Graph and Network Methods Applied to Financial Markets – Literature Review

Marek GRÓF, Marcel REŠOVSKÝ
Technical University of Košice, Faculty of Economics
marek.grof@tuke.sk, marcel.resovsky@tuke.sk

Abstract

This paper presents a summary of the research performed in the field of applying graph theory methodology in the study of financial markets. Literature review concerning the methods of enforcing a Minimum Spanning Tree structure among the financial assets and the methodology of using a selected correlation threshold to identify possible relations among these assets is presented.

Keywords: Minimum Spanning Tree, correlation threshold, stocks, markets, financial network

1 INTRODUCTION

The financial markets represent a widely studied subject within economic theory. Various methods are used to study the price evolution of assets. One of the relatively new approaches is the method of studying the formation and existence of network structures within the market itself. This paper aims to summarize the research performed in the field of applying graph theory methodology in the study of financial markets, specifically the methods of enforcing a Minimum Spanning Tree structure among the financial assets and the methodology of using a selected correlation threshold to identify possible relations among these assets. Both these methods can be highly valuable in the fields of portfolio management and risk management for investors.

2 MINIMUM SPANNING TREE METHOD

Bonanno et al. studied topological properties of financial network of 1071 stocks that were traded at the New York Stock Exchange (NYSE) between 1987 and 1998. Based on daily price return $r_i(t)$ time series of asset $i$, $N \times N$ correlation matrix can be calculated. Bonanno et al. did not provide the exact method of calculating the price return and correlation matrix explicitly, but most authors (for example Huang et al., Onnela et al., Tabak et al, Tse et al.) computed price return of asset $i$ using logarithmic
return \( r_i(t) = \ln P_i(t) - \ln P_i(t - 1) \) and subsequently calculated the correlation between two assets \( \rho_{ij} \) as:

\[
\rho_{ij} = \frac{\langle r_i^t r_j^t \rangle \langle r_i^t \rangle \langle r_j^t \rangle}{\sqrt{\langle (r_i^t)^2 \rangle \langle (r_j^t)^2 \rangle - \langle (r_i^t) (r_j^t) \rangle^2}}
\]  

(1)

where \( \langle \cdot \rangle \) represents the average over the studied time period. Then, from the correlation coefficient \( \rho_{ij} \), metric distance \( d_{ij} \) between stocks \( i \) and \( j \) in graph \( G \) representing financial network can be expressed as:

\[
d_{ij} = \sqrt{2(1 - \rho_{ij})}.
\]  

(2)

Distance matrix \( D \) containing metric distances \( d_{ij} \) between assets in graph \( G \) is then used to form a minimum spanning tree (MST)\(^1\). Bonanno et al. compared MST created from time series of real price returns \( r_i(t) \) with a) time series of price returns obtained from a simple model \( r_i(t) = \epsilon_i(t) \), where \( \epsilon_i(t) \) is Gaussian random variable with zero mean and unit variance; b) time series obtained from one factor model that can be written as follows:

\[
r_i(t) = \alpha_i + \beta_i r_M(t) + \epsilon_i(t),
\]  

(3)

where \( r_i(t) \) is the return of asset \( i \) at day \( t \), \( r_M(t) \) is the return of a market factor (authors considered S&P 500 as the market factor) and \( \epsilon_i(t) \) is the white noise. Parameters \( \alpha_i \) and \( \beta_i \) in the one factor model can be estimated applying the ordinary least square method on the time series of real price returns and then these estimated parameters are used to generate artificial price returns of the one factor model. Topological differences between MST constructed from real price returns and price returns generated by the one factor model can be seen in Figure 1.

\(^1\) Minimum spanning tree (MST) is connected acyclic graph that contains all vertices of the graph and minimizes the sum of weights of edges. Each vertex in MST is connected to any other vertex without forming loops or cycles. Methods for MST construction are known as nearest neighbor single linkage cluster algorithms. The most often used algorithms are Prims and Kruskals.
From Figure 1 it is obvious that topological properties of correlation based MST of real price returns and price returns obtained from the one factor model are different. The first MST has a hierarchical structure, the second one has a starlike structure with a central node. Bonnano et al. also compares the MST of the real and artificial data using topological characteristics. The first one used is the distribution of vertex degrees (distribution of the number of links originating in each vertex). “In many real networks it has been shown that the degree is distributed according to power law\(^2\) signaling the presence of long range correlations.” (Bonanno) The presence of power law in vertex degree distribution demonstrates the conclusions of the research of Onnela et al., Namaki et al., Bonanno et al. realized on financial networks. On the other hand, based on analyzing the Chinese stock market, Huang et al. concluded that the power law model is invalid in financial networks. The second topological characteristic that Bonanno et al. studied in the financial network, drainage basin area, is more frequently used for oriented graphs of river networks. For the purposes of correlation based MST analysis it can be characterized as the “number of steps each node is far from the most connected node (sink).” (Bonanno)

