Network Models of Massive Datasets

Vladimir Boginski[1], Sergiy Butenko[2], and Panos M. Pardalos[1]

[1]Center for Applied Optimization,
Industrial and Systems Engineering Department
University of Florida, Gainesville, FL 32611
{vb,pardalos}@ufl.edu
[2]Industrial Engineering Department,
Texas A&M University, College Station, TX
butenko@tamu.edu

Abstract. We give a brief overview of the methodology of modeling massive datasets arising in various applications as networks. This approach is often useful for extracting non-trivial information from the datasets by applying standard graph-theoretic techniques. We also point out that graphs representing datasets coming from diverse practical fields have a similar power-law structure, which indicates that the global organization and evolution of massive datasets arising in various spheres of life nowadays follow similar natural principles.

1. Introduction

Dealing with massive datasets of diverse nature and origin is an essential part of many practical applications arising in government and military systems, telecommunications, biotechnology, medicine, finance, astrophysics, ecology, geographical information systems, etc. [2, 10] Understanding the structural properties of a certain dataset is in many cases the task of a crucial importance.

The analysis of massive datasets arising in real-world applications is challenging due to several reasons. One of the important issues addressed in the literature is associated with the excessive size of the datasets, many of which cannot fit into the computer's internal memory. It leads to using external memory devices for the storage of some part of the data, which negatively affects the performance of algorithms applied for processing the data. The research area that addresses this type of problems deals with so-called *External Memory Algorithms* [3]. However, in many cases the enormous size of the dataset vanishes the power of even efficient external memory algorithms. Therefore, it is often very helpful to use an appropriate mathematical model, which can significantly simplify the analysis of a dataset and even theoretically predict some of its properties. Thus, another fundamental problem that arises here is *modeling* massive datasets.

In this paper, we will concentrate on one of the aspects of this problem, which deals with *network representation* of real-world datasets. According to this approach, a certain dataset is represented as a *network*, or *graph*, with certain attributes associated with its vertices and edges.

Studying the structure of a graph representing a dataset is often important for understanding the internal properties of the application it represents, as well as for improving storage organization and information retrieval. One can visualize a graph as a set of dots and links connecting them, which makes this representation convenient and easily understandable.

Network models allow one to extract information from massive datasets using various standard concepts from graph theory. In many cases, one can investigate specific properties of a dataset by detecting special formations in the corresponding graph, for instance, *connected components, spanning trees, cliques* and *independent sets.* In particular, cliques and independent sets are often used for solving the important clustering problem arising in data mining, which essentially represents partitioning the set of elements of a certain dataset into a number of subsets (clusters) of objects according to some similarity (or dissimilarity) criterion.

These concepts are associated with a number of network optimization problems that will be discussed later in this paper.

Another aspect of investigating graph models of massive datasets is studying the degree distribution of the constructed real-world graphs. The degree distribution is an important characteristic of a dataset represented by a graph. It represents the large-scale pattern of connections in the graph, which reflects the global properties of the dataset. One of the important results discovered during the last several years is the observation that many real-life massive graphs representing the datasets coming from diverse areas (Internet, telecommunications, finance, biology, sociology) follow the power-law model [4]. The interesting fact that graphs representing completely different datasets have a similar well-defined power-law structure has been widely reflected in the literature [5, 8, 7, 10, 23, 31, 32]. It indicates that the global organization and evolution of massive datasets arising in various spheres of life nowadays follow similar laws and patterns. This fact served as a motivation to introduce a concept of “self-organized networks”.

Later in this paper, we will discuss in more detail various aspects of modeling massive datasets as graphs. To illustrate the practical importance of graph-theoretic techniques, we will briefly describe several examples of real-life applications of these approaches associated with the datasets arising in telecommunications, finance and biomedicine.

The remainder of the paper is organized as follows. In Section 2, we briefly overview the basic definitions from graph theory. Section 3 discusses the general characteristics of the networks representing real-world datasets. Sections 4, 5 and 6 present the results of applying graph-theoretic approaches to the analysis of different massive datasets. Finally, Section 7 concludes the discussion.

2. Graph Theory Basics

To give a brief introduction to graph theory, we introduce several basic definitions and notations. Denote by $G = (V, E)$ a simple undirected graph with the set of n vertices V and the set of edges E. A multi-graph is an undirected graph with multiple edges.

