

Contribution for  
Innovations in Financial and Economic  
Networks  
Anna Nagurney, editor

Vladimir Boginski, Sergiy Butenko and Panos M. Pardalos

July 26, 2003

# 1 On Structural Properties of the Market Graph

Vladimir Boginski, Sergiy Butenko and  
Panos M. Pardalos

Industrial and Systems Engineering Department  
University of Florida  
303 Weil Hall, Gainesville, FL 32611  
{vb, butenko, pardalos}@ufl.edu

## 1.1 Introduction

In the modern world of information, one often encounters serious computational challenges of dealing with massive data sets that arise in a great variety of scientific, engineering and commercial applications, such as government and military systems, telecommunications, medicine and biotechnology, astrophysics, ecology, geographical information systems, and finance (cf. Boginski et al. (2003), Abello et al. (2002)). The pervasiveness and complexity of the problems brought by massive data sets make it one of the most challenging and exciting areas of research for years to come.

In many practically important cases, a massive data set can be represented as a very large graph with certain attributes associated with its vertices and edges. These attributes may contain specific information characterizing the given application. Studying the structure of this graph is essential for understanding the structural properties of the application it represents, as well as for improving storage organization and information retrieval.

One of the examples of representing a massive data set as a graph is the *Call graph* arising in the telecommunications traffic data. In this graph, the vertices are telephone numbers, and two vertices are connected by an edge

if a call was made from one number to another within a certain period of time. This graph was studied in detail by Abello et al. (1999) and by Aiello et al. (2001). Interestingly enough, this graph demonstrated the properties that are described by the *power-law model* (Boginski et al., 2003).

Moreover, it turns out that the Web graph (where the vertices are websites, and the arcs are links between them) and the Internet graph (where the vertices are routers and the edges are cables in the physical network) also obey the power-law model (Broder et al. (2000), Faloutsos et al. (1999)).

In this paper, we will concentrate on another important real-life application - the graph representation of the stock market. Every year stock markets generate a huge amount of data, and this information can be used for constructing a graph that will reflect the market behavior.

Although not so obviously as in the examples of the Call graph and the Web graph, the stock market can be represented as a graph. A natural graph representation of the stock market is based on the cross correlations of price fluctuations. A *market graph* can be constructed as follows: each financial instrument is represented by a vertex, and two vertices are connected by an edge if the correlation coefficient of the corresponding pair of instruments (calculated for a certain period of time) exceeds a specified threshold  $\theta$ ,  $-1 \leq \theta \leq 1$ .

Nowadays, a great number of different instruments are traded in the US stock market, so the market graph representing them is very large. The market graph that we construct has 6546 vertices and several million edges.

In this paper, we present a detailed study of the properties of this graph. It turns out that the market graph can be rather accurately described by the power-law model. We analyze the distribution of the degrees of the vertices in this graph, the edge density of this graph with respect to the correlation threshold, as well as its connectivity and the size of its connected components.

Furthermore, we look for *maximum cliques* and *maximum independent sets* in this graph for different values of the correlation threshold. Analyzing cliques and independent sets in the market graph gives us a very valuable knowledge about the internal structure of the stock market. For instance, a clique in this graph will represent a set of instruments whose prices change similarly over time (a change of the price of any instrument in a clique is likely to affect all other instruments in this clique), and an independent set will consist of instruments that have negative correlations with respect to each other, therefore, they can be treated as a so-called diversified portfolio. Based on the information obtained from this analysis, we will be able to classify financial instruments into different groups, which will give us a deeper insight into the stock market structure.

## 1.2 Notations and Definitions

To facilitate the further discussion, let us introduce some formal definitions and notations. We will denote by  $G = (V, E)$  a simple undirected graph with the set of  $n$  vertices  $V$  and the set of edges  $E$ .

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  $G$ .

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  $\gamma$ -*clique*  $C_\gamma$ , also called a *quasi-clique*, is a subset of  $V$  such that  $G(C_\gamma)$  has at least  $\lfloor \gamma q(q-1)/2 \rfloor$  edges, where  $q$  is the cardinality of  $C_\gamma$ .

The maximum clique problem is known to be NP-hard (see Garey and Johnson (1979)). Moreover, it turns out that even the approximation of the maximum clique is NP-hard. For instance, Arora and Safra (1992) proved that for some positive  $\epsilon$  the approximation of the maximum clique within a factor of  $n^\epsilon$  is NP-hard. Recently, Håstad (1999) has shown that in fact for any  $\delta > 0$  the maximum clique is hard to approximate in polynomial time within a factor  $n^{1-\delta}$ .

