Available online at www.sciencedirect.com
SCIENCE DIRECT®

Computers & Operations Research 33 (2006) 3171–3184

computers & operations research

www.elsevier.com/locate/cor

Mining market data: A network approach

Vladimir Boginski[a,*], Sergiy Butenko[b], Panos M. Pardalos[a]

[a]*Department of Industrial and Systems Engineering, University of Florida, 303 Weil Hall, Gainesville, FL 32611, USA*
[b]*Department of Industrial Engineering, Texas A&M University, 236E Zachry Engineering Center, College Station, TX 77843-3131, USA*

Available online 23 February 2005

Abstract

We consider a network representation of the stock market data referred to as the market graph, which is constructed by calculating cross-correlations between pairs of stocks based on the opening prices data over a certain period of time. We study the evolution of the structural properties of the market graph over time and draw conclusions regarding the dynamics of the stock market development based on the interpretation of the obtained results.
© 2005 Elsevier Ltd. All rights reserved.

Keywords: Market graph; Data mining; Stock price fluctuations; Cross-correlation; Graph theory; Degree distribution; Power-law model; Clustering; Clique; Independent set

1. Introduction

One of the most important problems in the modern finance is finding efficient ways of summarizing and visualizing the stock market data that would allow one to obtain useful information about the behavior of the market. Nowadays, a great number of stocks are traded in the US stock markets; moreover, this number steadily increases. The amount of data generated by the stock market every day is enormous. This data is usually visualized by thousands of plots reflecting the price of each stock over a certain period of time. The analysis of these plots becomes more and more complicated as the number of stocks grows.

An alternative way of summarizing the stock prices data, which was recently developed, is based on representing the stock market as a graph (network). This graph is referred to as the *market graph*.

* Corresponding author.
E-mail addresses: vb@ufl.edu (V. Boginski), butenko@tamu.edu (S. Butenko), pardalos@ufl.edu (P.M. Pardalos).

0305-0548/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.cor.2005.01.027

It should be noted that the approach of representing a dataset as a network becomes more and more extensively used in various practical applications, finance being one of them [1–6]. This methodology allows one to visualize a dataset by representing its elements as vertices and observe certain relationships between them. Studying the structure of a graph representing a dataset is important for understanding the internal properties of the application it represents, as well as for improving storage organization and information retrieval. One can easily imagine a graph as a set of dots (vertices) and links (edges) connecting them, which in many cases makes this representation convenient and easily understandable.

A natural graph representation of the stock market is based on the cross-correlations of price fluctuations. The market graph is constructed as follows: a vertex represents each stock, and two vertices are connected by an edge if the correlation coefficient of the corresponding pair of stocks (calculated over a certain period of time) exceeds a specified threshold $\theta \in [-1, 1]$.

The formal procedure of constructing the market graph is as follows. Let $P_i(t)$ denote the price of the instrument i on day t. Then $R_i(t) = \ln P_i(t)/P_i(t-1)$ defines the logarithm of return of the instrument i over the one-day period from $(t-1)$ to t. The correlation coefficient between instruments i and j is calculated as

$$C_{ij} = \frac{\langle R_i R_j \rangle - \langle R_i \rangle \langle R_j \rangle}{\sqrt{\langle R_i^2 - \langle R_i \rangle^2 \rangle \langle R_j^2 - \langle R_j \rangle^2 \rangle}}, \tag{1}$$

where $\langle R_i \rangle$ is defined simply as the average return of the instrument i over N considered days (i.e., $\langle R_i \rangle = (1/N)\sum_{t=1}^{N} R_i(t)$) [7].

An edge connecting stocks i and j is added to the graph if $C_{ij} \geqslant \theta$, which means that the prices of these two stocks behave similarly over time, and the degree of this similarity is defined by the chosen value of θ. Therefore, studying the pattern of connections in the market graph would provide helpful information about the internal structure of the stock market.

In our previous research, we have investigated various properties of the market graph constructed using the data for 500 recent consecutive trading days in 2000–2002 [8,9]. In this work, it has been observed that the distribution of the correlation coefficients calculated for all possible pairs of stocks using (1) has the shape similar to a part of a normal distribution with the mean approximately equal to 0.05, and that for the values of the correlation threshold $\theta \geqslant 0.2$ the degree distribution of the market graph follows the *power-law model* [2]. According to this model, the probability that a vertex has a degree k (i.e., there are k edges emanating from it) is

$$P(k) \propto k^{-\gamma} \tag{2}$$

or, equivalently,

$$\log P(k) \propto -\gamma \log k \tag{3}$$

which shows that this distribution can be plotted as a straight line in the logarithmic scale.

An interesting fact is that besides the market graph, many other graphs arising in diverse areas [2,3,5,6,10–15] also have a well-defined power-law structure. This fact served as a motivation to introduce a concept of "self-organized networks", and it turns out that this phenomenon also takes place in finance.

Another contribution of Boginski et al. [8,9] is in a suggestion to relate some correlation-based properties of the market to certain *combinatorial properties* of the corresponding market graph. For example, the authors attacked the problem of finding large groups of highly correlated instruments by applying simple algorithms for finding *cliques* and *independent sets* in the market graph constructed using different values of the correlation threshold.