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 the 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 (frequency) 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. In the case of a directed graph, the concept of degree distribution is generalized: one can distinguish the distribution of *in-degrees* and *out-degrees*, which deal with the number of edges ending at and starting from a vertex, respectively.

The *distance* between two vertices is the number of edges in the shortest path between them (it is equal to infinity for vertices representing different connected components). The *diameter* of a graph G is usually defined as the maximal distance between pairs of vertices of G. Note, that in the case of a disconnected graph the usual definition of the diameter would result in the infinite diameter, therefore the following definition is in order. By the diameter of a disconnected graph we will mean the maximum finite shortest path length in the graph (which is the same as the largest of diameters of the graph's connected components).

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* $C\gamma$, 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 $C\gamma$.

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 $\overline{G}(V, \overline{E})$ which is defined as follows. If an edge $(i,j) \in E$, then $(i,j) \notin \overline{E}$, and if $(i,j) \notin E$ then $(i,j) \in \overline{E}$. Clearly, a maximum clique in $\overline{G}$ is a maximum independent set in G, so the maximum clique and maximum independent set problems can be easily reduced to each other.

A *legal (proper) coloring* of G is an assignment of colors to its vertices so that no pair of adjacent vertices has the same color. A coloring induces naturally a partition of the vertex set such that the elements of each set in the partition are pairwise nonadjacent (i.e., they form independent sets); these sets

are precisely the subsets of vertices being assigned the same color. If there exists a coloring of G that uses no more than k colors, we say that G admits a k-coloring (G is k-colorable). The minimal k for which G admits a k-coloring is called the *chromatic number* and is denoted by $\chi(G)$. The graph coloring problem is to find $\chi(G)$ as well as the partition of vertices induced by a $\chi(G)$-coloring. The graph coloring problem considered for the complementary graph G is referred to as the *minimum clique partition* problem in the original graph G (since an independent set in $\overline{G}$ is a clique in G).

The maximum clique and the graph coloring problems are NP-hard [20]. Moreover, it turns out that that these problems are difficult to approximate [6, 22, 21]. This makes these problems especially challenging in large graphs.

For other standard definitions which are used in this paper the reader is referred to a standard textbook in Graph Theory [9].

3. General Characteristics of Real-World Networks

As it was pointed out above, massive datasets arising in various spheres of life can be represented as networks. One of the most well-known examples of this approach is representing the World Wide Web as a massive graph (known as the Web graph) [14]. Other examples include the call graph arising in the telecommunications traffic data [1], the market graph representing the structure of financial markets [11, 12], as well as social networks where real people are the vertices [13, 23, 31, 32].

These graphs have been empirically studied, and one interesting result was obtained. It turns out that all these graphs coming from diverse applications follow the *power-law model* [4, 5, 7, 10, 14, 16, 23, 31, 32], which states that the probability that a vertex of a graph has a degree k (i.e., there are k edges emanating from it) is

$$P(k) \propto k^{-\gamma}. \tag{1}$$

Equivalently, one can represent it as

$$\text{Log } P \propto -\gamma \log k, \tag{2}$$

which demonstrates that this distribution forms a straight line in the logarithmic scale, and the slope of this line is equal to the value of the parameter γ.

Another interesting observation is the fact that the aforementioned graphs tend to be *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 rather high in these graphs.

These networks are also associated with a well-known “small-world” hypothesis, which claims that despite the large number of vertices, the distance between any two vertices (or, the *diameter* of the graph) is small [31].

One more characteristic that should be mentioned here is referred to as the *scale-free* property of the power-law distribution. A power-law dependency of the form $F(x) \propto x^{-\gamma}$ is scale-free in the sense that it remains the same if x is multiplied by some constant. This property implies that the power-law structure of a certain network should not depend on of the size of the network. Clearly, the considered 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 [7], 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 last principle is in many cases rather natural and easy to understand. For instance, if one considers the social network representing the actors' collaboration (known as the Hollywood graph, where vertices are Hollywood actors, and two vertices are connected by an edge if these two actors have ever appeared in the same movie), it is clear that new actors usually appear in the movies with famous actors who have acted in many movies, and therefore have high degrees. Similar situation is typical for other collaboration networks incorporating scientists, sportsmen, etc. Another example is associated with the Web graph: new websites usually have the links to popular websites.

It should be also noted that the growth and preferential attachment principles can be applied to constructing a formal mathematical procedure of generating power-law graphs satisfying the scale-free property [7].

