By comparing the time correlation in the adjusted close price of stocks within the same group, we show that our clustering method results in a better classification of DJIA components than the methods of industry clustering and hierarchical clustering. With this integrated method, we have constructed a two-dimensional map of the DJIA network for visualization, and have related the first and second coordinates of DJIA components in the map to, respectively, their long-term profitability and property. Our analyses show very strong correlations for the sectors of Energy, Basic Materials, Technology, Capital Goods and Consumer/Non-Cyclical, significant correlations for the sectors of Services and Financial and a poor correlation for the Healthcare sector.

Keywords: visualization; clustering; financial networks; stock profitability.

1. Introduction

Stock markets are examples of complex networks, in which the relationship among various stock components can be studied by analysing large amounts of stock price data. Time series forecasting takes the analysis from historical time series data and tries to predict the time evolution of these financial networks. This prediction is of great importance to investors or traders, who are keen to maximize their financial profit. Such an analysis of financial time series data has attracted substantial attention in recent years. However, due to a wide range of factors influencing the fluctuation of the stock markets, it is still difficult to accurately predict the future trend of stock prices based on historical data analysis.

The analysis of financial time series data is helpful in identifying homogeneous groups of stocks. A reliable classification scheme of stocks can provide a synthetic and informative description of complex financial databases, and the resulted groupings have significant implications for portfolio management. Such an approach of statistical data analysis has various other applications, including machine learning,

data mining, pattern recognition, image analysis, information retrieval and bioinformatics. A number of clustering methods have been developed, as a tool for handling large and heterogeneous collections of systems, such as hierarchical clustering (HC) [1–3], minimum spanning tree (MST) [4–6], K-means clustering [2,3,7] and affinity propagation [8,9]. However, there are two major flaws in these clustering methods, including user intrusion and excessive computation time for large data sets. For example, HC recursively merges all components into a cluster, and thus requires human intervention to stop the merging process, when a specific level of classification is reached. A user-defined level of classification, in general, is arbitrary, and not related to the characteristic properties of the investigated system. Moreover, HC involves the calculation of similarities between all pairs of components for each clustering level, and is not efficient for handling large data sets. Several algorithms have been proposed for reducing computation time in cluster analysis by using parallel computation techniques or at the cost of the clustering results [2,10,11]. However, most clustering algorithms still require pre-given assumptions as inputs for the classification of complex systems, including the number of clusters, cluster sizes and boundary conditions. Therefore, it is desirable to have a convenient and operable clustering algorithm that can efficiently cluster large complex systems.

Stocks are often classified based on companies' property, value or expected returns [12–14]. Early attempts to partition stocks into similar groups, proposed by Elton and Gruber [15] and Farrell [16], used heuristic approaches. A more popular method of stock classification is based on the industrial affiliation of stocks, and it demonstrates that industrial factors capture a large portion of the extra-market correlation in stock returns [17]. A variety of industry classification systems have been adopted for defining stock affiliation, such as the Standard Industrial Classification, the Global Industry Classification System and the Fama and French [18] system. Weiner [19] evaluated several industrial classification schemes, and found a number of drawbacks for each one. On the other hand, a significant amount of statistical classification models have been applied to common stock analysis, in the areas including common stock investment categories, price-earnings and return-risk equity classification, information content and return performance, and capital structure [20]. Nevertheless, few of these studies are wholly convincing, due to methodological issues and/or mediocre results. Therefore, there is great incentive for further research to improve stock classification schemes.

We have previously developed the minimum span clustering (MSC) algorithm for the classification of large complex systems, and have demonstrated an excellent MSC classification of the social science network consisting of 1575 *Social Sciences Citation Index* (*SSCI*) journals [21]. Our MSC results showed that the social science network contains 4 resolution scales, including 294 batches of journals at the highest scale, 65 categories of journal batches at the second, 15 research groups of journal categories at the third, and 3 knowledge domains at the lowest resolution. By comparing the relatedness of journals within clusters, the MSC method is shown to present a better classification of SSCI journals than the methods of ISI's heuristic approach and HC. Particularly, no free tuning variable is required in the MSC method. In this study, based on the time series of stock trading prices, we hereby propose an approach for clustering and visualizing financial networks by combining dynamic time warping (DTW) [22], MSC, MST, Sammon mapping (SM) and Classical Multi-dimensional Scaling (CMDS). Although the proposed MSC method has been shown to adequately cluster networks of size about 2000 nodes, we applied MSC to study the Dow Jones Industrial Average (DJIA) network due to the following reasons: (1) the network has a relatively small number of components, which allows the possibility to construct a two-dimensional map of DJIA components without substantial distortion, and the properties of the two principal axes of the map can be thoroughly investigated (2) the components of DJIA do not change frequently and (3) DJIA has closely tracked the performance of a broader value-weighted index of stocks. It has been recently observed that information concerning highly followed stocks tends to price other

TABLE 1 *Components of the DJIA network*

Stock	Code	Sector
3M Co.	MMM	Capital Goods
Alcoa, Inc.	AA	Basic Materials
American Express Company	AXP	Financial
AT&T, Inc.	T	Services
Bank of America Corporation	BAC	Financial
Boeing Co.	BA	Capital Goods
Caterpillar Inc.	CAT	Capital Goods
Chevron Corp.	CVX	Energy
Cisco Systems, Inc.	CSCO	Technology
E I DuPont de Nemours & Co.	DD	Basic Materials
Exxon Mobil Corporation	XOM	Energy
General Electric Co.	GE	Financial
Hewlett-Packard Company	HPQ	Technology
Intel Corporation	INTC	Technology
International Business Machines Corp.	IBM	Technology
Johnson & Johnson	JNJ	Healthcare
JPMorgan Chase & Co.	JPM	Financial
McDonald's Corp.	MCD	Services
Merck & Co. Inc.	MRK	Healthcare
Microsoft Corporation	MSFT	Technology
Mondelez International, Inc.	MDLZ	Consumer/Non-Cyclical
Pfizer Inc.	PFE	Healthcare
Procter & Gamble Co.	PG	Consumer/Non-Cyclical
The Coca-Cola Company	KO	Consumer/Non-Cyclical
The Home Depot, Inc.	HD	Services
The Travelers Companies, Inc.	TRV	Financial
United Technologies Corp.	UTX	Capital Goods
Verizon Communications Inc.	VZ	Services
Wal-Mart Stores Inc.	WMT	Services
Walt Disney Co.	DIS	Services

less followed stocks in the same industry, and such a spill over effect of information becomes more prominent in industries where analysts follow fewer stocks [23]. Therefore, we expect that an accurate classification and visualization of the DJIA network could serve as a seed to further investigate broader stock networks, such as the New York Stock Exchange network.