The main purpose of the present paper is to reveal the *dynamics* of the changes in structural properties of the market graph over time. We consider the graphs constructed based on the stock prices data for different time periods during 1998–2002 and study the *evolution* of certain characteristics of these graphs. This approach helps us to obtain useful information about the trends that take place in the modern stock market.

2. Basic concepts from graph theory and their data mining interpretation

To give a brief introduction to graph theory, we present several basic definitions and notations. We will also discuss the interpretation of the introduced concepts from the data mining perspective.

Let $G=(V,E)$ be an undirected graph with the set of n vertices V and the set of edges $E=\{(i,j): i,j\in V\}$.

2.1. Connectivity and degree distribution

The graph $G=(V,E)$ is *connected* if there is a path from any vertex to any vertex in the set V. If the graph is disconnected, it can be decomposed into several connected subgraphs, which are referred to as the *connected components* of G.

The *degree* of a vertex is the number of edges emanating from it. For every integer number k one can calculate the number of vertices $n(k)$ with the degree equal to k, and then get the probability that a vertex has the degree k as $P(k)=n(k)/n$, where n is the total number of vertices. The function $P(k)$ is referred to as the *degree distribution* of the graph. The degree distribution is an important characteristic of a graph representing a dataset. It reflects the overall pattern of connections in the graph, which in many cases reflects the global properties of the dataset this graph represents. As it was indicated above, many real-life massive graphs representing the datasets coming from diverse areas (Internet, telecommunications, finance, biology, sociology) have degree distributions that follow the *power-law* model (2).

An important characteristic of the power-law model that should be mentioned here is its *scale-free* property. This property implies that the power-law structure of a certain network should not depend on the size of the network. Clearly, real-world networks dynamically grow over time, therefore, the growth process of these networks should obey certain rules in order to satisfy the scale-free property. In [11], the authors point out the necessary properties of the evolution of the real-world networks: *growth* and *preferential attachment*. The first property implies the obvious fact that the size of these networks continuously grows (i.e., new vertices are added to a network, which means that new elements are added to the corresponding dataset). The second property represents the idea that new vertices are more likely to be connected to old vertices with high degrees. The fact that many networks representing different datasets obey the power-law model indicates that the global organization and evolution of massive datasets arising in various spheres of life nowadays follow similar laws and patterns.

2.2. Cliques and independent sets

Given a subset $S \subseteq V$, by $G(S)$ we denote the subgraph induced by S. A subset $C \subseteq V$ is a *clique* if $G(C)$ is a complete graph, i.e., it has all possible edges. The maximum clique problem is to find the largest clique in a graph.

The following definitions generalize the concept of clique. Namely, instead of cliques one can consider dense subgraphs, or *quasi-cliques*. A *γ-clique C_γ*, also called a *quasi-clique*, is a subset of V such that $G(C_\gamma)$ has at least $\gamma q(q-1)/2$ edges, where q is the cardinality (i.e., number of vertices) of C_γ.

An *independent set* is a subset $I \subseteq V$ such that the subgraph $G(I)$ has no edges. The maximum independent set problem can be easily reformulated as the maximum clique problem in the *complementary* graph $\bar{G}(V, \bar{E})$, which is defined as follows. If an edge $(i, j) \in E$, then $(i, j) \notin \bar{E}$, and if $(i, j) \notin E$, then $(i, j) \in \bar{E}$. Clearly, a maximum clique in $\bar{G}$ is a maximum independent set in G, so the maximum clique and maximum independent set problems can be easily reduced to each other.

Locating cliques (quasi-cliques) and independent sets in a graph representing a dataset provides important information about this dataset. Intuitively, edges in such a graph would connect vertices corresponding to "similar" elements of the dataset, therefore, cliques (or quasi-cliques) would naturally represent dense clusters of similar objects. On the contrary, independent sets can be treated as groups of objects that differ from every other object in the group, and this information is also important in some applications. Clearly, it is useful to find a maximum clique or independent set in the graph, since it would give the maximum possible size of the groups of "similar" or "different" objects.

The maximum clique problem (as well as the maximum independent set problem) is known to be NP-hard [16]. Moreover, it turns out that these problems are difficult to approximate [17,18]. This makes these problems especially challenging in large graphs. However, as we will see later, a special structure of the market graph allows us to get the exact solution of the maximum clique problem.

2.3. Clustering via clique partitioning

The problem of locating cliques and independent sets in a graph can be naturally extended to finding an optimal *partition* of a graph into a minimum number of distinct cliques or independent sets. These problems are referred to as *minimum clique partition* and *graph coloring*, respectively. Various mathematical programming formulations of these problems can be found in [19]. Clearly, as in the case of maximum clique and maximum independent set problems, minimum clique partition and graph coloring are reduced to each other by considering the complimentary graph, and both of these problems are NP-hard [16]. Solving these problems for the graphs representing real-life datasets is important from the data mining perspective, in particular, in solving the *clustering* problem.

The essence of *clustering* is partitioning the elements in a certain dataset into several distinct subsets (clusters) grouped according to an appropriate *similarity criterion* [20]. Identifying the groups of objects that are "similar" to each other but "different" from other objects in a given dataset is important in many practical applications. The clustering problem is challenging due to the fact that the number of clusters and the similarity criterion are usually not known a priori.