In the next sections, we present several examples of representing datasets of different origin as large graphs. As we will see, the graph representation is often convenient for visualizing the dataset represented by a graph, and in many cases it provides a deeper insight into its structural properties. Moreover, the considered graphs have the power-law structure, which indicates that the evolution of datasets arising in different applications follows similar natural principles discussed above.

4. Call Graph

One of the examples of applying graph-theoretic techniques to analyzing massive datasets is the *call graph* representing telecommunications traffic data, which was studied by Abello, et al. [1] and Aiello et al. [4]. 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.

The considered one-day call graph representing the data from AT&T telephone billing records had 53,767,087 vertices and over 170 millions of edges. This graph appeared to have 3,667,448 connected components, most

of them tiny; only 302,468 (or 8%) components had more than 3 vertices. A giant connected component with 44,989,297 vertices was computed [1].

The maximum clique problem and problem of finding large quasi-cliques with pre-specified density were considered in this giant component. These problems were addressed using a greedy randomized adaptive search procedure (GRASP) [18, 19]. In short, GRASP is an iterative method that at each iteration constructs, using a greedy function, a randomized solution and then finds a locally optimal solution by searching the neighborhood of the constructed solution. This is a heuristic approach which gives no guarantee about quality of the solutions found, but proved to be practically efficient for many combinatorial optimization problems. To make application of optimization algorithms in the considered large component possible, the authors use some suitable graph decomposition techniques employing external memory algorithms [3].

Abello et al. ran 100,000 GRASP iterations taking 10 parallel processors about one and a half days to finish. Of the 100,000 cliques generated, 14,141 appeared to be distinct, although many of them had vertices in common. The authors suggested that the graph contains no clique of a size greater than 32. Finally, large quasi-cliques with density parameters γ = 0.9, 0.8, 0.7 and 0.5 for the giant connected component were computed. The sizes of the largest quasi-cliques found were 44, 57, 65 and 98, respectively.

It is also important to investigate the degree distribution of the Call graph. According to [4], the distribution of in-degrees and out-degrees of this graph, as well as the distribution of the sizes of the connected components, can be very well represented by a power law.

5. Market Graph

In this section, we describe the recently developed methodology utilizing a representation of the stock market as a large graph based on the correlation matrix corresponding to the set of stocks traded in the U.S. stock market. This graph is referred to as the *market graph*.

The procedure of constructing this graph is relatively simple. Clearly, the set of vertices of this graph corresponds to the set of stocks. For each pair of stocks *i* and *j*, the correlation coefficient C_{ij} is calculated using the following standard procedure.

Let $P_i(t)$ denote the price of the instrument *i* at time *t*. Then

$$R_i(t, \Delta t) = ln \frac{P_i(t + \Delta t)}{P_i(t)}$$

defines the logarithm of return of the stock i over the period from *(t)* to $t + \Delta t$.

The elements of the correlation matrix *C* representing correlation coefficients between all pairs of stocks *i* and *j* are calculated as

$$C_{ij} = \frac{E(R_i R_j) - E(R_i)E(R_j)}{\sqrt{Var(R_i)Var(R_j)}}, \quad (3)$$

where $E(R_i)$ is defined simply as the average return of the instrument i over T considered time periods (i.e., $E(R_i) = \frac{1}{N}\sum_{t=1}^{T} R_i(t)$) [26, 27, 30].

If one specifies a certain *threshold* θ, $-1 \leq \theta \leq 1$, then an undirected edge connecting the vertices i and j is added to the graph if the corresponding correlation coefficient C_{ij} is greater than or equal to θ. The value of θ is usually chosen to be significantly larger than zero, and in this case an edge between two vertices reflects the fact that the corresponding stocks are significantly correlated.

Boginski et al. [11, 12]studied the properties of the market graph constructed using this procedure based on the time series of the prices of more than 6000 stocks traded in the U.S. stock market observed over several partially overlapping 500-day periods during 2000-2002. The intervals between consecutive observations were equal to one day (i.e., the coefficients C_{ij} were calculated using formula (3) with T = 500 and Δt = 1 day). These studies produced several interesting results that are discussed in the next subsections. It should be noted that since the size of the market graph in significantly smaller than the size of the call graph, the analysis of the market graph can be performed in much more detail.

