Physica A 392 (2013) 2226–2234

Contents lists available at SciVerse ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Minimal spanning tree problem in stock networks analysis: An efficient algorithm

Maman Abdurachman Djauhari *, Siew Lee Gan

Department of Mathematical Sciences, Faculty of Science, Universiti Teknologi Malaysia, 81310 UTM Skudai, Johor Bahru, Johor Darul Takzim, Malaysia

ARTICLE INFO

Article history:
Received 21 July 2012
Received in revised form 26 November 2012
Available online 3 January 2013

Keywords:
Adjacency matrix
Euclidean distance
Membership function
Sub-dominant ultrametric
Ultrametric distance

ABSTRACT

Since the last decade, minimal spanning trees (MSTs) have become one of the main streams in econophysics to filter the important information contained, for example, in stock networks. The standard practice to find an MST is by using Kruskal's algorithm. However, it becomes slower and slower when the number of stocks gets larger and larger. In this paper we propose an algorithm to find an MST which has considerably promising performance. It is significantly faster than Kruskal's algorithm and far faster if there is only one unique MST in the network. Our approach is based on the combination of fuzzy relation theory and graph theoretical properties of the forest of all MSTs. A comparison study based on real data from four stock markets and four types of simulated data will be presented to illustrate the significant advantages of the proposed algorithm.

1. Introduction

There are two main streams in stock networks analysis. The first is about the mathematical law that governs the time series data of stock prices. That law has a long history before it becomes a main stream in the study of stock prices. The empirical study is dated back to 1863 when Jules Regnault formulated 'the law on the square root of time' [1]. It says that if $p_i(t)$ is the price of stock i at time t and $p_i(t)$ is a Brownian motion process, then the deviation $|p_i(t+\Delta t) - p_i(t)|$ is about $\sqrt{\Delta t}$. But, it was Bachelier in 1900 who first derived a mathematical model of the dynamic of stock prices. He assumed that the changes in stock prices are subject to fixed mathematical laws. With this assumption he then used the increments of Brownian motion to model absolute price changes. Since this model could make the prices of stocks become negative [2], instead of modeling the absolute price changes, now it is common to model the relative price changes by using geometric Brownian motion (GBM) process. This model was popularized by Paul Samuelson who received the Nobel Prize in Economics in 1970 [1]. The popularity is due to the fact that this model facilitates a convenient work environment. More specifically, let $p_i(t)$ be a GBM process and $r_i(t)$ be the logarithm of i-th stock price return at time t. Then $r_i(t)$ follows a Gaussian distribution. Thus, under this model, similarity among stocks viewed as time series could be measured by the Pearson correlation coefficient.

The second is the search for the tool to filter the important information in stock networks. Only recently has the graph theoretical approach been used to filter the information contained in stock networks governed by GBM process. It was pioneered by Mantegna in 1999 [3]. He used the minimal spanning tree (MST) to study the topological properties of stocks and the corresponding sub-dominant ultrametric (SDU) to understand the hierarchical tree of stocks. After the work of Borůvka in 1926, people had to wait for 73 years before MST became another mainstream in stock networks analysis. Nowadays, the role of MST together with SDU as information filter can be found in many areas of financial industries and also

* Corresponding author. Tel.: +60 17 5200795; fax: +60 7 55 66162.
E-mail addresses: maman@utm.my (M.A. Djauhari), slgan3@live.utm.my (S.L. Gan).

0378-4371/$ – see front matter
doi:10.1016/j.physa.2012.12.032

in many areas of scientific investigations. See, for example, in general complex system [4], foreign exchange [5], economy analysis [6], portfolio analysis [7], risk assessment in term of single linkage [8], trading [9], and volatility [10].

Unlike the use of GBM process which has been accepted as the appropriate model to describe stock price, the construction of MST and SDU to filter the important information in stock networks is still in development. A significantly faster algorithm than the known ones is always sought after. It is in this spirit that this paper is presented; the spirit to contribute in searching for a better solution to the MST problem.