The following section describes the financial time series data of the DJIA network, and our proposed distance measure for the data, which is calculated using DTW. In Section 3, we then delineate the general procedure to classify and visualize the DJIA network by combining MSC, MST and SM-optimized CMDS. Section 4 presents our results of DJIA classification and visualization, as well as our interpretation of the two-dimensional map of the network. Finally, Section 5 gives our conclusion.

2. Distance measure for financial time series

The analysis of financial time series data is of primary significance in the economic world, which requires both static and dynamic information. For instance, DJIA is an index which shows how 30 US

blue-chip stocks, as shown in Table 1, have been traded during a standard trading session in the stock market. Traditionally, these 30 components are classified into eight different sectors (static information), including basic materials (AA, DD), Capital Goods (BA, CAT, MMM, UTX), Consumer/Non-Cyclical (MDLZ, KO, PG), Energy (CVX, XOM), Financial (AXP, BAC, GE, JPM and TRV), Healthcare (JNJ, MRK, PFE), Services (DIS, HD, MCD, T, VZ, WMT), and Technology (CSCO, HPQ, IBM, INTC, MSFT). In each trading day, the price of its stock components starts to vary at 9:30 AM (the opening price), fluctuates every minute (the highest and lowest prices) and closes at 4:00 PM (the closing price). The movement in the price of a stock component, over time, can be described by an open-high-low-close price chart. In addition, to account for all the corporate actions such as stock splits, dividends/distributions and rights offerings, an adjusted close price (ACP), denoted as $\{p(t)\}$, is often used when examining historical returns or performing a detailed analysis on historical returns. Here, the daily return of a stock is defined as $r(t) = \log(p(t)) - \log(p(t-1))$. Thus, the dynamic information of DJIA components is provided through the financial time series recording the multivariate data (denoted as $\{\vec{p}(t)\}$ of these five prices (open, high, low, close, adjusted close) for every trading day. In this study, the financial time series of DJIA components' prices and trading volume have been downloaded from YAHOO finance (http://finance.yahoo.com), with a range lying between 31 July 2007 and 18 July 18 (a total of 1000 trading days).

Table 2 summarizes the statistics (mean, standard deviation, skewness, kurtosis and Ljung–Box test for autocorrelation (24)) for the daily return of 30 DJIA components. During the time range of our collected data, 11 components have a negative mean of returns, primarily belonging to sectors of Financial and Technology. The standard deviation (0.018–0.052) of those stocks with a negative mean return is usually larger than that (0.012–0.026) of stocks with a positive mean return, implying a larger volatility for stocks in these two sectors. Furthermore, skewness indicates evidence of asymmetry in the distribution of our return data, and kurtosis refers to the degree of peak in the distribution. A positive skewness implies frequent small losses and a few extreme gains, while a negative skewness implies just the opposite. For DJIA, two-thirds of all components have been found to have positive skewness, and all 30 components have a kurtosis >3 (the highest being BAC, CVX, XOM, JNJ, KO and TRV). The observed leptokurtosis of return distributions in DJIA components means that these distributions have fatter tails, and that there is larger chance of extreme outcomes, compared with a normal distribution. This phenomenon has been observed before, and is known to be typical of most asset returns, such as stock, bond, commodity and energy returns. The Ljung–Box test shows that there are no significant autocorrelations up to order 20 in the returns of IBM and JPM.

In stock-based normalization, we first calculate the mean stock price μ_i and standard deviation σ_i for each stock i, and obtain the normalized financial time series for the price of DJIA components from $p_i(t)$:

$$P_i(t) = \frac{p_i(t) - \mu_i}{\sigma_i}. \tag{1}$$

Then, we quantify the degree of dissimilarity (distance) among the derived normalized time series. Being symmetric and non-negative, dissimilarity is small (close to zero) when two time series are similar to each other, while large when two time series differ greatly. For a pair of time series i and j, a popular choice of the dissimilarity measure ($d_{i,j}$) is the Pearson correlation coefficient ($\rho_{i,j}$), i.e. $d_{i,j} = [2(1 - \rho_{i,j})]^{0.5}$. Although this metric can be useful in clarifying the structure of stock returns movement, there are several drawbacks in its definition. First, it does not take into account the stochastic volatility dependence of these time series. Two time series could be strongly correlated, but have very different internal stochastic dynamics [25]. Secondly, the difference in the response time of stock prices

TABLE 2 *Statistics of stock returns for DJIA components*

Stock	Mean × 100	Std. dev. × 100	Skewness	Kurtosis	$Q(20)$
MMM	0.0172	1.75	−0.057	4.29	36.4*
AA	−0.0844	3.78	−0.161	5.26	48.6*
AXP	−0.0048	3.57	0.096	5.11	54.6*
T	−0.0039	1.83	0.687	8.96	95.4*
BAC	−0.1471	5.21	−0.138	10.40	67.6*
BA	−0.0293	2.29	0.219	3.46	35.1*
CAT	0.0426	2.64	0.097	3.35	45.6*
CVX	0.0353	2.18	0.229	12.56	118.9*
CSCO	−0.0621	2.34	−0.743	8.02	34.7*
DD	0.0305	2.32	−0.288	3.97	69.7*
XOM	0.0058	2.03	0.170	12.55	44.7*
GE	−0.0592	2.74	0.070	6.16	38.3*
HPQ	−0.0242	2.15	0.170	5.32	147.7*
INTC	0.0057	2.33	−0.088	3.69	40.9*
IBM	0.0532	1.66	0.171	4.36	25.2
JNJ	0.0230	1.22	0.665	14.34	85.4*
JPM	−0.0022	3.88	0.325	7.65	26.6
MCD	0.0716	1.48	0.044	4.41	43.4*
MRK	−0.0160	2.13	−0.549	8.52	33.3*
MSFT	−0.0005	2.17	0.353	7.73	56.2*
MDLZ	0.0227	1.48	−0.331	5.13	86.1*
PFE	0.0021	1.80	−0.041	4.80	33.3*
PG	0.0148	1.37	−0.186	6.80	57.0*
KO	0.0373	1.48	0.684	11.15	64.8*
HD	0.0094	2.28	0.470	3.25	47.7*
TRV	0.0213	2.60	0.338	14.65	119.5*
UTX	0.0276	1.91	0.494	5.69	61.9*
VZ	0.0076	1.77	0.378	6.78	55.4*
WMT	0.0234	1.46	0.197	7.36	89.5*
DIS	0.0209	2.23	0.424	5.89	49.3*