These facts together with practical evidence (see Johnson and Trick (1996)) demonstrate that the maximum clique problem is hard to solve even in graphs of a moderate size. The theoretical complexity and the huge sizes of data make this problem especially hard in large graphs.

Another concept closely related to a clique is an *independent set* which 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 in this sense these problems are equivalent, and the complexity results obtained for the maximum clique problem are also valid for the maximum independent set problem.

## 1.3 Theoretical Models of Real-Life Graphs

In many cases, especially if the size of real-life graphs is large enough, it is useful to construct a proper theoretical model of these graphs. This may help to avoid computational difficulties and to better understand the structure of a real-life graph.

The first attempt to build such a model was based on the concept of so-

called *uniform random graphs*. The classical theory of uniform random graphs founded by Erdős and Rényi (1959, 1960, 1961) deals with several standard models. Two of such models are  $G(n, m)$  and  $\mathcal{G}(n, p)$  (Bollobás (1978, 1985)). The first model assigns the same probability to all graphs with  $n$  vertices and  $m$  edges, while in the second model each pair of vertices is chosen to be linked by an edge randomly and independently with probability  $p$ .

In most cases for each natural  $n$  a probability space consisting of graphs with exactly  $n$  vertices is considered, and the properties of this space as  $n \rightarrow \infty$  are studied. It is said that a typical element of the space or *almost every (a.e.)* graph has property  $Q$  when the probability that a random graph on  $n$  vertices has this property tends to 1 as  $n \rightarrow \infty$ . They also say that the property  $Q$  holds *asymptotically almost surely (a.a.s.)*.

The asymptotical properties of uniform random graphs have been well studied. Among the results obtained in this field we should mention the connectivity threshold and the emergence of a giant connected component. More specifically, a uniform random graph  $\mathcal{G}(n, p)$  is a.a.s. connected if  $p > \frac{\log n}{n}$  and it is a.a.s. disconnected otherwise. Also, if  $p$  is in the range  $\frac{1}{n} < p < \frac{\log n}{n}$ , the graph  $\mathcal{G}(n, p)$  a.a.s. has a unique *giant connected component*.

Although the uniform random graph model captured some properties of the massive graphs arising in practical applications (such as the emergence of a giant connected component), empirical measurements showed a significant difference in the structures of a real graph and a corresponding uniform random graph with the same number of vertices and edges. For instance, this model cannot handle the property of *clustering* that often takes place in the real graphs (Watts and Strogatz (1998), Watts (1999)). This property means that the probability of the event that two given vertices are connected by an edge is higher if these vertices have a common neighbor (i.e. a vertex which is connected by an edge with both of these vertices). The probability that for a given vertex its two neighbors are connected by an edge is called the *clustering coefficient*. It can be easily seen that in the case of the model  $\mathcal{G}(n, p)$  the clustering coefficient is equal to the parameter  $p$ , since the probability that each pair of vertices is connected by an edge is independent of all other vertices. In real-life graphs, the value of the clustering coefficient turns out to be much higher than the value of the parameter  $p$  of the uniform random graphs with the same number of vertices and edges.

However, the most important drawback of the uniform random graph model is the difference in the vertices degree distribution, compared to the real massive graphs. It can be easily shown that as the number of vertices in a uniform random graph increases, the distribution of the degrees of the vertices tends to the well-known *Poisson distribution* with the parameter  $np$  which represents the average degree of a vertex. But the empirical experiments show that in many real-life graphs the degree distribution obeys a *power law*. That is why another theoretical model was developed for describing real-life graphs - the *power-law random graph model*.

The basic idea of the power-law random graph model  $P(\alpha, \beta)$  is as follows. If  $y$  is the number of nodes with degree  $x$ , then according to this model