5.1. Edge Density of the Market Graph as a Characteristic of Collective Behavior of Stocks

Changing the values of the correlation threshold θ allows one to construct market graphs where the connections between the vertices reflect different degrees of correlation between the corresponding stocks. It is easy to see that the number of connections (edges) in the market graph decreases as the threshold value θ increases.

The ratio of the number of edges in the graph to the maximum possible number of edges is called the *edge density.* The edge density of the market graph is essentially a measure of the fraction of pairs of stocks exhibiting a similar behavior over time. As it was pointed out above, specifying different values of θ allows one to define different "levels" of this similarity. Figure 5.1 shows the plot of the edge density of the market graph as a function of θ.

On the other hand, one can look at the changes of the edge density of the market graph over time. In [12] these dynamics were analyzed for 11overlapping 500-day periods in 2000-2002, where the 1st period was the earliest, and the 11th period was the latest. In order to take into account only highly correlated pairs of stocks, a considerably large value of θ (θ =0.5) was specified. It turned out that the edge density of the market graph corresponding to the latest period was more than 8 times higher than for the first period. The

corresponding plot is shown in Figure 5.1. The dramatic jump of the edge density suggests that 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, and the structure of the market becomes not purely random. However, one may argue that this “globalization” can also be explained by the specifics of the time period considered in the analysis, the later half of which is characterized by a general downhill movement of the stock prices.

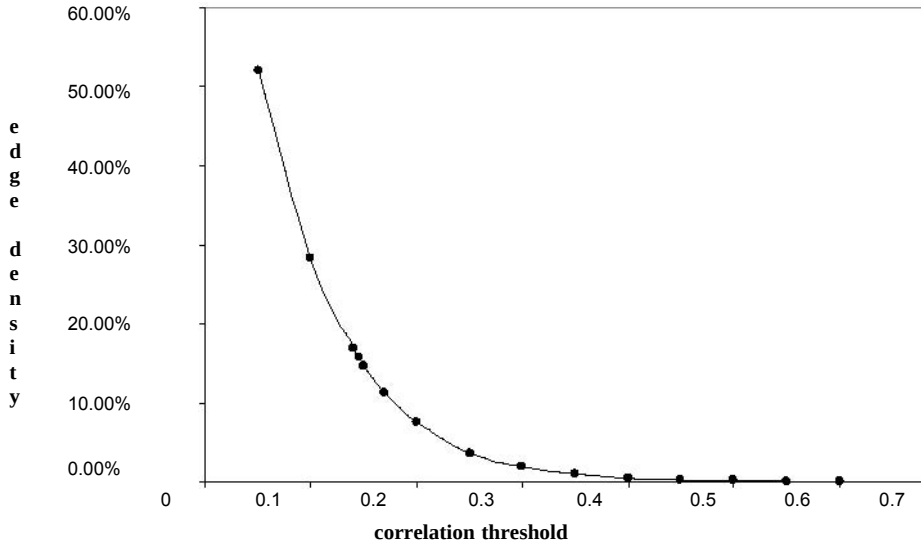


Fig. 1. Edge density of the market graph for different values of the correlation threshold.

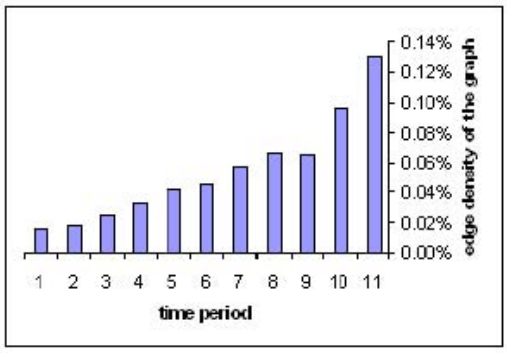


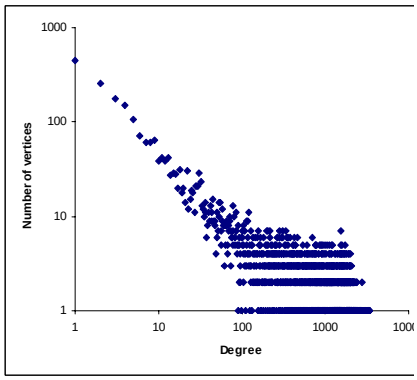
Fig. 2. Evolution of the edge density of the market graph during 2000-2002.