*Significant at the 1% (5%) level; $Q(20)$ is the Ljung–Box statistic with 20 lags.

to change of information is not considered. Thirdly, it cannot serve as a direct tool for comparing and grouping stocks with unequal sample sizes. Fourthly, the visualization of stocks based on this definition is often distorted, since the distance between two stocks is constrained to lie within the interval [0, 1].

In this article, we propose a different measure for dissimilarity between a pair of financial time series with DTW. First introduced in 60s, and extensively explored in 70s, DTW has a wide range of uses, including handwriting and online signature matching, sign language and gestures recognition, data mining and time series clustering, computer vision and animation, music and signal processing and protein sequence alignment. As schematically illustrated in Fig. 1, DTW uses a dynamic programming approach to align non-linearly two time series of multivariate data in the time dimension so that their dissimilarity is efficiently minimized. Consider three different time series, named I, II and III, of unequal sample sizes, as shown in Fig. 1(A). In this case, to information change in financial markets, the response of time series I is different from that of II and III, while the difference between II and III is only their

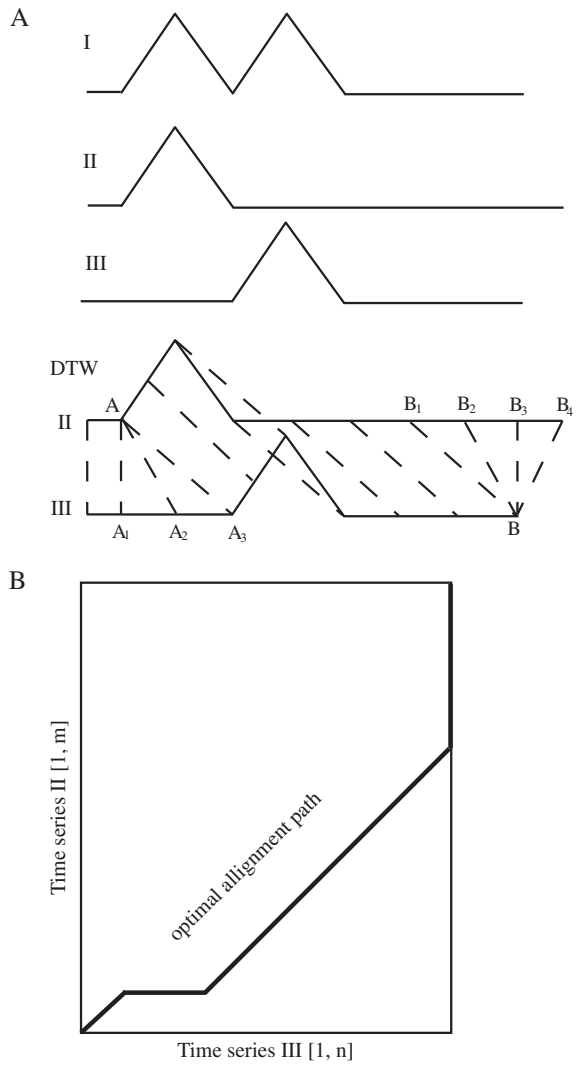

FIG. 1. Schematic illustration of the DTW method. (A) Three time series I, II and III are compared for their dissimilarity (distance) of, and DTW suggests II and III are closer to each other. (B) DTW finds the optimal alignment path on the plane of time series II and III.

response time. Clearly, it is inappropriate to conclude that time series I is more similar to II than III from the dissimilarity measure with the Pearson correlation coefficient. On the other hand, as demonstrated in Fig. 1(A), the DTW process of comparing II and III shows a stretching point A in II (corresponding to A_1, A_2 and A_3 in III) and a stretching point B in III (corresponding to B_1, B_2, B_3 and B_4 in II). In this way, it has been found that time series II and III are closely related to each other. Specifically, the pattern detection process involves searching time series III ($^{\mathrm{III}}P_1$, $^{\mathrm{III}}P_2$,..., $^{\mathrm{III}}P_i$,..., $^{\mathrm{III}}P_n$) for the template time series II ($^{\mathrm{II}}P_1$, $^{\mathrm{II}}P_2$,..., $^{\mathrm{II}}P_j$,..., $^{\mathrm{II}}P_m$). As shown in Fig. 1(B), the time series III and II can be arranged to form a $n \times m$ grid, in which each grid point, (i, j), corresponds to an alignment between elements II_i

and III_j. An optimal alignment path, $\mathbf{W} = (w_1, w_2, \ldots, w_k, \ldots, w_l)$, maps the elements of III and II, so that the dissimilarity between them is minimized. That is to say, $\mathbf{W}$ is a sequence of grid points, and w_k corresponds to a point (i, j). To formulate a dynamic programming problem, we define a dissimilarity measure between two elements as

$$\delta(i,j) = \begin{cases} ({}^{\mathrm{III}}P_i - {}^{\mathrm{II}}P_j)^2, & \text{for single variable data} \\ ({}^{\mathrm{III}}\vec{P}_i - {}^{\mathrm{II}}\vec{P}_j)2, & \text{for multivariable data} \end{cases} \tag{2}$$

and the overall dissimilarity, between time series II and III, is

$$d_{\mathrm{II,III}} = \min_W \left[\sum_{k=1}^{l} \delta(w_k) \right]. \tag{3}$$

3. Clustering and visualizing the DJIA network

The DJIA financial network is rather complex. Its visualization can be implemented by first clustering the network into domains using MSC, delineating the general structure of these domains with MST and visualizing the relationship among various domains in a map using SM-optimized CMDS.