$$y = e^\alpha / x^\beta. \quad (1.1)$$

Equivalently, we can write

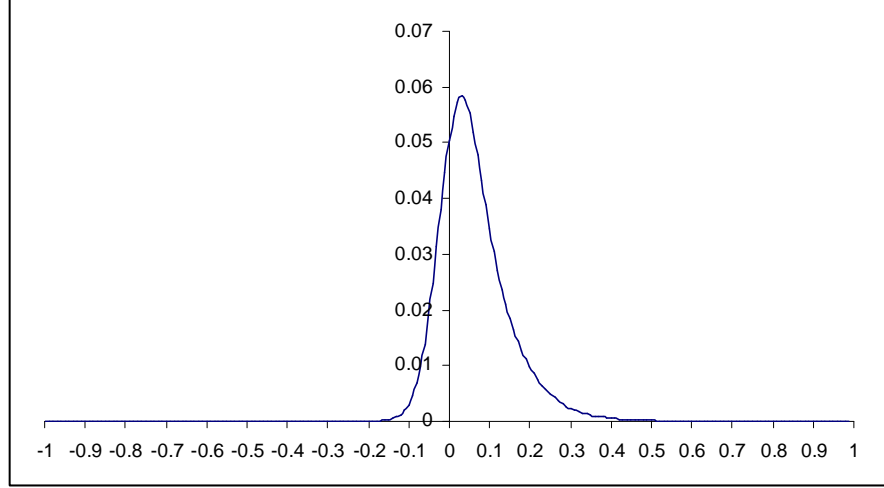
$$\log y = \alpha - \beta \log x. \quad (1.2)$$

This representation is more convenient in the sense that the relationship between  $y$  and  $x$  can be plotted as a straight line on a log-log scale, so that  $(-\beta)$  is the slope, and  $\alpha$  is the intercept.

An important aspect related to the power-law model is the relationship between the parameter  $\beta$  and the size of the largest connected component. It has been theoretically shown that in a power-law graph a giant connected component *a.a.s.* emerges at  $\beta \simeq 3.47875$ , and a graph *a.a.s.* becomes connected when  $\beta < 1$ . More specifically, the following results are valid (Aiello et al, 2001):

- If  $0 < \beta < 1$ , then a power-law graph is *a.a.s.* connected (i.e. there is only one connected component of size  $n$ ).
- If  $1 \leq \beta < 2$ , then a power-law graph *a.a.s.* has a giant connected component (the component size is  $\Theta(n)$ ), and the second largest connected component *a.a.s.* has a size  $\Theta(1)$ .
- If  $2 < \beta < \beta_0 = 3.47875$ , then a giant connected component *a.a.s.* exists, and the size of the second largest component *a.a.s.* is  $\Theta(\log n)$ .
- $\beta = 2$  is a special case when there is *a.a.s.* a giant connected component, and the size of the second largest connected component is  $\Theta(\log n / \log \log n)$ .
- If  $\beta > \beta_0 = 3.47875$ , then there is *a.a.s.* no giant connected component.

As it was pointed out above, the power-law model proved to be suitable for the Call graph and the Web graph. Since these graphs are very large, the asymptotical theoretical results obtained for this model are applicable to these graphs. Although the market graph is significantly smaller than the Web and the Call graphs, and one cannot expect the same level of agreement between theoretical and practical results, this graph is continuously growing, since a lot of new instruments appear in the stock market. Does the market graph obey the power-law model? In the next section we present a detailed analysis of the market graph which will give an answer to this question.



**Fig. 1.1.** Distribution of correlation coefficients in the stock market

## 1.4 Structure of the Market Graph

### 1.4.1 Constructing the market graph

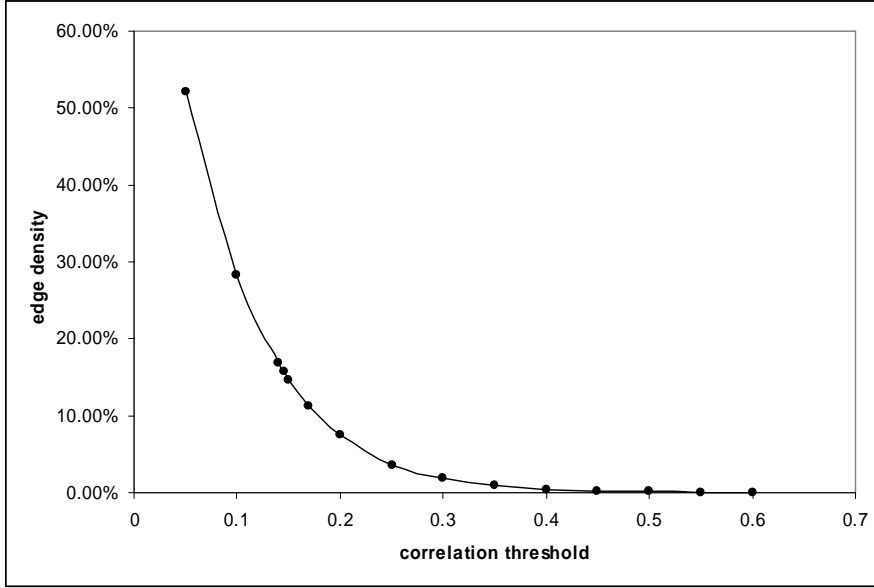
The market graph that we study in this paper represents the set of financial instruments traded in the US stock markets. More specifically, we consider 6546 instruments and analyze daily changes of their prices over a period of 500 consecutive trading days in 2000-2002. Based on this information, we calculate the cross-correlations between each pair of stocks using the following formula (Mantegna and Stanley, 2000):

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

where  $R_i(t) = \ln \frac{P_i(t)}{P_i(t-1)}$  defines the return of the stock  $i$  for day  $t$ .  $P_i(t)$  denotes the price of the stock  $i$  on day  $t$ .