5.2. Global Pattern of Connections in the Market Graph

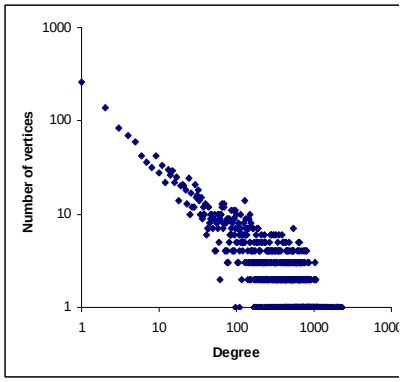
The edge density of the market graph discussed in the previous subsection is a global characteristic of connections between stocks, however, it does not reflect the *pattern* of these connections. For this purpose, the concept of degree *distribution* defined above is utilized.

It turns out that the degree distribution of the market graph has a highly specific power-law structure.

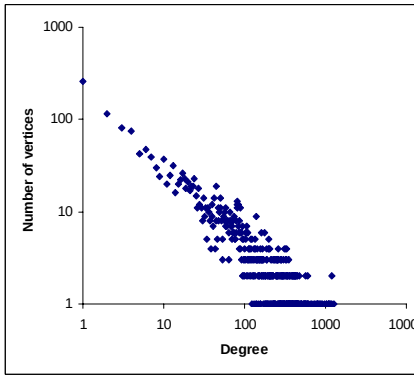
According to [11, 12], the power-law structure of the market graph is stable for different values of θ, as well as for different considered time periods. Figure 3 demonstrates the degree distribution of the market graph (in the logarithmic scale) for several values of θ. In [12], the authors considered the degree distribution of the market graph for 11 overlapping time periods, and the distributions corresponding to four of these periods are shown in Figure 4. As one can see, all these plots are approximately straight lines in the logarithmic scale, which coincides with (2).



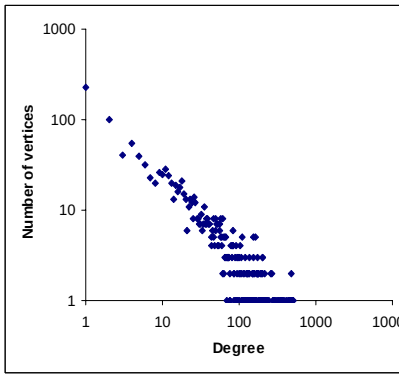
(a)



(b)



(c)



(d)

Fig. 3. Degree distribution of the market graph for a 500-day periods in 2001-2002 corresponding to (a) θ =0.3, (b) θ =0.4, (c) θ =0.5, (d) θ =0.3.

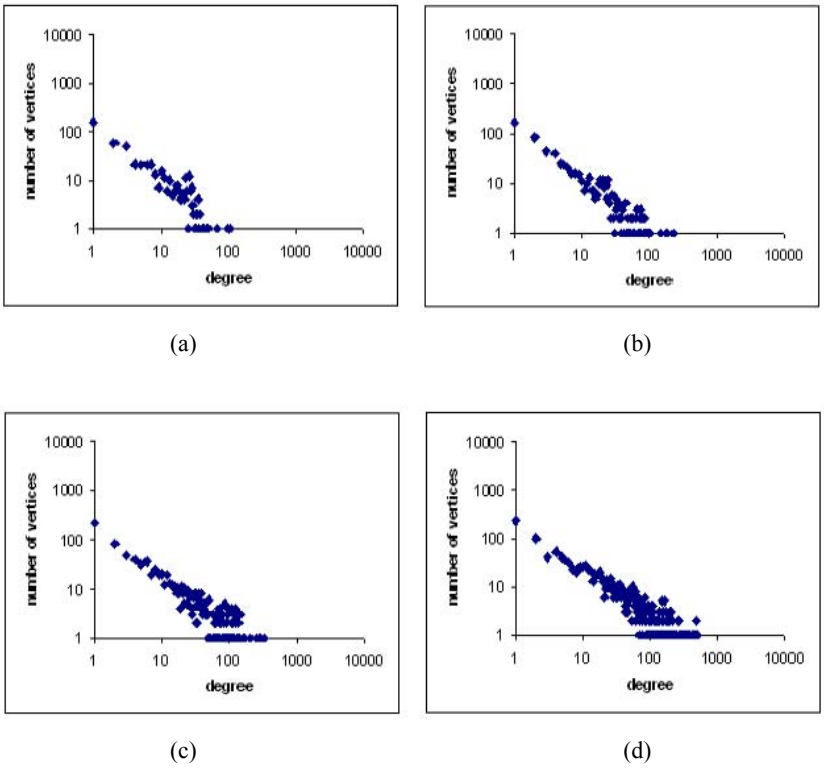


Fig. 4. Degree distribution of the market graph for different 500-day periods in 2000-2002 with θ =0.5: (a) period 1, (b) period 4, (c) period 7, (d) period 11.