If a dataset is represented as a graph, where each data element corresponds to a vertex, the clustering problem essentially deals with decomposing this graph into a set of subgraphs (subsets of vertices), so that each of these subgraphs would correspond to a specific cluster. Clearly, since the data elements assigned to the same cluster should be "similar" to each other, the goal of clustering can be achieved by finding

Table 1
Dates and mean correlations corresponding to each 500-day shift

Period #	Starting date	Ending date	Mean correlation
1	09/24/1998	09/15/2000	0.0403
2	12/04/1998	11/27/2000	0.0373
3	02/18/1999	02/08/2001	0.0381
4	04/30/1999	04/23/2001	0.0426
5	07/13/1999	07/03/2001	0.0444
6	09/22/1999	09/19/2001	0.0465
7	12/02/1999	11/29/2001	0.0545
8	02/14/2000	02/12/2002	0.0561
9	04/26/2000	04/25/2002	0.0528
10	07/07/2000	07/08/2002	0.0570
11	09/18/2000	09/17/2002	0.0672

a clique partition of the graph, and the number of clusters will be equal to the number of cliques in the partition.

3. Evolution of the market graph

In order to investigate the dynamics of the market graph structure, we chose the period of 1000 trading days in 1998–2002 and considered eleven 500-day shifts within this period. The starting points of every two consecutive shifts are separated by the interval of 50 days. Therefore, every pair of consecutive shifts had 450 days in common and 50 days different. Dates corresponding to each shift and the corresponding mean correlations are summarized in Table 1.

This procedure allows us to accurately reflect the structural changes of the market graph using relatively small intervals between shifts, but at the same time one can maintain sufficiently large sample sizes of the stock prices data for calculating cross-correlations for each shift. We should note that in our analysis we considered only stocks which were among those traded as of the last of the 1000 trading days, i.e., for practical reasons we did not take into account stocks which had been withdrawn from the market. However, these could be included in a more detailed analysis to obtain a better global picture of market evolution.

3.1. Global characteristics of the market graph: correlation distribution, degree distribution, edge density

The first subject of our consideration is the distribution of correlation coefficients between all pairs of stocks in the market. As it was mentioned above, this distribution on $[-1, 1]$ had a shape similar to a part of normal distribution with mean close to 0.05 for the sample data considered in [8,9]. One of the interpretations of this fact is that the correlation of most pairs of stocks is close to zero, therefore, the structure of the stock market is substantially random, and one can make a reasonable assumption that the prices of most stocks change independently. As we consider the evolution of the correlation distribution over time, it turns out that the shape of this distribution remains stable, which is illustrated by Fig. 1.

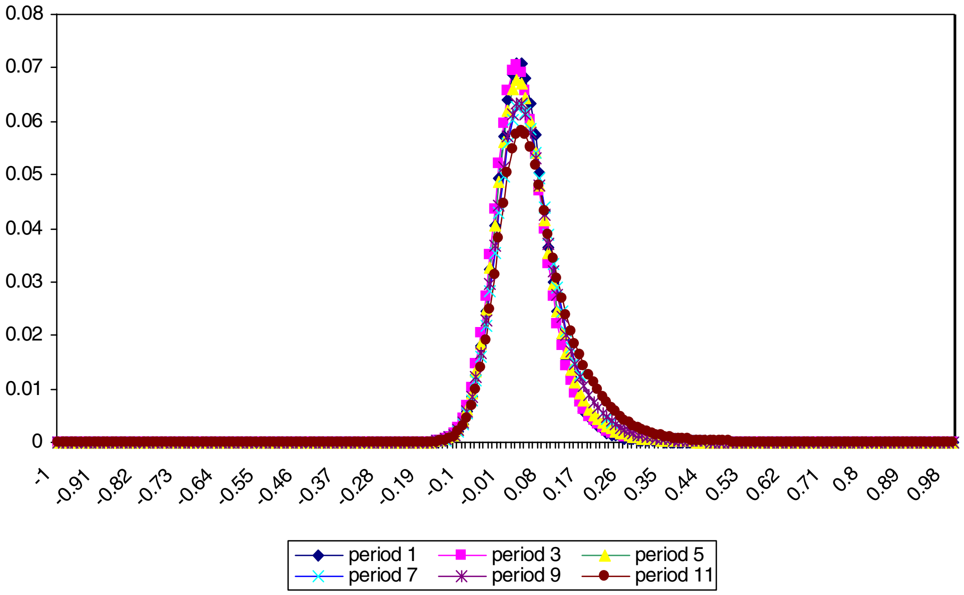

Fig. 1. Distribution of correlation coefficients in the US stock market for several overlapping 500-day periods during 2000–2002 (period 1 is the earliest, period 11 is the latest).

The stability of the correlation coefficients distribution of the market graph intuitively motivates the hypothesis that the degree distribution should also remain stable for different values of the correlation threshold. To verify this assumption, we have calculated the degree distribution of the graphs constructed for all considered time periods. The correlation threshold $\theta = 0.5$ was chosen to describe the structure of connections corresponding to significantly high correlations. Our experiments show that the degree distribution is similar for all time intervals, and in all cases it is well described by a power law. Fig. 2 shows the degree distributions (in the logarithmic scale) for some instances of the market graph (with $\theta = 0.5$) corresponding to different intervals. As one can see, all these plots can be well approximated by straight lines, which means that they represent the power-law distribution, as it follows from (3).