The correlation coefficients  $C_{ij}$  can vary from -1 to 1. Figure 1.1 shows the distribution of the correlation coefficients based on the prices data for the years 2000-2002. It can be seen that this plot has a shape similar to the normal distribution with the mean 0.05.

The main idea of constructing a market graph is as follows. Let the set of financial instruments represent the set of vertices of the graph. Also, we specify a certain threshold value  $\theta$ ,  $-1 \leq \theta \leq 1$  and add an undirected edge connecting the vertices  $i$  and  $j$  if the corresponding correlation coefficient  $C_{ij}$  is greater than or equal to  $\theta$ . Obviously, different values of  $\theta$  define the market graphs with the same set of vertices, but different sets of edges.



**Fig. 1.2.** Edge density of the market graph for different values of the correlation threshold.

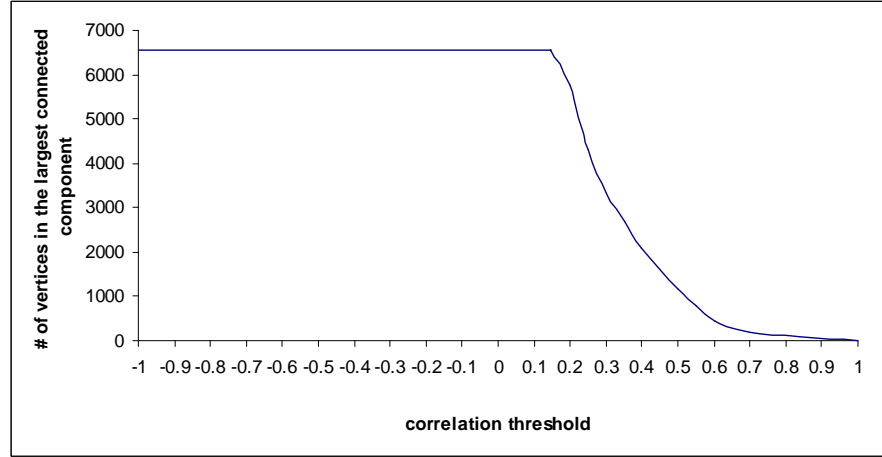
It is easy to see that the number of edges in the market graph decreases as the threshold value  $\theta$  increases. In fact, our experiments show that the edge density of the market graph decreases exponentially w.r.t.  $\theta$ . The corresponding graph is presented on Figure 1.2.

#### 1.4.2 Connectivity of the market graph

In Section 1.3 we mentioned the connectivity thresholds in random graphs. The main idea of this concept is finding a threshold value of the parameter of the model ( $p$  in the case of uniform random graphs, and  $\beta$  in the case of power-law graphs) that will define if the graph is connected or not. Moreover, if the graph is disconnected, another threshold value can be defined to determine if the graph has a giant connected component or all of its connected components have a small size.

For instance, in the case of the power-law model  $\beta = 1$  is a threshold value that determines the connectivity of the power-law graph, i.e. the graph is a.a.s. connected if  $\beta < 1$ , and it is a.a.s. disconnected otherwise. Similarly,  $\beta \approx 3.47875$  defines the existence of a giant connected component in the power-law graph.

Now a natural question arises: what is the connectivity threshold for the market graph? Since the number of edges in the market graph depends on the chosen correlation threshold  $\theta$ , we should find a value  $\theta_0$  that determines



**Fig. 1.3.** Plot of the size of the largest connected component in the market graph as a function of correlation threshold  $\theta$ .

the connectivity of the graph. As it was mentioned above, the smaller value of  $\theta$  we choose, the more edges the market graph will have. So, if we decrease  $\theta$ , after a certain point, the graph will become connected. We have conducted a series of computational experiments for checking the connectivity of the market graph using the breadth-first search technique, and we obtained a relatively accurate approximation of the connectivity threshold:  $\theta_0 \simeq 0.14382$ . Moreover, we investigated the dependency of the size of the largest connected component in the market graph w.r.t.  $\theta$ . The corresponding plot is shown on Figure 1.3.