1.1. The role of MST in stock networks analysis

Network among stocks is a complex system [4]. It is usually represented as correlation networks where the interrelationship between two different stocks is quantified by using Pearson coefficient of correlation (PCC) between their respective logarithm of price return. This linear relationship is justified by the use of GBM process to model the mathematical law of $p_i(t)$. Under that model, the random variable $r_i(t)$ follows a Gaussian distribution and thus the similarity between two different stocks can be represented in terms of PCC. The network is then summarized in the form of a correlation matrix [3,7,11] and analyzed through the corresponding distance (dissimilarity, in general) matrix. This is to understand the topological structure and economic classification of the stocks [12]. If the topological structure is helpful in identifying the topological properties of the stocks such as stocks' centrality, which are useful in the study of the relative position of a particular stock with respect to the others, economic classification is to identify the stocks that have similar behavior [3,11].

The topological properties are usually identified by using the network topology representing an MST in the distance networks [13] while the economic classification in the form of an indexed hierarchical tree is constructed based on the SDU of the distance networks. Due to that strategic role of MST and SDU, it is important to note here three properties of MST in conjunction with the SDU. First, once an MST has been obtained, then it can be used to find the SDU. This is the current practice in stock networks analysis. Second, although the MST in a given distance networks might not be unique, the SDU always is. Third, the SDU can actually be obtained without passing through MST. A recent result is presented by Djauhari in Ref. [14]. He shows that if one has obtained the SDU, then only the forest of all MSTs can directly be identified and not an MST except of course when there is only one single MST in the network. We will discuss further the last two properties in Section 3.

1.2. The MST problem

Recent work on the history of the MST problem shows that there were a variety of independent discoveries of the algorithms and ideas to solve that problem [15]. See also Refs. [16,17] for other important historical background on that problem. The first efficient solution, which has become popular since the last decade, is credited to the work of Borůvka in 1926. Since Borůvka's algorithm requires that all edges have distinct weights, which is very rare in a real situation, in the current practice among stock players and researchers the MST in stock networks analysis is usually obtained from the two most suggested algorithms in the literature, i.e., Kruskal's algorithm and Prim's algorithm [3,11,16]. See also Zhang et al. Ref. [6] who use these algorithms in economy analysis and also Ulusoy et al. [9] in trading.

The popularity of those two algorithms motivates Huang et al. [18] to conduct a study to compare their computational complexity. They have reported the conditions where one algorithm is superior to the other. However, many researchers agree that the prominent role is played by Kruskal's algorithm. It is mathematically very appealing [15] and easy to formulate although it is not the fastest algorithm [17]. This is perhaps the reason why many researchers prefer to use Kruskal's algorithm instead of Prim's algorithm. See, for example, Refs. [5,19,20].

In this paper an efficient algorithm which considerably improves the running time of Kruskal's algorithm will be proposed. Since Kruskal's algorithm determines an MST directly from the distance networks, its performance becomes slower and slower when the number of stocks gets larger and larger, say of the order of hundreds. The proposed algorithm does not search MST directly from the distance networks. It searches MST in two steps. The first step is to find the forest of all possible MSTs (or the 'forest' in brief) in the distance networks. The second is the search for an MST from the forest. The details are in Section 3. This method will reduce considerably the running time needed to obtain an MST and the SDU simultaneously.

In the next section, we begin our discussion by recalling the construction process of the SDU from the fuzzy relation viewpoint discussed in Ref. [14]. This will allow us to see in greater depth the properties of distance networks which will be useful to find the SDU efficiently. Based on the SDU, in Section 3 we construct the forest, then find an MST in the sub-graph of the forest obtained after a repeated process of leaves (nodes of degree one) removal by using Kruskal's algorithm, and finally construct an MST in the network. A comparison study with Kruskal's algorithm and also Prim's algorithm in terms of the running time, discussed in Section 4, will illustrate the significant advantages of the proposed algorithm and concluding remarks in the last section will close this presentation.

2. Construction of SDU from fuzzy relation viewpoint

In this section we recall and discuss briefly the standard procedure to filter the important information contained in stock networks. Let E be a set of n stocks in a portfolio under study, and again $p_i(t)$ and $r_i(t)$ be the price of stock i and the logarithm