The cross-correlation distribution and the degree distribution of the market graph represent the general characteristics of the market, and the aforementioned results lead us to the conclusion that the global structure of the market is stable over time. However, as we will see now, some global changes in the stock market structure do take place. In order to demonstrate it, we look at another characteristic of the market graph—its edge density. The edge density of a graph is the ratio of the number of edges in this graph to the maximum possible number of edges, which is equal to $n(n-1)/2$, where n is the number of vertices in the graph.

In our analysis of the market graph, we chose a relatively high correlation threshold $\theta = 0.5$ that would ensure that we consider only the edges corresponding to the pairs of stocks, which are significantly correlated with each other. In this case, the edge density of the market graph would represent the proportion of those pairs of stocks in the market, whose price fluctuations are similar and influence each other. The subject of our interest is to study how this proportion changes during the considered period of time.

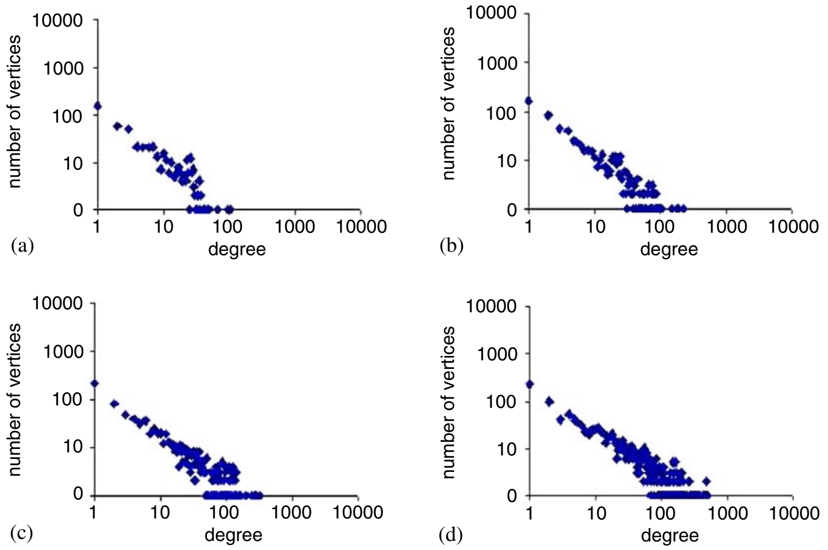

Fig. 2. Degree distribution of the market graph for different 500-day periods in 2000–2002 with $\theta = 0.5$: (a) period 1, (b) period 4, (c) period 7, and (d) period 11.

Table 2
Number of vertices and number of edges in the market graph ($\theta = 0.5$) for different periods

Period	Number of vertices	Number of edges	Edge density (%)
1	5430	2258	0.015
2	5507	2614	0.017
3	5593	3772	0.024
4	5666	5276	0.033
5	5768	6841	0.041
6	5866	7770	0.045
7	6013	10,428	0.058
8	6104	12,457	0.067
9	6262	12,911	0.066
10	6399	19,707	0.096
11	6556	27,885	0.130

Table 2 summarizes the obtained results. As it can be seen from this table, both the number of vertices and the number of edges in the market graph increase as time goes. Obviously, the number of vertices grows since new stocks appear in the market, and we do not consider those stocks which ceased to exist by the last of 1000 trading days used in our analysis, so the maximum possible number of edges in the graph increases as well. However, it turns out that the number of edges grows faster; therefore, the edge density of the market graph increases from period to period. As one can see from Fig. 3(a), the greatest increase of the edge density corresponds to the last two periods. In fact, the edge density for the latest interval is approximately 8.5 times higher than for the first interval! This dramatic jump suggests that

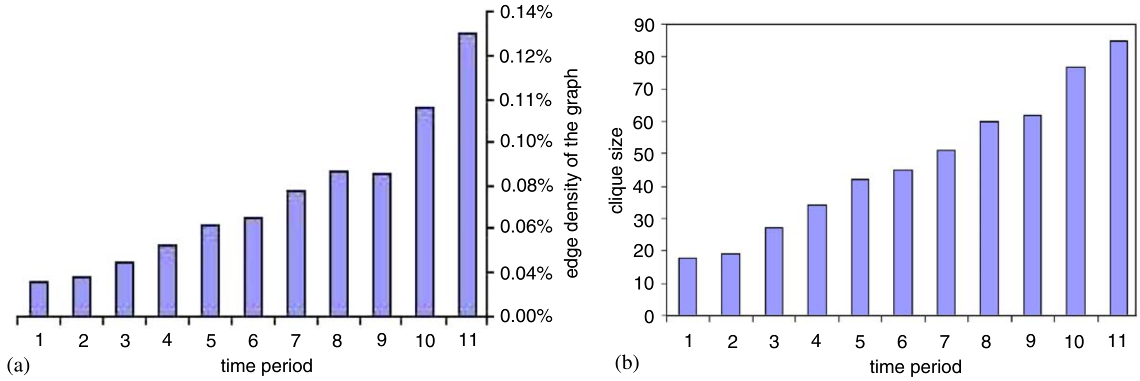

Fig. 3. Evolution of the edge density (a) and maximum clique size (b) in the market graph ($\theta = 0.5$).

there is a trend to the "globalization" of the modern stock market, which means that nowadays more and more stocks significantly affect the behavior of the others.