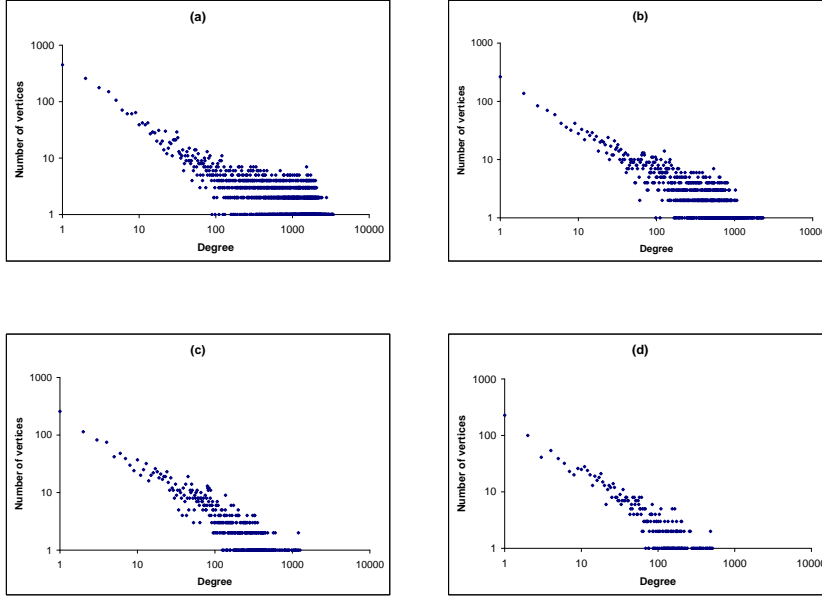
### 1.4.3 Degree distributions in the market graph

As it was shown in the previous section, the power-law model fairly well describes some of the real-life massive graphs, such as the Web graph and the Call graph. In this section, we will show that the market graph also obeys the power-law model.

It should be noted that since we consider a set of market graphs, where each graph corresponds to a certain value of  $\theta$ , the degree distributions will be different for each  $\theta$ .

The results of our experiments turned out to be rather interesting.

If we specify a small value of the correlation threshold  $\theta$ , such as  $\theta = 0$ ,  $\theta = 0.05$ ,  $\theta = 0.1$ ,  $\theta = 0.15$ , the distribution of the degrees of the vertices is very “noisy” and does not have any well-defined structure. Note that for these values of  $\theta$  the market graph is connected and has a high edge density. The market graph structure seems to be very difficult to analyze in these cases.

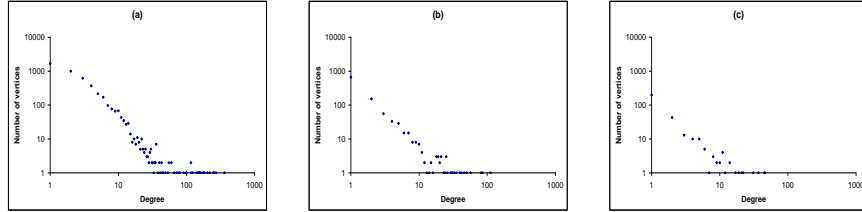


**Fig. 1.4.** Degree distribution of the market graph for (a)  $\theta = 0.2$ ; (b)  $\theta = 0.3$ ; (c)  $\theta = 0.4$ ; (d)  $\theta = 0.5$

However, the situation changes drastically if a higher correlation threshold is chosen. As the edge density of the graph decreases, the degree distribution more and more resembles a power law. In fact, for  $\theta \geq 0.2$  this distribution is approximately a straight line in the log-log scale, which is exactly the power law distribution, as it was shown in Section 1.3. Figure 1.4 demonstrates the degree distributions of the market graphs for some values of the correlation threshold.

An interesting observation is that the slope of the lines (which is equal to the parameter  $\beta$  of the power-law model) is rather small. It can be seen from Formula 1.1 that in this case the graph will contain many vertices with a high degree. This fact is important for the next subject of our interest - finding maximum cliques in the market graph. Intuitively, one can expect a large clique in a graph with a small value of the parameter  $\beta$ . As we will see in the next section, this assumption is true for the market graph.

Another combinatorial optimization problem associated with the market graph is finding maximum independent sets in the graphs with a negative correlation threshold  $\theta$ . Clearly, instruments in an independent set will be negatively correlated with each other, and therefore form a diversified portfolio.