To classify the DJIA network, we applied MSC to decompose it into several groups. In the MSC procedure, one divides a network into several clusters, whose members are associated with their closest neighbours. Each constructed cluster is then considered as a coarse-grained component and the resulted coarse-grained network can be further clustered by MSC. The details of implementing MSC can be obtained from our earlier publication [21]. In addition, the structure of the DJIA network can be described by constructing its MST with the Kruskal algorithm [4,5]. MST is a spanning tree for which the sum of distances among connected components is the smallest, and the Kruskal algorithm constructs the MST by connecting components in the order of increasing distance but avoiding those connections which forms loops in the network graph.

Although MST delineates the general structure of the DJIA network, specific relationships among unconnected components are not preserved in the graph. In order to construct a two-dimensional map of the DJIA network preserving all relationships among network components, we apply CMDS to transform the high-dimensional structure of the DJIA network to its low-dimensional representation. CMDS first calculates eigenvalues and eigenvectors of the distance matrix by principal coordinate analysis, and then projects the DJIA network onto a two-dimensional plane spanned by the two dominant principal axes, which usually describe important features of the network. In other words, CMDS finds a set of two-dimensional vectors $\{\mathbf{x}^\mu\}$ such that the squared distance matrix between the $\{\mathbf{x}^\mu\}$ points matches $\{d_{i,j}^2\}$ in equation (3) as closely as possible. The quality of this two-dimensional representation of the DJIA network in general depends on the distribution of eigenvalues, particularly the largest two eigenvalues. To minimize possible distortion of network structure, the coordinates of network components in the two-dimensional map are further optimized by SM, which tries to preserve the structure of inter-component distances in high-dimensional space on the low-dimension projection. Specifically, the optimization is implemented by minimizing the Sammon stress

$$E = \frac{1}{\sum_{i<j} d_{i,j}} \sum_{i<j} \frac{di,j - d(x^i, x^j)}{d_{i,j}}, \tag{4}$$

where the summation runs over the dataset under investigation, and $d(a,b)$ is the distance between points a and b.

4. Results and discussion

To demonstrate the applicability of MSC in clustering financial networks, we consider the classification of 30 DJIA components using 3 sets of financial time series, including ACP, trading volume and five prices data. The last set is a multivariate time series consisting of daily information about five different prices (open, high, low, close, adjusted close) of DJIA components. Since DJIA's size is relatively small, only one resolution level of MSC is implemented in this study. In our MSC classification scheme, as shown in Table 3, 30 DJIA components are divided into 7 groups with ACP data, 3 groups with trading volume data and 5 groups with five prices data. For the classification with trading volume, Group 3 contains MSFT and INTC, Group 2 contains T, VZ, CSCO and HPQ, while the rest of DJIA components is in Group 1. This result suggests that these technology-related components have a unique pattern of trading volume. In addition, the classification with ACP data is roughly consistent with that with five prices data: Groups 4 and 5 in the ACP-based classification are merged into Group 4 of the five prices-based classification, and Group 6 in the ACP-based classification is divided and merged with Groups 1 and 2 in the five-prices based classification. The classification with five prices data includes: Group 1 containing MSFT, WMT, PG, JNJ and MRK; Group 2 containing MMM, AA, AXP, BA, CAT, DD, GE, HD, INTC, JPM, TRV, UTX and DIS; Group 3 containing PFE, XOM, T and VZ; Group 4 containing CVX, KFT, IBM, KO and MCD; and Group 5 containing BAC, CSCO and HPO.

Two major factors have been found to contribute to the five-prices based classification of DJIA components, including their "profitability" and "property". Here, "profitability" of components refers to the change of their weekly ACP after a specific range of time, while "property" of components is defined by their industrial category as listed in Table 1. It is observed that, if two components are similarly profitable or have similar property, their price pattern would be similar to each other. For example, KO and IBM are, respectively, the most profitable companies in the Consumer/Non-Cyclical and Technology sectors, and the time correlation of their ACP time series from 31 July 2007 to 18 July 2011 is 0.88. In addition, T and VZ both provide telecommunication services, and the time correlation of their ACP time series in the above time range is 0.86. In Table 3, we also display the change in the weekly ACP of stocks in a 3-year range (02 January 2009 to 31 December 2011), in which only AA, BAC and HPQ have a loss on their weekly ACP. To rationalize the MSC classification, we examine the average change of the weekly ACP and its standard deviation in the above 3-year range for each classified group. For the original DJIA network, the average change of the weekly ACP is 43.72%, and its standard deviation is 46.77%. For MSC classified groups, the average change in the weekly ACP and its standard deviation are the following: 25.1 and 12.5% for Group 1, 62.9 and 48.4% for Group 2, 27.1 and 12.1% for Group 3, 74.3 and 28.2% for Group 4, -26.9 and 34.3% for Group 5. In general, the standard deviation of the weekly ACP change in most MSC groups (except for Group 2) is considerably smaller than that of the original DJIA network, suggesting that our classification scheme can be well explained by the profitability of DJIA components. In terms of the stock price performance, Group 4 is considered as a winner group while Group 5 is a loser group. The other factor that also affects our classification scheme is the property of DJIA components. For example, all components in the sector of Capital Goods are clustered into Group 2. However, since the profitability factor seems to play a more important role in our classification scheme, there exists significant discrepancy between our MSC classification and the classification by industry clustering. Note that profitability and property of components, respectively, correspond to the largest and second largest eigenvalue in the CMDS mapping, which will be discussed later in this section.