It should be noted that the increase of the edge density could be predicted from the analysis of the distribution of the cross-correlations between all pairs of stocks. From Fig. 1, one can observe that even though the distributions corresponding to different periods have a similar shape and the same mean, the "tail" of the distribution corresponding to the latest period (period 11) is somewhat "heavier" than for the earlier periods, which means that there are more pairs of stocks with higher values of the correlation coefficient.

3.2. Clique and independent set size dynamics in the market graph

In this subsection, we analyze the evolution of the size of the maximum clique in the market graph over the considered period of time.

Since a clique is a set of completely interconnected vertices, any stock that belongs to the clique is highly correlated with *all* other stocks in this clique; therefore, a stock is assigned to a certain group only if it demonstrates a behavior similar to *all* other stocks in this group. Clearly, the size of the maximum clique is an important characteristic of the stock market, since it represents the maximum possible group of similar objects (i.e., mutually correlated stocks).

A standard integer programming formulation [21] was used to compute the exact maximum clique in the market graph, however, before solving this problem, we applied a greedy heuristic for finding a lower bound of the clique number, and a special preprocessing technique which reduces the problem size. To find a large clique, we apply the standard "best-in" greedy algorithm based on degrees of vertices. Let C denote the clique. Starting with $C = \emptyset$, we recursively add to the clique a vertex $v_{\max}$ of largest degree and remove all vertices that are not adjacent to $v_{\max}$ from the graph. After running this algorithm, we applied the following preprocessing procedure [1]. We recursively remove from the graph all of the vertices which are not in C and whose degree is less than $|C|$, where C is the clique found by the greedy algorithm.

Denote by $G' = (V', E')$ the graph induced by remaining vertices. Then the maximum clique problem can be formulated and solved for G'. The following integer programming formulation was

Table 3
Greedy clique size and the clique number for different time periods ($\theta = 0.5$)

Period	$\|V\|$	Edge dens. in G	Clustering coefficient	$\|C\|$	$\|V'\|$	Edge dens. in G'	Clique no.
1	5430	0.00015	0.505	15	76	0.286	18
2	5507	0.00017	0.504	18	43	0.731	19
3	5593	0.00024	0.499	26	49	0.817	27
4	5666	0.00033	0.517	34	70	0.774	34
5	5768	0.00041	0.550	42	82	0.787	42
6	5866	0.00045	0.558	45	86	0.804	45
7	6013	0.00058	0.553	51	110	0.769	51
8	6104	0.00067	0.566	60	114	0.819	60
9	6262	0.00066	0.553	62	107	0.869	62
10	6399	0.00096	0.486	77	134	0.841	77
11	6556	0.00130	0.452	84	146	0.844	85

used [21]:

$$\begin{aligned} \text{maximize} \quad & \sum_{i=1}^{|V'|} x_i \\ \text{s.t.} \quad & x_i + x_j \leqslant 1, \quad (i, j) \notin E', \\ & x_i \in \{0, 1\}. \end{aligned}$$

It should be noted that in the case of market graph instances with a high positive correlation threshold, the aforementioned preprocessing procedure is very efficient and significantly reduces the number of vertices in a graph [8]. This can be intuitively explained by the fact that these instances of the market graph are *clustered* (i.e., two vertices in a graph are more likely to be connected if they have a common neighbor), so the *clustering coefficient*, which is defined as the probability that for a given vertex its two neighbors are connected by an edge, is much higher than the edge density in these graphs (see Table 3). This characteristic is also typical for other power-law graphs arising in different applications.

After reducing the size of the original graph, the resulting integer programming problem for finding a maximum clique can be relatively easily solved using the CPLEX integer programming solver [22]. It should be also noted that the aforementioned heuristic algorithm was rather efficient: in fact, in 7 out of 11 cases the greedy heuristic was able to find the exact solution of the maximum clique problem.

Table 3 presents the sizes of the maximum cliques found in the market graph for different time periods. As in the previous subsection, we used a relatively high correlation threshold $\theta = 0.5$ to consider only significantly correlated stocks. As one can see, there is a clear trend of the increase of the maximum clique size over time, which is consistent with the behavior of the edge density of the market graph discussed above (see Fig. 3(b)). This result provides another confirmation of the globalization hypothesis discussed above.

Another related issue to consider is how much the structure of maximum cliques is different for the various time periods. Table 4 presents the stocks included into the maximum cliques for different time

Table 4
Structure of maximum cliques for different time periods ($\theta = 0.5$)