**Fig. 1.5.** Degree distribution of the complementary market graph for (a)  $\theta = -0.15$ ; (b)  $\theta = -0.2$ ; (c)  $\theta = -0.25$

However, we can consider a *complementary graph* for a market graph with a negative value of  $\theta$ . In this graph, an edge will connect instruments  $i$  and  $j$  if the correlation between them  $C_{ij} < \theta$ . Recall that a maximum independent set in the initial graph is a maximum clique in the complementary graph, so the maximum independent set problem can be reduced to the maximum clique problem in the complementary graph.

Therefore, it is also useful to investigate the degree distributions of these complementary graphs. As it can be seen from Figure 1.1, the distribution of the correlation coefficients is almost symmetric around  $\theta = 0.05$ , so for the values of  $\theta$  close to 0 the edge density of both the initial and the complementary graph is high enough. So, for these values of  $\theta$  the degree distribution of a complementary graph is also “noisy” as in the case of the corresponding initial graph.

As  $\theta$  decreases (i.e. increases in the absolute value), the degree distribution of a complementary graph tends to the power law. The corresponding graphs are shown on Figure 1.5. However, in this case, the slope of the line in the log-log scale (the value of the parameter  $\beta$ ) is higher than in the case of positive values of  $\theta$ . It means that there are not many vertices with a high degree in these graphs, so the size of a maximum clique should be significantly smaller than in the case of the market graphs with a positive correlation threshold.

It is also interesting to compare the difference in clustering coefficients of a market graph (with positive values of  $\theta$ ) or its complement (with negative values of  $\theta$ ) (see Table 1.1).

Intuitively, the large clustering coefficients should correspond to graphs with larger cliques, therefore, from Table 1.1 one should expect that the cliques in market graph with positive  $\theta$  are much larger than the independent sets in market graph with negative  $\theta$ . This prediction will be confirmed in the next section, where we present the computational results of solving the maximum clique and maximum independent set problems in the market graph.

**Table 1.1.** Clustering coefficients of the market graph (\* - complementary graph)

$\theta$	clustering coef.
-0.15*	$2.64 \times 10^{-5}$
-0.1*	0.0012
0.3	0.4885
0.4	0.4458
0.5	0.4522
0.6	0.4872
0.7	0.4886

#### 1.4.4 Cliques and independent sets in the market graph

As it was mentioned above, the maximum clique and the maximum independent set problems are NP-hard. It makes these problems especially challenging in large graphs. The maximum clique problem admits an integer programming formulation, however, in the case of the graph with 6546 vertices this integer programming problem cannot be solved in a reasonable time. Therefore, we used a greedy heuristic for finding a lower bound of the clique number, and a special preprocessing technique which reduces a problem size.

To find a large clique, we apply the following greedy algorithm. Starting with an empty set, we recursively add to the clique a vertex from the neighborhood of the clique adjacent to the most vertices in the neighborhood of the clique. If we denote by  $N(i) = \{j | (i, j) \in E\}$  the set of neighbors of  $i$  in  $G = (V, E)$ , then the neighborhood of a clique  $C$  is  $\bigcap_{i \in C} N(i)$ , and we obtain the following algorithm:

$$C = \emptyset, G_0 = G;$$

**do**

$$G_0 = \bigcap_{i \in C} N(i) \setminus C;$$

$$C = C \cup j, \text{ where } j \text{ is a vertex of largest degree in } G_0;$$

**until**  $G_0 = \emptyset$ .

After running this algorithm, we applied the following preprocessing procedure (Abello et al., 1999). 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 above algorithm. This simple procedure enabled us to significantly reduce the size of the maximum clique search space. Let us denote by  $G'(V', E')$  the graph induced by remaining vertices. Table 1.2

presents the sizes of the cliques found using the greedy algorithm, and sizes of the graphs remaining after applying the preprocessing procedure.

**Table 1.2.** Sizes of cliques found using the greedy algorithm and sizes of graphs remaining after applying the preprocessing technique

$\theta$	edge density	clique size	$ V' $	edge dens. in $G'$
0.35	0.0090	168	535	0.6494
0.4	0.0047	104	405	0.6142
0.45	0.0024	109	213	0.8162
0.5	0.0013	84	146	0.8436
0.55	0.0007	61	102	0.8701
0.6	0.0004	45	70	0.8758
0.65	0.0002	23	80	0.5231
0.7	0.0001	21	33	0.7557

In order to find the maximum clique of  $G'$  (which is also the maximum clique in the original graph  $G$ ), we used the following integer programming formulation of the maximum clique problem (Bomze et al., 1999):

$$\begin{aligned} & \text{maximize} && \sum_{i=1}^n x_i \\ & \text{s.t.} && x_i + x_j \leq 1, (i, j) \notin E'; \\ & && x_i \in \{0, 1\}, i = 1, \dots, n. \end{aligned}$$

We used CPLEX to solve this integer program for some of the considered instances.

