Physica A 419 (2015) 108–114

Contents lists available at ScienceDirect

Physica A

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

Optimality problem of network topology in stocks market analysis

Maman Abdurachman Djauhari [a,*], Siew Lee Gan [b]

[a] Institute for Mathematical Research, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor Darul Ehsan, Malaysia

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

HIGHLIGHTS

- The MST-based network topology is not optimal in terms of topological properties.
- Therefore, the economic interpretation of that network might be misleading.
- A set of optimality criteria and a selection method of an optimal MST are developed.
- The advantage of the proposed method is illustrated by using NYSE data.

ARTICLE INFO

Article history:
Received 20 January 2014
Received in revised form 24 September 2014
Available online 16 October 2014

Keywords:
Business sectors
Closeness centrality
Correlation networks
Market index
Stock networks

ABSTRACT

Since its introduction fifteen years ago, minimal spanning tree has become an indispensible tool in econophysics. It is to filter the important economic information contained in a complex system of financial markets' commodities. Here we show that, in general, that tool is not optimal in terms of topological properties. Consequently, the economic interpretation of the filtered information might be misleading. To overcome that non-optimality problem, a set of criteria and a selection procedure of an optimal minimal spanning tree will be developed. By using New York Stock Exchange data, the advantages of the proposed method will be illustrated in terms of the power-law of degree distribution.

1. Introduction

The stock price, as showed in stocks markets, is the reflection of the corresponding company. However, its day-to-day behavior is not merely constrained by the company's own fundamentals; it is influenced by the other companies traded in the market and by the economic factors [1]. Those interrelations among stocks are well-known [2] and quantified in terms of Pearson correlation coefficient [3,4]. It is then customary to summarize those interrelationships in the form of a symmetric matrix C of size $(n \times n)$, called correlation matrix [2,5–8], where n is the number of stocks under study. Specifically, let $p_i(t)$ and $r_i(t)$ be the price of stock i and the logarithm of ith stock's price return at time t, respectively. Thus,

$$r_i(t) = \ln\ p_i(t) - \ln\ p_i(t-1)$$

* Corresponding author. Tel.: +60 17 5200795; fax: +60 3 89466973.
E-mail addresses: maman_abd@upm.edu.my, maman_djauhari@yahoo.com (M.A. Djauhari), slgan3@live.utm.my (S.L. Gan).

http://dx.doi.org/10.1016/j.physa.2014.09.060
0378-4371/

Table 1
Distribution of stocks in each sector.

Shape	Industry sector	Number of stocks
□ : Square	Oil and gas	12
■ : Black square	Industrials	16
○ : Circle	Consumer goods	12
● : Black circle	Telecommunications	2
△ : Triangle	Financials	19
▲ : Black triangle	Health care	12
◇ : Diamond	Consumer services	11
◆ : Black diamond	Technology	4
⊠ : Crossed square	Basic materials	6
⊗ : Crossed circle	Utilities	4
	Total number of stocks	98

for all $i = 1, 2, \ldots, n$. The element of the ith row and jth column of C is [3,4,9],

$$c(i,j) = \frac{\langle r_i r_j \rangle - \langle r_i \rangle \langle r_j \rangle}{\sqrt{\langle r_i^2 \rangle - \langle r_i \rangle^2}\sqrt{\langle r_j^2 \rangle - \langle r_j \rangle^2}}$$

where $\langle r_i \rangle$ is the average of $r_i(t)$ for all t.

As C represents a complex network among stocks [10,11], the current practice in analyzing C is to filter the important economic information contained therein by using minimal spanning tree (MST) [2–12] and sub-dominant ultrametric (SDU) [2,5–8,12]. If MST could help to filter the topological structure of the stocks and explain the relation between that structure and economic classification of the stocks [13], the economic classification itself is given by the SDU [3,4,10] in the form of an indexed hierarchical tree.

Since its introduction 15 years ago by Mantegna [3], MST has become an indispensible tool not only in econophysics to filter important information contained in correlation networks [4,9,14] but also in many areas of scientific investigations. For example, in complex systems [7,15], degree distribution of individual stock listed on the S&P500 and KOSPI200 [7,8], emerging markets [16–19], equity markets [2,20], financial markets [21–23], foreign exchange [8,24], portfolio analysis [23], risk assessment [25], trading [26], and volatility [2].

From the literature, we learn that until the present day, the standard practice to find MST is by using Kruskal's algorithm [3–5,7,8,12,26–32] or Prim's algorithm [6,7,18,26,27,29,32]. Interestingly, those algorithms give the same result [29]. This motivates Huang et al. [33] to compare those algorithms to each other in term of computational complexity. They had reported the condition where one algorithm is superior to the other.

Here we show that, according to the topological properties, the MST issued from any of those algorithms in general does not provide an optimal network topology except when there is only one unique MST in the network. Thus, if there is more than one MST, it might produce misleading economic interpretation. This is due to the fact that the algorithms provide only one MST among all possible MSTs that might exist in the network [34]. In the next section, (i) a preliminary study on New York Stock Exchange (NYSE) data, which shows that an MST issued from Kruskal's algorithm might not be optimal, will be reported and (ii) a hypothetical example will show that the result of Kruskal's algorithm depends on the way we order the stocks to define the row and column of the network C. To overcome the non-optimality problem, in Section 3 which is the technical part of the paper, a set of optimality criteria is introduced and a selection procedure of optimal MST is proposed. Later on, a case study of NYSE will be presented and discussed. Finally, concluding remarks in the last section will close this presentation.

2. Motivation: case of NYSE

In this section a preliminary study on NYSE data will be reported. Initially, there were 100 stocks traded there. However, due to the availability of data in the time period of investigation, there are only $n = 98$ stocks to be analyzed. The daily data of closing prices [11] during the whole year of 2012 were used. They were downloaded from website of finance [35] while the list of stocks and the 10 business sectors are provided in the database of NYSE [36]. The distribution of stocks in each sector, represented in different shapes, is given in Table 1.

In what follows we show the non-optimality of MST issued from Kruskal's algorithm. Since $n = 98$, the correlation network consists of $98(98-1)/2 = 4735$ interrelations. To analyze such complex networks, as mentioned earlier, the current practice is to find an MST by using Kruskal's algorithm. If the correlation matrix C is defined in such a way that the ith row (and column) corresponds to the ith stock listed in those databases [35,36], the result is presented in Fig. 1.

Our preliminary study shows that this MST is not the only one contained in the correlation network C. There are a lot of possible MSTs in C. Thus, it is hard to believe that this MST is an optimal one. As a matter of fact, if there are more than one MST in the network, in relation with network topological properties there are two important properties of Kruskal's algorithm that must be taken into consideration; (i) among all possible MSTs, it gives one MST only, and (ii) the result depends on the way we order the stocks to define the row and column of C.