The stability of the degree distribution of the market graph implies that there are highly specific patterns underlying the stock price fluctuations.

5.3. Interpretation of Cliques and Independent Sets in the Market Graph

Another significant result of [11] is a suggestion to relate some correlation-based properties of the stock market to certain *combinatorial* properties of the corresponding market graph. For example, the authors attacked the problem of finding large groups of highly-correlated stocks by applying simple algorithms for finding large cliques in the market graph constructed using a relatively large value of correlation threshold. As it was mentioned above, a clique is a set of *completely interconnected* vertices, therefore, partitioning the market graph into

large cliques defines a natural classification of stocks into dense clusters, where any stock that belongs to the clique is highly correlated with all other stocks in this clique. The fact that all stocks in a clique are correlated with each other is very important: it shows that this technique provides a classification of stocks, in which a stock is assigned to a certain group only if it demonstrates a behavior which is similar to *all* other stocks in this group. The possibility to consider quasi-cliques instead of cliques in this classification should also be mentioned. This would allow one to construct larger groups of "similar" stocks while the density of connection within these groups would remain high enough.

Interestingly, the size of the maximum clique in the market graph was rather large even for a high correlation threshold. The details of these numerical experiments can be found in [11]. For example, for θ =0.6 the edge density of the market graph is only 0.04%, however, a large clique of size 45 was detected in this graph. It should be noted that even though the maximum clique problem is NP-hard, the exact solutions of this problem were found for different instances of the market graph, which was which was possible because of the fact that the market graph is clustered (i.e., it contains dense groups of connected vertices).

Independent sets in the market graph are also important for practical purposes. Since an independent set is a set of vertices which are not connected with any other vertex in this set, independent sets in a market graph with a negative value of θ correspond to sets of stocks whose price fluctuations are negatively correlated, or *fully diversified portfolios.* Therefore, finding large independent sets in the market graph provides a new technique of choosing diversified portfolios. However, it turns out that the sizes of independent sets detected in the market graph are significantly smaller than clique sizes [11], which indicates that one would not expect to find a large diversified portfolio in the modern stock market.

The results described in this subsection provide another argument in support of the idea of the globalization of the stock market, which was proposed above based on the analysis of the edge density of the market graph.

The methodology of finding cliques and independent sets in the market graph also provides an efficient tool of performing *clustering* based on the stock market data, i.e., partitioning the set of stocks into clusters of "similar" objects.

The choice of the grouping criterion is natural: "similar" or "different" financial instruments are determined according to the correlation between their price fluctuations. Clearly, 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). If independent sets are used instead of cliques, it would represent the partition of the market into a set *of distinct diversified portfolios*. In this case the minimum possible number of clusters is equal to a minimum number of distinct independent sets, or the *chromatic number* corresponding to the optimal solution of the graph coloring problem.

6. Brain Networks

Another application of network approaches to data analysis is associated with studying human brain. The enormous number of neurons and dynamic nature of connections between them makes the analysis of brain function especially challenging. Analyzing the connectivity of the brain using graph-theoretic approaches is an important practical task, and the results of this analysis can be applied in treatment of various types of brain disorders, for instance, epileptic seizures.

Obviously, the analysis of the graph representing all the neurons as vertices cannot be empirically performed, since the number of neurons in the brain and the connections between them is too large: according to [28], the number of neurons is estimated to be 8.3×10^{9}, and the number of connections is approximately 6.6×10^{13}. However, one can still judge about some properties of this graph, for instance, the fact that despite its low edge density, this graph is clustered, which is also typical for other real-life graphs. Various aspects of the analysis of brain connectivity are discussed in [24].

In order to perform a more detailed quantitative analysis of the graph representing the brain, one can consider much smaller graphs by treating certain groups of neurons (functional units of the brain) as vertices and investigate the connections between these functional units. For instance, this approach was applied to the connectivity analysis of the cortical visual system of the macaque monkey, which was represented by the graph with 32 vertices [17].