Period	Stocks included into maximum clique
1	BK, EMC, FBF, HAL, HP, INTC, NCC, NOI, NOK, PDS, PMCS, QQQ, RF, SII, SLB, SPY, TER, WM
2	ADI, ALTR, AMAT, AMCC, ATML, CSCO,KLAC, LLTC, LSCC, MDY, MXIM, NVLS, PMCS, QQQ, SPY, SUNW, TXN, VTSS, XLNX
3	AMAT, AMCC, CREE, CSCO, EMC, JDSU, KLAC, LLTC, LSCC, MDY, MXIM, NVLS, PHG, PMCS, QLGC, QQQ, SEBL, SPY, STM, SUNW, TQNT, TXCC, TXN, VRTS, VTSS, XLK, XLNX
4	AMAT, AMCC, ASML, ATML, BRCM, CHKP, CIEN, CREE, CSCO, EMC, FLEX, JDSU, KLAC, LSCC, MDY, MXIM, NTAP, NVLS, PMCS, QLGC, QQQ, RFMD, SEBL, SPY, STM, SUNW, TQNT, TXCC, TXN, VRSN, VRTS, VTSS, XLK, XLNX
5	ALTR, AMAT, AMCC, ASML, ATML, BRCM, CIEN, CREE, CSCO, EMC, FLEX, IDTI, IRF, JDSU, JNPR, KLAC, LLTC, LRCX, LSCC, LSI, MDY, MXIM, NTAP, NVLS, PHG, PMCS, QLGC, QQQ, RFMD, SEBL, SPY, STM, SUNW, SWKS, TQNT, TXCC, TXN, VRSN, VRTS, VTSS, XLK, XLNX
6	ADI, ALTR, AMAT, AMCC, ASML, ATML, BEAS, BRCM, CIEN, CREE, CSCO, CY, ELX, EMC, FLEX, IDTI, ITWO, JDSU, JNPR, KLAC, LLTC, LRCX, LSCC, LSI, MDY, MXIM, NTAP, NVLS, PHG, PMCS, QLGC, QQQ, RFMD, SEBL, SPY, STM, SUNW, TQNT, TXCC, TXN, VRSN, VRTS, VTSS, XLK, XLNX
7	ALTR, AMAT, AMCC, ATML, BEAS, BRCD, BRCM, CHKP, CIEN, CNXT, CREE, CSCO, CY, DIGL, EMC, FLEX, HHH, ITWO, JDSU, JNPR, KLAC, LLTC, LRCX, LSCC, MDY, MERQ, MXIM, NEWP, NTAP, NVLS, ORCL, PMCS, QLGC, QQQ, RBAK, RFMD, SCMR, SEBL, SPY, SSTI, STM, SUNW, SWKS, TQNT, TXCC, TXN, VRSN, VRTS, VTSS, XLK, XLNX
8	ALTR, AMAT, AMCC, AMKR, ARMHY, ASML, ATML, AVNX, BEAS, BRCD, BRCM, CHKP, CIEN, CMRC, CNXT, CREE, CSCO, CY, DIGL, ELX, EMC, EXTR, FLEX, HHH, IDTI, ITWO, JDSU, JNPR, KLAC, LLTC, LRCX, LSCC, MDY, MERQ, MRVC, MXIM, NEWP, NTAP, NVLS, ORCL, PMCS, QLGC, QQQ, RFMD, SCMR, SEBL, SNDK, SPY, SSTI, STM, SUNW, SWKS, TQNT, TXCC, TXN, VRSN, VRTS, VTSS, XLK, XLNX
9	ADI, ALTR, AMAT, AMCC, ARMHY, ASML, ATML, AVNX, BDH, BEAS, BHH, BRCM, CHKP, CIEN, CLS, CREE, CSCO, CY, DELL, ELX, EMC, EXTR, FLEX, HHH, IAH, IDTI, IIH, INTC, IRF, JDSU, JNPR, KLAC, LLTC, LRCX, LSCC,LSI, MDY, MXIM, NEWP, NTAP, NVLS, PHG, PMCS, QLGC, QQQ, RFMD, SCMR, SEBL, SNDK, SPY, SSTI, STM, SUNW, SWKS, TQNT, TXCC, TXN, VRSN, VRTS,

Table 4 (*continued*)

Period	Stocks included into maximum clique
	VTSS, XLK, XLNX
10	ADI, ALTR, AMAT, AMCC, AMD, ASML, ATML, BDH, BHH, BRCM, CIEN, CLS, CREE, CSCO, CY, CYMI, DELL, EMC, FCS, FLEX, HHH, IAH, IDTI, IFX, IIH, IJH, IJR, INTC, IRF, IVV, IVW, IWB, IWF, IWM, IWV, IYV, IYW, IYY, JBL, JDSU, KLAC, KOPN, LLTC, LRCX, LSCC, LSI, LTXX, MCHP, MDY, MXIM, NEWP, NTAP, NVDA, NVLS, PHG, PMCS, QLGC, QQQ, RFMD, SANM, SEBL, SMH, SMTC, SNDK, SPY, SSTI, STM, SUNW, TER, TQNT, TXCC, TXN, VRTS, VSH, VTSS, XLK, XLNX
11	ADI, ALA, ALTR, AMAT, AMCC, AMD, ASML, ATML, BDH, BEAS, BHH, BRCM, CIEN, CLS, CNXT, CREE, CSCO, CY, CYMI, DELL, EMC, EXTR, FCS, FLEX, HHH, IAH, IDTI, IIH, IJH, IJR, INTC, IRF, IVV, IVW, IWB, IWF, IWM, IWO, IWV, IWZ, IYV, IYW, IYY, JBL, JDSU, JNPR, KLAC, KOPN, LLTC, LRCX, LSCC, LSI, LTXX, MCRL, MDY, MKH, MRVC, MXIM, NEWP, NTAP, NVDA, NVLS, PHG, PMCS, QLGC, QQQ, RFMD, SANM, SEBL, SMH, SMTC, SNDK, SPY, SSTI, STM, SUNW, TER, TQNT, TXN, VRTS, VSH, VTSS, XLK, XLNX

periods. It turns out that in most cases stocks that appear in a clique in an earlier period also appear in the cliques in later periods.