In addition to the clustering of DJIA components with MSC or industry clustering, a widely used clustering method is HC, which builds a binary tree of the data that successively merges similar groups of components. The clustering results of the DJIA network by using HC with five prices data are

TABLE 3 *MSC classification results with three sets of financial data, including the time series of ACP, trading volume and five prices*

ACP		Volume		Five prices		
Company	Group	Company	Group	Company	Group	Price change* (%)
WMT	1	WMT	1	MSFT	1	36.9
PG	1	PG	1	WMT	1	12.25
JNJ	1	JNJ	1	PG	1	16.61
MMM	2	MMM	1	JNJ	1	19.9
AA	2	AA	1	MRK	1	40
AXP	2	AXP	1	MMM	2	50.22
BA	2	BA	1	AA	2	−25.09
CAT	2	CAT	1	AXP	2	161.29
DD	2	DD	1	BA	2	76.84
GE	2	GE	1	CAT	2	110.66
HD	2	HD	1	DD	2	99.38
TRV	2	TRV	1	GE	2	17.33
UTX	2	UTX	1	HD	2	92.23
DIS	2	DIS	1	INTC	2	76.12
T	3	BAC	1	JPM	2	9.38
VZ	3	MRK	1	TRV	2	42.42
PFE	3	PFE	1	UTX	2	43.41
XOM	3	XOM	1	DIS	2	62.97
CVX	4	CVX	1	PFE	3	35.5
IBM	4	IBM	1	XOM	3	11.67
KO	4	KO	1	T	3	23.4
MCD	5	MCD	1	VZ	3	37.96
MDLZ	5	MDLZ	1	CVX	4	54.47
INTC	6	JPM	1	MDLZ	4	53.72
JPM	6	T	2	IBM	4	122.46
MRK	6	VZ	2	KO	4	67.45
MSFT	6	CSCO	2	MCD	4	73.34
BAC	7	HPQ	2	BAC	5	−60.71
CSCO	7	INTC	3	CSCO	5	7.84
HPQ	7	MSFT	3	HPQ	5	−27.85

*Price change of stocks is calculated for the time range between 02 January 2009 and 31 December 2011.

demonstrated in Fig. 2, in which the *x*-axis shows various groups of DJIA components and the *y*-axis shows the distance between neighbouring groups. HC has an explicit procedure and clearly interpretable results, but its decision of how many groups to use is arbitrary. To compare with MSC results, the five HC classified groups of DJIA are underlined in Fig. 2. The HC classification includes three isolated components (Groups II, III and V), a small group (Group IV) and a large group (Group I). As far as the weekly ACP change is concerned, Groups II and III are winners, while Groups IV and V are losers. For comparison purposes, members of Groups IV and V in HC belong to the same loser group (Group 5) in MSC, while members of Groups II and III in HC belong to Group 2 in MSC. Overall, MSC provides a better classification than HC, since the majority of DJIA components is not classified in HC. Unlike HC, MSC requires no input of the number of clusters, which is a free parameter in HC. Furthermore,

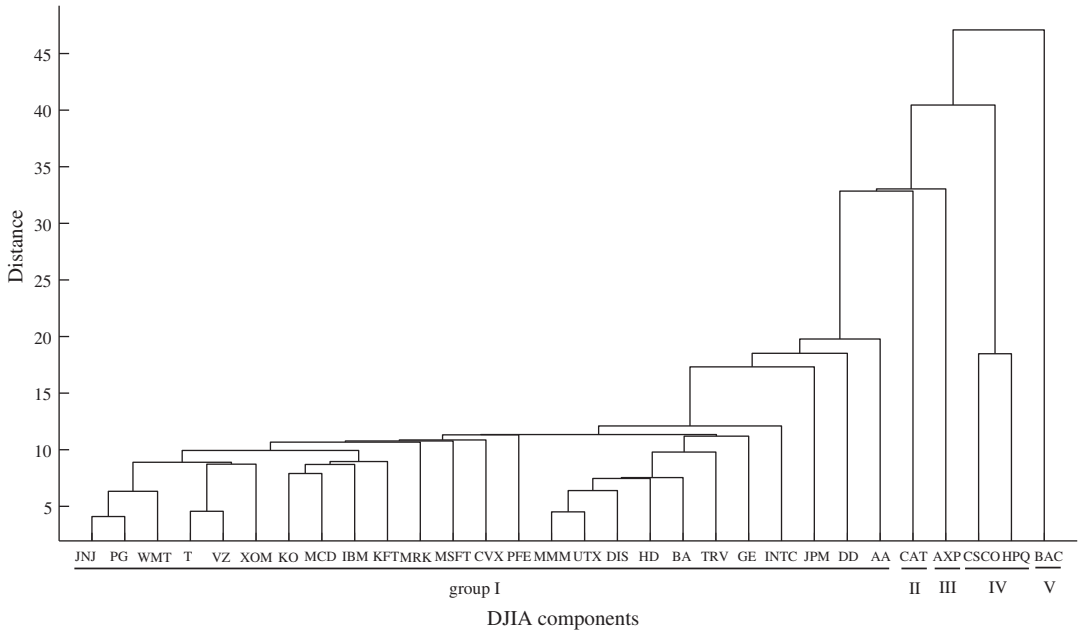

FIG. 2. HC classification of the DJIA network based on the DTW distance between its components. To compare with the MSC results, here the number of HC groups is manually set to 5.

the members within the same cluster predicted by MSC tend to be more closely related to each other than those predicted by HC. This last comparison will be further demonstrated in Fig. 3.

The MSC classification method has been shown previously to optimize the overall relatedness between components within the same group for the social science network [21]. For the DJIA network, in Fig. 3, we calculate the accumulated distribution of time correlation between components within the same group for MSC, HC, industry clustering and the original DJIA network. In MSC, 80% of the calculated time correlations have a value >0.5, indicating a strong correlation between components within the same group. However, this percentage is only $\sim$72% in HC (with five groups), and 66% in both the original network and industry clustering. The comparison in Fig. 3 concludes that, for the time correlation in the ACP of components within the same group, MSC $>$ HC $>$ industry clustering $\geqslant$ original DJIA network.