Eguiluz et al. [15] studied a relatively large graph corresponding to 147,456 functional units of the human brain, which were selected by dividing the entire brain into a set of 64 × 64 × 36 voxels of a small size. Signals representing the activity of each functional unit were recorded over a certain time period, and these time series were then used for constructing the set of edges of the graph representing these brain sites. The authors utilized the same idea that was used for creating the market graph described in the previous section. In this case, the correlation between each pair of brain units was calculated according to (3) using the time series representing the signals obtained from these units, and the corresponding vertices were connected by an edge if the correlation exceeded a specified threshold value. Interestingly, the degree distribution of the resulting graphs constructed for different correlation thresholds also has the power-law structure with the parameter $\gamma \approx 2$ [15].

It should be noted that the standard correlation coefficient is not the only possible measure of the similarity in the behavior of a pair of brain sites. In [25], the authors apply the statistical concept of *T-index* as the quantitative representation of the entrainment of a pair of functional units of the brain at a certain time moment, which is then utilized in a mathematical programming model for studying the predictability of epileptic seizures.

Clearly, investigating various characteristics of brain networks is a very important practical task. For instance, one would be interested in locating spanning trees and cliques in these networks, which may provide a new insight

into the process of signal propagation between the neurons. This information could be helpful in studying various types of brain disorders.

7. Concluding Remarks

We discussed the network representation approach to studying massive datasets. This paper does not attempt to exhaust this rich and multifaceted research area, but rather represents a brief review of recent results concerning some applications of interest. A common characteristic of the considered examples is a highly dynamic nature of the studied datasets and, as a result, of the corresponding networks. Many other real-life networks representing massive data sets have been actively studied in the past few years. For a recent review of complex networks and their applications in sociology, information technology and biology the reader is referred to [29].

8. References

1. J. Abello, P.M. Pardalos, and M.G.C. Resende. On maximum clique problems in very large graphs, DIMACS Series, 50, American Mathematical Society, 119-130 (1999).
2. J. Abello, P.M. Pardalos, and M.G.C. Resende, editors. Handbook of Massive Data Sets. Kluwer Academic Publishers (2002).
3. J. Abello and J. S. Vitter (eds.). External Memory Algorithms. Vol. 50 of DIMACS Series in Discrete Mathematics and Theoretical Computer Science. American Mathematical Society, 1999.
4. W. Aiello, F. Chung, and L. Lu. A random graph model for power law graphs, Experimental Math. 10, 53-66 (2001).
5. R. Albert, and A.-L. Barabasi. Statistical mechanics of complex networks. Reviews of Modern Physics 74: 47-97 (2002).
6. S. Arora and S. Safra. Approximating clique is NP-complete. Proceedings of the 33rd IEEE Symposium on Foundations on Computer Science, pages 2-13, 1992.
7. A.-L. Barabasi and R. Albert. Emergence of scaling in random networks. Science **286**: 509–511 (1999).
8. A.-L. Barabasi. Linked. Perseus Publishing (2002).
9. C. Berge. Graphs and Hypergraphs. North-Holland Mathematical Library, 6, 1976.
10. V. Boginski, S. Butenko, and P.M. Pardalos. Modeling and Optimization in Massive Graphs. In: Novel Approaches to Hard Discrete Optimization, P. M. Pardalos and H. Wolkowicz, eds. American Mathematical Society, 17-39 (2003).
11. V. Boginski, S. Butenko, and P.M. Pardalos. On Structural Properties of the Market Graph. In: Innovations in Financial and Economic Networks, A. Nagurney (Ed.), Edward Elgar Publishers, 29–45 (2003).

12. V. Boginski, S. Butenko, and P.M. Pardalos. Network-based Techniques in the Analysis of the Stock Market. In: Supply Chain and Finance, P. M. Pardalos, A. Migdalas, G. Baourakis (Eds.), World Scientific, 1–14 (2003).

13. V. Boginski, S. Butenko, and P.M. Pardalos. Collaboration Networks in Sports. In: Economics, Management and Optimization in Sports, S. Butenko, J. Gil-Lafuente, P. M. Pardalos (Eds.), Springer-Verlag, 265–278 (2004).

14. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, J.Wiener.Graph structure in the Web, Computer Networks **33**: 309– 320 (2000).