There are some other interesting observations about the structure of the maximum cliques found for different time periods. It can be seen that all the cliques include a significant number of stocks of the companies representing the "high-tech" industry sector. As the examples, one can mention well-known companies such as Sun Microsystems, Inc., Cisco Systems, Inc., Intel Corporation, etc. Moreover, each clique contains stocks of the companies related to the semiconductor industry (e.g., Cypress Semiconductor Corporation, Cree, Inc., Lattice Semiconductor Corporation, etc.), and the number of these stocks in the cliques increases with the time. These facts suggest that the corresponding branches of industry expanded during the considered period of time to form a major cluster of the market.

In addition, we observed that in the later periods (especially in the last two periods) the maximum cliques contain a rather large number of exchange traded funds, i.e., stocks that reflect the behavior of certain indices representing various groups of companies. It should be mentioned that all maximum cliques contain Nasdaq 100 tracking stock (QQQ), which was also found to be the vertex with the highest degree (i.e., correlated with the most stocks) in the market graph [8].

Another natural question that one can pose is how the size of independent sets (i.e., diversified portfolios in the market) changes over time. As it was pointed out in [8,9], finding a maximum independent set in the market graph turns out to be a much more complicated task than finding a maximum clique. In particular, in the case of solving the maximum independent set problem (or, equivalently, the maximum clique problem in the complementary graph), the preprocessing procedure described above does not reduce the size of the original graph. This can be explained by the fact that the clustering coefficient in the complementary market graph with $\theta = 0$ is much smaller than in the original graph corresponding to $\theta = 0.5$ (see Table 5).

Table 5
Size of independent sets found using the greedy heuristic ($\theta = 0.0$)

Period	Number of vertices	Edge density	Clustering coefficient	Independent set size
1	5430	0.258	0.293	11
2	5507	0.275	0.307	11
3	5593	0.281	0.307	10
4	5666	0.265	0.297	11
5	5768	0.260	0.292	11
6	5866	0.254	0.288	11
7	6013	0.228	0.269	11
8	6104	0.227	0.268	10
9	6262	0.238	0.277	12
10	6399	0.228	0.269	12
11	6556	0.201	0.245	11

Edge density and clustering coefficient are given for the complement graph.

Similar to [8], we calculate maximal independent sets (a maximal independent set is an independent set that is not a subset of another independent set) in the market graph using the above greedy algorithm. As one can see from Table 5, the sizes of independent sets found in the market graph for $\theta = 0$ are rather small, which is consistent with the results of [8].

The choice of $\theta = 0$ is rather too extreme for practical purposes, however an independent set found for this value of θ could serve as a "base" for forming a diversified portfolio. For example, an independent set could be extended by adding extra vertices selected in such a way that the resulting set of vertices is "almost" independent. One possibility is, starting with an independent set I, recursively add a vertex to I that has no more than k neighbors in I, where k is some small positive integer. The resulting set would correspond to a $(k+1)$-*plex* in the complement graph. A subset of vertices C of cardinality m is called a k-plex if it has at least $m - k$ neighbors in C. Note that if $k = 1$ we obtain the definition of clique. The concept of k-plex is quite popular in the study of social networks, where a k-plex corresponds to a so-called cohesive subgroup [23]. Another idea along these lines is to use *quasi-independent sets* to represent diversified portfolios. Similarly to a quasi-clique, a γ-quasi-independent set is defined as a subset of vertices whose induced subgraph has at most $(1-\gamma)q(q-1)/2$ edges.

3.3. *Minimum clique partition of the market graph*

Besides analyzing the maximum cliques in the market graph, one can also divide the market graph into the smallest possible set of distinct cliques. As it was pointed out above, the partition of a dataset into sets (clusters) of elements grouped according to a certain criterion is referred to as *clustering*.

Clearly, the methodology of finding cliques in the market graph provides an effective tool of performing clustering based on the stock market data. The choice of the grouping criterion is clear and natural: "similar" financial instruments are determined according to the correlation between their price fluctuations. Moreover, the minimum number of clusters in the partition of the set of financial instruments is equal to the minimum number of distinct cliques that the market graph can be divided into (the minimum clique partition problem).

Table 6
The largest clique size and the number of cliques in computed clique partitions ($\theta = 0.05$)

Period	Number of vertices	Edge density	Largest clique in the partition	# of cliques in the partition
1	5430	0.400	469	494
2	5507	0.377	552	517
3	5593	0.379	636	513
4	5666	0.405	743	503
5	5768	0.413	789	501
6	5866	0.425	824	496
7	6013	0.469	929	471
8	6104	0.475	983	470
9	6262	0.456	997	509
10	6399	0.474	1159	501
11	6556	0.521	1372	479

For finding a clique partition, we choose the instance of the market graph with a low correlation threshold $\theta = 0.05$ (the mean of the correlation coefficients distribution shown in Fig. 1), which would ensure that the edge density of the considered graph is high enough and the number of isolated vertices (which would obviously form distinct cliques) is small.

