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.

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
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,