15. V. M. Eguiluz, G. Cecchi, D. R. Chialvo, M. Baliki, A. V. Apkarian. Scale-free structure of brain functional networks. http://arxiv.org/abs/cond-mat/0309092 (2003).

16. M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the Internet topology, ACM SICOMM (1999).

17. D.J. Felleman and D.C. Van Essen, 1991. Distributed Hierarchical Processing in the Primate Cerebral Cortex. Cereb. Cortex, 1, 1–47.

18. T. A. Feo and M. G. C. Resende, Greedy randomized adaptive search procedures, Journal of Global Optimization **6** (1995) 109-133.

19. T. A. Feo, M. G. C. Resende and S. H. Smith, A greedy randomized adaptive search procedure for maximum independent set, Operations Research **42** (1994) 860-878.

20. Garey, M.R. and Johnson, D.S., 1979. Computers and Intractability: A Guide to the Theory of NP-completeness, Freeman.

21. M.R. Garey and D.S. Johnson. The complexity of near-optimal coloring. Journal of the ACM, 23: 43-49, 1976.

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

23. B. Hayes. Graph Theory in Practice. American Scientist, **88**: 9-13 (Part I), 104-109 (Part II) (2000).

24. C.C. Hilgetag, R. K¨otter, K.E. Stephen, O. Sporns, 2002. Computational Methods for the Analysis of Brain Connectivity, In: G. A. Ascoli, ed., Computational Neuroanatomy, Humana Press.

25. L.D. Iasemidis, P.M. Pardalos, J.C. Sackellares, D-S. Shiau, 2001. Quadratic binary programming and dynamical system approach to determine the predictability of epileptic seizures. J. Combinatorial Optimization 5, 9-26

26. L. Laloux, P. Cizeau, J.-P. Bouchad and M. Potters. Noise Dressing of Financial Correlation Matrices. Phys. Rev. Lett. **83**(7), 1467–1470 (1999).

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

28. J.M. Murre and D.P. Sturdy, 1995. The Connectivity of the Brain: Multi-Level Quantitative Analysis. Biol. Cybern., 73, 529–545.

29. M. E. J. Newman. The Structure and Function of Complex Networks, SIAM Review, 45: 167-256, 2003.

30. V. Plerou, P. Gopikrishnan, B. Rosenow, L.A.N. Amaral, and H.E. Stanley. Universal and Nonuniversal Properties of Cross Correlations in Financial Time Series. Phys. Rev. Lett. **83**(7), 1471–1474 (1999).

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

32. D. Watts and S. Strogatz. Collective dynamics of 'small-world' networks. Nature **393**: 440-442 (1998).

Vladimir Boginski is a PhD candidate and graduate assistant at theDepartment of Industrial and Systems Engineering, University of Florida. He received his bachelor's degree in Applied Mathematics from Moscow Institute of Physics and Technology, and master's degree in Industrial and Systems Engineering from the University of Florida. His research areas include network optimization and data mining.

Sergiy Butenko is an Assistant Professor of Industrial Engineering at Texas A&M University. He received his master's degree in Mathematics from Kiev Taras Shevchenko University in Ukraine, and PhD degree in Industrial and Systems Engineering from the University of Florida. Dr. Butenko's primary research interests are in the areas of optimization, operations research and mathematical programming.

Dr. Panos Pardalos is Professor of Industrial and Systems Engineering at the University of Florida. He is also affiliated faculty member of the Computer Science Department, the Hellenic Studies Center, and the Biomedical Engineering Program. He obtained a PhD degree from the University of Minnesota in Computer and Information Sciences. He has received numerous awards including, University of Florida Research Foundation Professor, Foreign Member of the Royal Academy of Doctors (Spain), Foreign Member Lithuanian Academy of Sciences, Ukrainian Academy of Sciences, and Foreign Member of the Petrovskaya Academy of Sciences and Arts (Russia). Dr.Pardalos is a world leading expert in global and combinatorial optimization. He is the editor-in-chief of the Journal of Global Optimization and the Journal of Computational Management Science, managing editor of several book series, and a member of the editorial board of ten international journals. He is the author of 7 books and the editor of more than 40 books. He has written numerous articles and developed several well-known software packages. His research is supported by National Science Foundation and other government organizations. His recent research interests include network design problems, optimization in telecommunications, e-commerce, and massive computing.