Table 1.3 summarizes the sizes of the maximum cliques found in the graph for different values of  $\theta$ . It turns out that these cliques are rather large. In fact, even for  $\theta = 0.6$ , which is a very high correlation threshold, the clique of size 45 was found.

These results are in agreement with the discussion in the previous section, where we analyzed the degree distributions of the market graphs with positive values of  $\theta$  and came to the conclusion that the cliques in these graphs should be large.

The financial interpretation of the clique in the market graph is that it defines the set of stocks whose price fluctuations exhibit a similar behavior. Our results show that in the modern stock market there are large groups of instruments that are correlated with each other.

Next, we consider the maximum independent set problem in the market graphs with nonpositive values of the correlation threshold  $\theta$ . As described in the previous section, this problem can be easily represented as a maximum

**Table 1.3.** Sizes of the maximum cliques in the market graph with different values of the correlation threshold

$\theta$	edge density	clique size
0.35	0.0090	193
0.4	0.0047	144
0.45	0.0024	109
0.5	0.0013	85
0.55	0.0007	63
0.6	0.0004	45
0.65	0.0002	27
0.7	0.0001	22

**Table 1.4.** Sizes of independent sets found using the greedy algorithm

$\theta$	edge density	indep. set size
0.05	0.4794	36
0.0	0.2001	12
-0.05	0.0431	5
-0.1	0.005	3
-0.15	0.0005	2

clique problem in a complementary graph. Interestingly, the preprocessing procedure that was very helpful for finding maximum cliques in original graphs was absolutely useless in the case with their complements, therefore we conclude that the maximum independent set appears to be more difficult to compute than the maximum clique in the market graph. Table 1.4 presents the results obtained using the greedy algorithm described above.

As one can see, the sizes of the computed independent sets are very small, which coincides with the prediction that was made in the previous section based on the analysis of the degree distributions.

From the financial point of view, the independent set in the market graph represents “the most diversified” portfolio, where all instruments are negatively correlated with each other. It turns out that choosing such a portfolio is not an easy task, and one cannot expect to easily find a large group of negatively correlated instruments.

#### 1.4.5 Instruments corresponding to high-degree vertices

Up to this point, we studied the properties of the market graph as one big system, and did not consider the characteristics of every vertex in this graph.

However, a very important practical issue is to investigate the degree of each vertex in the market graph and to find the vertices with high degrees, i.e. the instruments that are highly correlated with many other instruments in the market. Clearly, this information will help us to answer a very important question: which instruments most accurately reflect the behavior of the market?

For this purpose, we chose the market graph with a high correlation threshold ( $\theta = 0.6$ ), calculated the degrees of each vertex in this graph and sorted the vertices in the decreasing order of their degrees.

Interestingly, even though the edge density of the considered graph is only 0.04% (only highly correlated instruments are connected by an edge), there are many vertices with degrees greater than 100.

According to our calculations, the vertex with the highest degree in this market graph corresponds to the NASDAQ 100 Index Tracking Stock. The degree of this vertex is 216, which means that there are 216 instruments that are highly correlated with it. An interesting observation is that the degree of this vertex is twice higher than the number of companies whose stock prices the NASDAQ index reflects, which means that these 100 companies greatly influence the market.

In Table 1.5 we present the “top 25” instruments in the U.S. stock market, according to their degrees in the considered market graph. The corresponding symbols definitions can be found on several websites, for example <http://www.nasdaq.com>. Note that most of them are indices that incorporate a number of different stocks of the companies in different industries. Although this result is not surprising from the financial point of view, it is important as a practical justification of the market graph model.

## 1.5 Conclusion

In this paper, we presented a detailed study of the properties of the market graph. Finding cliques and independent sets in the market graph gives us a new tool of the analysis of the market structure by classifying the stocks into different groups.

As it was pointed out above, our experiments show that the distribution of the correlation coefficients between the stocks in the US stock market remains very stable over time. Therefore, the results of the analysis of the market graph can be used for predicting the behavior of the stock market in the future.