We use the standard greedy heuristic to compute a clique partition in the market graph: recursively find a maximal clique and remove it from the graph, until no vertex remain. Cliques are computed using the previously described greedy algorithm. The corresponding results for the market graph with threshold $\theta = 0.05$ are presented in Table 6. Note that the size of the largest clique in the partition is increasing from one period to another, with the largest clique in the last period containing about three times as many vertices as the corresponding clique in the first partition. At the same time, the number of cliques in the partition is comparable for different periods, with a slight overall trend towards decrease, whereas the number of vertices is increasing as time goes.

4. Concluding remarks

Graph representation of the stock market data and interpretation of the properties of this graph gives a new insight into the internal structure of the stock market. In this paper, we have studied different characteristics of the market graph and their evolution over time and came to several interesting conclusions based on our analysis. It turns out that the power-law structure of the market graph is quite stable over the considered time intervals; therefore one can say that the concept of self-organized networks, which was mentioned above, is applicable in finance, and in this sense the stock market can be considered as a "self-organized" system.

Another important result is the fact that the edge density of the market graph, as well as the maximum clique size, steadily increase during the last several years, which supports the well-known idea about the globalization of economy which has been widely discussed recently.

We have also indicated the natural way of dividing the set of financial instruments into groups of similar objects (clustering) by computing a clique partition of the market graph. This methodology can be extended by considering quasi-cliques in the partition, which may reduce the number of obtained clusters.

Acknowledgements

The authors would like to thank Sigurdur Ólafsson for his encouragement and the two anonymous referees for their constructive comments.

References

[1] Abello J, Pardalos PM, Resende MGC. On maximum clique problems in very large graphs. DIMACS Series, vol. 50. Providence, RI: American Mathematical Society; 1999. p. 119–30.
[2] Aiello W, Chung F, Lu L. A random graph model for power-law graphs. Experimental Mathematics 2001;10:53–66.
[3] Hayes B. Graph theory in practice. American Scientist 2000;88:9–13 (Part I), 104–9 (Part II).
[4] Jeong H, Tomber B, Albert R, Oltvai ZN, Barabasi A-L. The large-scale organization of metabolic networks. Nature 2000;407:651–4.
[5] Watts D. Small worlds: the dynamics of networks between order and randomness. Princeton, NJ: Princeton University Press; 1999.
[6] Watts D, Strogatz S. Collective dynamics of 'small-world' networks. Nature 1998;393:440–2.
[7] Mantegna RN, Stanley HE. An introduction to econophysics: correlations and complexity in finance. Cambridge: Cambridge University Press; 2000.
[8] Boginski V, Butenko S, Pardalos PM. On structural properties of the market graph. In: Nagurney A, editor. Innovations in financial and economic networks. Edward Elgar Publishers; 2003.
[9] Boginski V, Butenko S, Pardalos PM. Statistical analysis of financial networks. Computational Statistics and Data Analysis 2005;48(2):431–43.
[10] Albert R, Barabasi A-L. Statistical mechanics of complex networks. Reviews of Modern Physics 2002;74:47–97.
[11] Barabasi A-L, Albert R. Emergence of scaling in random networks. Science 1999;286:509–11.
[12] Barabasi A-L. Linked. Perseus Publishing; 2002.
[13] Boginski V, Butenko S, Pardalos PM. Modeling and optimization in massive graphs. In: Pardalos PM, Wolkowicz H, editors. Novel approaches to hard discrete optimization. Providence, RI: American Mathematical Society; 2003. p. 17–39.
[14] Broder A, Kumar R, Maghoul F, Raghavan P, Rajagopalan S, Stata R, Tomkins A, Wiener J. Graph structure in the Web. Computer Networks 2000;33:309–20.
[15] Faloutsos M, Faloutsos P, Faloutsos C, On power-law relationships of the Internet topology. ACM SICOMM, 1999.
[16] Garey MR, Johnson DS. Computers and intractability: a guide to the theory of NP-completeness. New York: Freeman; 1979.
[17] Arora S, Safra S. Approximating clique is NP-complete. Proceedings of the 33rd IEEE symposium on foundations on computer science 1992. p. 2–13.
[18] Håstad J. Clique is hard to approximate within $n^{1-\varepsilon}$. Acta Mathematica 1999;182:105–42.
[19] Pardalos PM, Mavridou T, Xue J. The graph coloring problem: a bibliographic survey. In: Du D-Z, Pardalos PM, editors. Handbook of combinatorial optimization, vol. 2. Dordrecht: Kluwer Academic Publishers; 1998. p. 331–95.
[20] Bradley PS, Fayyad UM, Mangasarian OL. Mathematical programming for data mining: formulations and challenges. INFORMS Journal on Computing 1999;11(3):217–38.
[21] Bomze IM, Budinich M, Pardalos PM, Pelillo M. The maximum clique problem. In: Du D-Z, Pardalos PM, editors. Handbook of combinatorial optimization. Dordrecht: Kluwer Academic Publishers; 1999. p. 1–74.
[22] ILOG CPLEX 7.0 Reference Manual, 2000.
[23] Wasserman S, Faust K. Social network analysis: methods and applications. Cambridge: Cambridge University Press; 1994.