To visualize the relationships among DJIA components, we have implemented the SM-optimized CMDS mapping of the DJIA network. Figure 4 integrates this CMDS mapping of the DJIA network with its MST diagram and MSC grouping results. Here, we denote members of five MSC groups with different symbols: ★ for Group 1, ● for Group 2, ⬢ for Group 3, ■ for Group 4 and ◉ for Group 5. In the MST diagram, we use solid lines to represent intra-group connections, and dashed lines for inter-group connections. From Fig. 4, it is clear that the results of MSC, MST and SM-optimized CMDS mapping are consistent with each other. In this map, related components tend to aggregate together if their profitability is not notably different. For example, the components in the Consumer/Non-Cyclical sector (PG, MDLZ and KO) as well as those in the Healthcare sector (MRK, JNJ and PFE) appear in the same region, although they are classified into different MSC groups. It is noticed that WMT locates nearby JNJ, PG and MDLZ, since the retailer giant is the main sales channel for the products

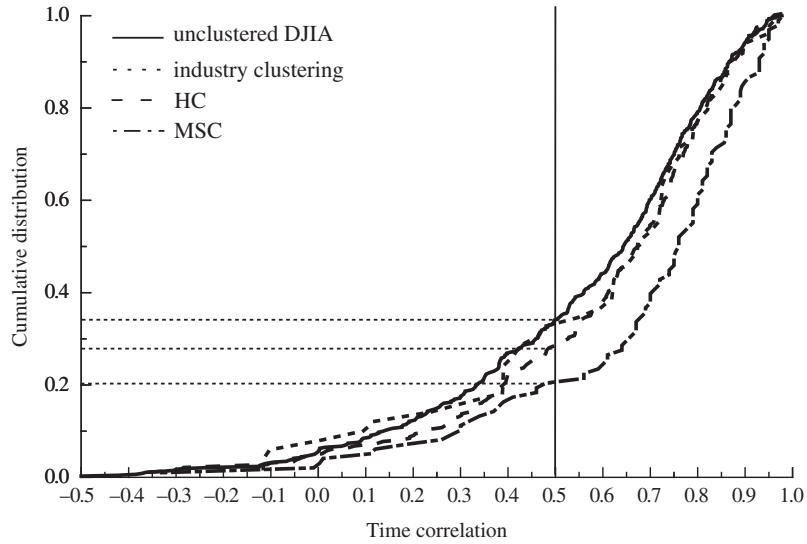

FIG. 3. Cumulative distributions of time correlation between components within the same group for various classification schemes, including MSC, HC, industry clustering and the original DJIA network.

of these three companies. We also notice that MSFT and INTC are close to each other in the map. Since 1989, the Wintel partnership has dominated the PC industry, with a share of personal computing platforms >75%, for ∼30 years. With the increasing popularity of smart phones and tablets, it would be interesting to investigate the future correlation of these two companies.

It is noteworthy that the winner group (Group 4) is located on the right-hand side of the map, while the loser group (Group 5) is on its opposite side. Apparently, the first coordinate of DJIA components (within the same industry sector) highly correlates to their profitability. In Table 4, we analyse the correlation between the first coordinate of DJIA components and the change of their weekly ACP after 1, 2, 3 and 3+ years. Here, we consider three different sets of initial ACP, that are in the week of 02 January 2009 (Set I), 02 March 2009 (Set II) and 01 May 2009 (Set III). We note that the DJIA index reached its lowest value since 2000 on 02 March 2009 due to the 2008 financial crisis, and the weekly ACP change after 02 March 2009 is thus considered as its rebound after hitting the minimum. As stocks tend to be under priced at the minimum (especially for financial components in the 2008 crisis), we also consider the weekly ACP 2 months before or after 02 March 2009 as the benchmark price in this investigation. For almost all sectors (except for Healthcare), our results show that the first coordinate of DJIA components has a strong correlation to the long-term profitability (change of their weekly ACP after 2 or 3 years), but a weak correlation to the short-term profitability (change of their weekly ACP after 1 year). Within these examined periods, the correlation coefficient is 1.00 for both Energy and Basic Materials, 0.91–0.98 for Technology, 0.89–1.00 for Capital Goods, 0.88–0.99 for Consumer/Non-cyclical and 0.72–0.90 for Services. For the Financial sector, the correlation is significant (0.74–0.88) for the benchmark dates 02 January 2009 and 01 May 2009, but is weak (0.31–0.48) for the benchmark date 02 March 2009. The weak correlation between the first coordinate of financial components and their weekly ACP change is mainly due to the drastic drop in the stock price of BAC after the 2008 financial crisis. For Healthcare components, we find no clear correlation (−0.9 to

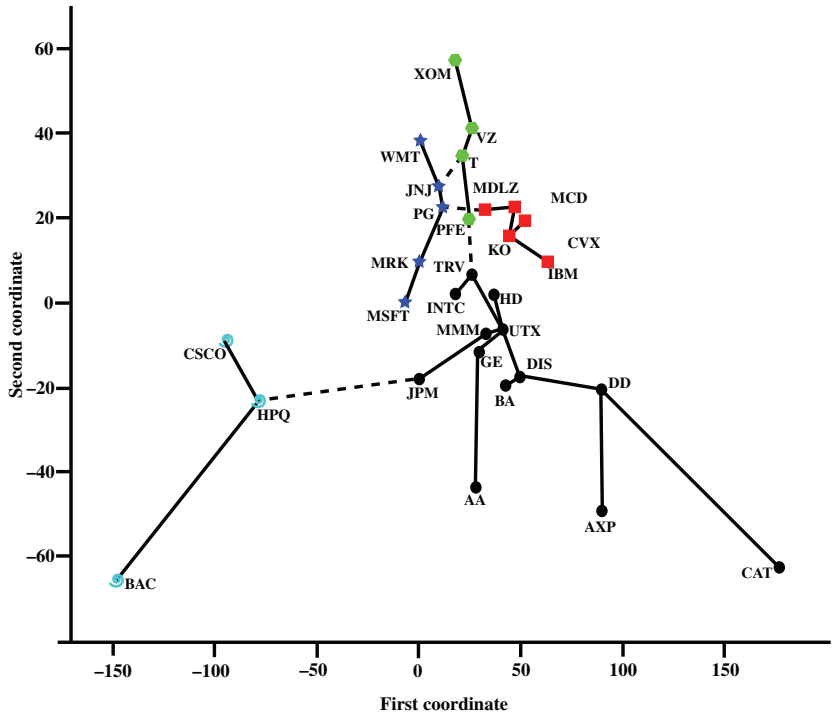

FIG. 4. Two-dimensional map of the DJIA network constructed using SM-optimized CMDS, in which the results of MST and MSC are also delineated. Members of 5 MSC groups are denoted with different symbols: ★ for Group 1, ● for Group 2, ⬣ for Group 3, ■ for Group 4 and 🌀 for Group 5. In the MST diagram, solid lines represent intra-group connections, and dashed lines represent inter-group connections.