Another important result obtained in this paper is that the power-law model, which well describes the massive graphs arising in telecommunications and Internet, is also applicable in finance. It confirms an amazing observation that a lot of real-life massive graphs have a similar power-law structure.

Although we addressed many issues in our analysis of the market graph, there are still a lot of open problems. For instance, since the independent

**Table 1.5.** Top 25 instruments with highest degrees ( $\theta = 0.6$ ).

symbol	vertex degree
QQQ	216
IWF	193
IWO	193
IYW	193
XLK	181
IVV	175
MDY	171
SPY	162
IJH	159
IWV	158
IVW	156
IAH	155
IYY	154
IWB	153
IYV	150
BDH	144
MKH	143
IWM	142
IJR	134
SMH	130
STM	118
IIH	116
IVE	113
DIA	106
IWD	106

sets in the market graph turned out to be very small, there is a possibility to consider quasi-cliques instead of cliques in the complementary graph. This will allow us to find larger diversified portfolios which is important from the practical point of view. Also, one can consider another type of the market graph based on the data of the liquidity of different instruments, instead of considering the returns. It would be very interesting to study the properties of this graph and compare it with the market graph considered in this paper. Therefore, this research direction is very promising and important for deeper understanding of the market behavior.

## Acknowledgments

The authors would like to thank the referees for their comments which helped to improve the quality of presentation.

## References

- J. Abello, P.M. Pardalos and M.G.C. Resende, On maximum clique problems in very large graphs, *DIMACS Series*, 50, American Mathematical Society, 1999, 119-130.
- J. Abello, P.M. Pardalos and M.G.C. Resende, editors, **Handbook of Massive Data Sets**, Kluwer Academic Publishers, 2002.
- W. Aiello, F. Chung, L. Lu, A random graph model for power law graphs, *Experimental Math.* **10** (2001) 53-66.
- S. Arora, C. Lund, R. Motwani, M. Szegedy, Proof verification and hardness of approximation problems, *Journal of the ACM* **45** (1998) 501-555.
- S. Arora and S. Safra, Approximating clique is NP-complete, *Proceedings of the 33rd IEEE Symposium on Foundations on Computer Science* (1992) 2-13.
- V. Boginski, S. Butenko, P. M. Pardalos, Modeling and Optimization in Massive Graphs, P. M. Pardalos and H. Wolkowicz, editors, **Novel Approaches to Hard Discrete Optimization**, American Mathematical Society, 2003, 17-39.
- B. Bollobás, **Extremal Graph Theory**, Academic Press, 1978.
- B. Bollobás, **Random Graphs**, Academic Press, 1985.
- I. M. Bomze, M. Budinich, P. M. Pardalos, and M. Pelillo, The maximum clique problem, D.-Z. Du and P. M. Pardalos, editors, **Handbook of Combinatorial Optimization**, Kluwer Academic Publishers, 1999, 1-74.
- A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, J. Wiener, Graph structure in the Web, *Computer Networks* **33** (2000) 309-320.
- P. Erdős and A. Rényi, On random graphs, *Publicationes Mathematicae* **6** (1959) 290-297.
- P. Erdős and A. Rényi, On the evolution of random graphs, *Publ. Math. Inst. Hungar. Acad. Sci.* **5** (1960) 17-61.
- P. Erdős and A. Rényi, On the strength of connectedness of a random graph, *Acta Math. Acad. Sci. Hungar.* **12** (1961) 261-267.
- M. Faloutsos, P. Faloutsos, C. Faloutsos, On power-law relationships of the Internet topology, *ACM SIGCOMM*, 1999.
- M.R. Garey and D.S. Johnson, **Computers and Intractability: A Guide to the Theory of NP-completeness**, Freeman, 1979.

J. Håstad, Clique is hard to approximate within  $n^{1-\epsilon}$ , *Acta Mathematica* **182** (1999) 105-142.

**ILOG CPLEX 7.0 Reference Manual**, 2000.

D. S. Johnson and M. A. Trick, editors, **Cliques, Coloring, and Satisfiability: Second DIMACS Implementation Challenge**, Vol. 26 of DIMACS Series, American Mathematical Society, 1996.

R. N. Mantegna and H. E. Stanley, **An Introduction to Econophysics: Correlations and Complexity in Finance**, Cambridge University Press, 2000.

D. Watts, **Small Worlds: The Dynamics of Networks Between Order and Randomness**, Princeton University Press, 1999.

D. Watts and S. Strogatz, Collective dynamics of ‘small-world’ networks, *Nature* **393** (1998) 440-442.