0.39) between their first coordinate and ACP change. Finally, in Table 4, we have also included recent financial data (up to 12 November 2012) for our correlation analysis between the first coordinate of DJIA components and their weekly ACP change. In general, we find very strong correlations for Energy, Basic Materials, Technology, Capital Goods and Consumer/Non-cyclical, significant correlations for Services and Financial, and a poor correlation for Healthcare.

Our analysis shows that the correlation between the first coordinate of DJIA components and their weekly ACP change is mainly impaired by BAC in Financial, by HD in Services, and by MRK in Healthcare. To examine whether these three stocks are mispriced, we study their relative valuations, including trailing P/E (Price/Earning per share), forward P/E, trailing P/S (Price/Sales per share) and P/BV (Price/Book Value per share), benchmarked against the rest members in their sector. In the week of 02 March 2009, BAC has trailing P/E 0.75, trailing P/S 0.18 and P/BV 0.14, which are significantly smaller than the corresponding benchmarks 1.47, 0.60 and 0.70 in Financial (forward P/E is not compared due to negative earnings of BAC in the forward year). This comparison explains the poor correlation between the first coordinate and the long-term profitability found in the Financial sector for time range Set II, since the benchmark price of BAC was largely underestimated. In the week of 12 November 2012, HD has trailing P/E 21.97, forward P/E 17.82, trailing P/S 1.28 and P/BV 5.22, while the corresponding benchmarks in services are 25.29, 13.33, 1.61 and 3.26. Since

TABLE 4 *Correlation coefficient between the first coordinate of components and their profitability*

Sector	Time range			
	02 January 2009 to 31 December 2009*	02 January 2009 to 31 December 2010*	02 January 2009 to 31 December 2011*	02 January 2009 to 31 November 2012*
Energy[#]	1.00	1.00	1.00	1.00
Basic Materials[#]	−1.00	1.00	1.00	1.00
Technology	0.56	0.98	0.94	0.91
Capital Goods	−0.39	0.99	0.89	0.83
Consumer/ Non-Cyclical	0.87	0.96	0.99	0.99
Services	0.64	0.81	0.83	0.44
Financial	0.55	0.74	0.87	0.84
Healthcare	−0.93	−0.90	−0.13	−0.22

Sector	Time range			
	02 March 2009 to 01 March 2010*	02 March 2009 to 01 March 2011*	02 March 2009 to 01 March 2012*	02 March 2009 to 01 March 2012*
Energy[#]	1.00	1.00	1.00	1.00
Basic Materials[#]	−1.00	1.00	1.00	1.00
Technology	−0.58	0.91	0.91	0.89
Capital Goods	0.90	1.00	0.99	0.96
Consumer/ Non-Cyclical	−0.67	0.97	0.88	0.95
Services	0.63	0.72	0.79	0.50
Financial	−0.64	0.31	0.48	0.46
Healthcare	−0.65	0.39	0.11	0.02

Sector	Time range			
	01 May 2009 to 30 April 2010*	01 May 2009 to 30 April 2011*	01 May 2009 to 30 April 2012*	01 May 2009 to 30 November 2012*
Energy[#]	1.00	1.00	1.00	1.00
Basic Materials[#]	1.00	1.00	1.00	1.00
Technology	−0.23	0.98	0.94	0.98
Capital Goods	0.72	0.99	0.98	0.99
Consumer/ Non-Cyclical	0.34	0.99	0.97	0.99
Services	0.76	0.87	0.90	0.87
Financial	−0.39	0.75	0.88	0.75
Healthcare	−0.78	0.24		0.24

[#]Both Energy and Basic Materials in DJIA contain only two components.
*Correlation coefficient is calculated for the price change within (beyond) the time range of collected financial data for this study.

only two relative valuation benchmarks show that HD is overpriced, it is not conclusive about the observed decrease in correlation for the Services sector. For the Healthcare sector, in the week of 12 November 2012, MRK has trailing P/E 20.07, forward P/E 12.00, trailing P/S 2.81 and P/BV 2.41, and the corresponding benchmarks are 20.77, 11.57, 2.87 and 2.77, respectively. Since the difference is small in both the relative valuation benchmarks and the first coordinate of Healthcare components, the poor correlation in the Healthcare sector might result from distortions in the two-dimensional projection.

It is also noticed that the second coordinate of the map in Fig. 4 corresponds to the property of DJIA components. Using the industry clustering listed in Table 1, we calculate the average location and the standard deviation for each sector: (38.60, 26.16) for Energy, (20.10, 3.46) for Consumer/Non-Cyclical, (20.10, 23.35) for Services, (18.77, 8.90) for Healthcare, (−3.91, 12.49) for Technology, (−23.60, 26.17) for Capital Goods, (−27.34, 29.02) for Financial and (−31.94, 15.64) for Basic Materials. The standard deviation in the second coordinate of each sector reflects the diversity of the defined sector in DJIA, which is smaller for Consumer/Non-Cyclical and Healthcare, and is larger for Financial and Capital Goods. Note that this proposed correspondence may not be exact due to the distortion in the two-dimensional projection, as evidenced by the fact that, in CMDS, the largest eigenvalue is 3.3 times of the second largest eigenvalue and is 16.4 times of the third largest eigenvalue.

5. Conclusion

In summary, we have hereby proposed an approach to cluster and visualize financial networks by integrating various analysing methods, including DTW, MSC, MST and SM-optimized CMDS. Overall, the MSC method does not require human intervention and is computationally efficient (the computing time is linear in the network size). Furthermore, it presents a better classification for equities than the methods of industry clustering and HC. In addition to the network clustering, our integrated approach can also be used to construct a map of a financial network, which allows us to visualize the relationship among equities intuitively. More importantly, our approach delineates the connection of the two principal coordinates of the map to the long-term profitability and property of equities.

To exemplify this approach, we have investigated the DJIA network using the financial time series data from 31 July 2007 to 18 July 2011. The distance between network components is measured by DTW using single-variable or multivariate data. From this distance matrix, we then classified the network by MSC, and described its topology by MST. Our results suggest that the grouping of network components is mainly by their profitability and property. For example, we identify a winner group in profitability which contains CVX, MDLZ, IBM, KO and MCD, as well as a loser group in profitability which contains BAC, CSCO and HPQ. SM-optimized CMDS mapping is used to construct a two-dimensional map for the visualization of the DJIA network, and its results are consistent with those of MSC and MST. More importantly, in this map, it is found that the first coordinate of DJIA components shows high correlation to their profitability, and the second coordinate of components is related to their industrial category. More specifically, for components in each sector, their weekly ACP change in long terms (2–4 years) highly correlates to their first coordinate in the map. Our analyses show very strong correlations for Energy, Basic Materials, Technology, Capital Goods and Consumer/Non-Cyclical, significant correlations for Services and Financial, and a poor correlation for Healthcare. Such a statistical analysis could provide an important tool for investors to examine their holdings and for companies to evaluate their policies.

Funding

This work is supported by the National Science Council of Taiwan under grant no. NSC 99-2112-M-003-011-MY3.

Acknowledgement

CMC wishes to thank Michael Plischke for his support and stimulating discussions at Simon Fraser University.

References

1. Ward Jr, J. H. (1963) Hierarchical grouping to optimize an objective function. *J. Amer. Statist. Assoc.*, **58**, 236–244.
2. Hastie, T., Tibshirani, R. & Friedman, J. (2008) *The Elements of Statistical Learning*, 2nd edn. Berlin, Heidelberg, New York: Springer.
3. Manning, C. D., Raghavan, P. & Schütze, H. (2008) *Introduction to Information Retrieval*. Cambridge: Cambridge University Press.
4. Samoylenko, I., Chao, T.-C., Liu, W.-C. & Chen, C.-M. (2006) *Visualizing the scientific world and its evolution. J. Amer. Soc. Inform. Sci. Technol.*, **57**, 1461–1469.
5. Cormen, T. H., Leiserson, C. E., Rivest, R. L. & Stein, C. (2001) *Introduction to Algorithms*, 2nd edn. Cambridge: MIT Press.
6. Paivinen, N. (2005) Clustering with a minimum spanning tree of scale-free-like structure. *Pattern Recogn. Lett.*, **26**, 921–930.
7. MacQueen, J. B. (1967) Some Methods for classification and Analysis of Multivariate Observations. *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley: University of California Press.
8. Frey, B. J. & Dueck, D. (2007) Clustering by passing messages between data points. *Science*, **315**, 972–976.
9. Chen, C. M. (2008) Classification of scientific networks using aggregated journal-journal citation relations in the Journal Citation Reports. *J. Amer. Soc. Inform. Sci. Technol.*, **59**, 2296–2304.
10. Kishida, K. (2010) High-speed rough clustering for very large document collections. *J. Amer. Soc. Inform. Sci. Technol.*, **61**, 1092–1104.
11. Fernandez, A. & Gomez, S. (2008) Solving non-uniqueness in agglomerative hierarchical clustering using multidendrograms. *J. Classification*, **25**, 43–65.
12. Bhojraj, S., Lee, C. M. C. & Oler, D. K. (2003) What's my line? A comparison of industry classification schemes for capital market research. *J. Account. Res.*, **41**, 745–774.
13. King, B. F. (1966) Market and industry factors in stock price behavior. *J. Bus.*, **39**, 139–190.
14. Chan, L. K. C., Lakonishok, J. & Swaminathan, B. (2007) Industry classifications and return comovement. *Financial Analysts J.*, **63**, 56–70.
15. Elton, E. J. & Gruber, M. J. (1970) Homogeneous groups and the testing of economic hypotheses. *J. Financial Quant. Anal.*, **4**, 581–602.
16. Farrell Jr, J. L. (1974) Analyzing covariation of returns to determine homogeneous stock groupings. *J. Bus.*, **47**, 186–207.
17. Connor, G. (1995) The three types of factor models: a comparison of their explanatory power. *Financial Analysts J.*, **51**, 42–46.
18. Fama, E. F. & French, K. R. (1997) Industry costs of equity. *J. Financial Econ.*, **43**, 153–193.
19. Weiner, C. (2005) The impact of industry classification schemes on financial research. *SFB 649 Discussion Paper 2005-062*. http://edoc.hu-berlin.de/series/sfb-649-papers/2005-62/PDF/62.pdf. Accessed on 16 July 2013.

20. Altman E. I. (1981) Statistical classification models applied to common stock analysis. *J. Bus. Res.*, **9**, 123–149.
21. Chang, Y. F. & Chen, C. M. (2011) Classification and visualization of the social science network by the minimum span clustering method. *J. Amer. Soc. Inform. Sci. Technol.*, **62**, 2404–2413.
22. Bellman, R. & Kalaba, R. (1959) A Mathematical Theory of Adaptive Control Processes. *Proc. Natl. Acad. Sci. USA*, **45**, 1288–1290.
23. Hameed, A., Morck, R., Shen, J. & Yeung, B. (2010) Information, analysts, and stock return comovement. *NBER Working Paper*.
24. Ljung, G. M. & Box, G. E. P. (1978) On a measure of lack of fit in time series models. *Biometrika*, **65**, 297–303.
25. Caiado, J. & Crato, N. (2010) Identifying common dynamic features in stock returns. *Quant. Finance*, **10**, 797–807.