Figure 2 depicts the distribution of vertex degrees for the real data and artificial data estimated using the random and one factor model. The real data is represented by a circle, data obtained from the random model is represented by a triangle and data estimated using the one factor model is depicted by a square. Grey area represents the

\(^2\) Vertex degree distribution function \(P_k\) follow power law if \(P_k \propto k^{-\gamma}\) where \(\gamma\) is a parameter.
95% confidence interval for the degree distribution of the one factor model. As we can see from Figure 2, the frequency distribution is very different for all of these three models. Models differ not only in the frequency distribution of vertices degrees, but also in the maximal value of vertex degree (115 for real data MST, 7.34±0.92 for the random model, 718±29 for the one factor model). Bonanno et al. concluded that the “topology of MST of one factor model is very different from the MST of real data” although one factor model is able to explain more than 80% of correlation coefficients observed in real data”, also claiming that the “comparison shows that the random and the one factor model fail to describe the topological properties of the MST extracted from a portfolio of stocks simultaneously traded in a financial market. Topological properties of this financial system can be therefore used to falsify widespread financial models.” (Bonanno)

![Figure 2 Distribution of vertex degree for the real data MST (circle), random model (triangle) and one factor model (square)](source: [1])

Onnela et al. investigated relations between 477 stocks at (NYSE) over the time period of 1980-1999. They divided the dataset into $M$ overlapping windows of width $T$. The decision to divide the data into time windows was made in order to smooth the data. The stock data analysis was realized using two different types of network structures: asset tree and asset graph where asset tree is constructed according to methodology of Bonanno et al. When constructing asset graph, only $N - 1$ shortest edges (strongest correlation connections between stocks) from distance matrix $D$ are
taken into consideration. The difference between the asset tree and the asset graph is that the mechanism of asset tree construction guarantees connectedness of this kind of graph. On the other hand, the asset graph does not have to be connected (some vertices can be isolated and there can be multiple components in this structure). To characterize the market structure as a whole Onnela et al. proposed the following normalized measures:

- **mean distance**
\[
\bar{d}(t) = \frac{1}{N(N-1)} \sum_{i,j \in D^t} d_{ij}^t
\]

- **mean correlation coefficient**
\[
\bar{\rho}(t) = \frac{1}{N(N-1)} \sum_{i,j \in D^t} \rho_{ij}^t
\]

- **normalized tree length**
\[
L_{MST}(t) = \frac{1}{N-1} \sum_{i,j \in T^t} d_{ij}^t
\]

In a similar manner as the normalized tree length in the asset tree, normalized graph length in the asset graph can be defined. Evolution of the network structure over time was investigated by Onnela et al. using single and multi-step survival ratios. Single-step survival ratio \(\sigma(t) = \frac{1}{N-1} |E^t \cap E^{t-1}|\) expresses the fraction of edges \(E^t\) of the network structure at time \(t\) that were also present at time \(t - 1\). Multi-step survival ratio \(\sigma(t,n) = \frac{1}{N-1} |E^t \cap E^{t-1} \cap E^{t-2} \cap \cdots \cap E^{t-n+1} \cap E^{t-n}|\) expresses the fraction of edges that are present in the network structure over the whole time period. The development of single and multi-step survival ratio as a function of time is depicted in Figure 3 and Figure 4. Single-step survival ratio is higher for the asset graph where 94.8% of connections survive to the next time period (for asset tree only 82.6% of connections survive). Similarly Figure 4 shows that the curve representing surviving connections of the asset graph decays more slowly than the curve for the asset tree over almost the whole period. Dashed line in Figure 4 represents half-life defined as “time in which half the number of initial connections have decayed.” (Onnela et al.) We can see that the half-life of connections in the asset graph is longer than the half-life of connections in the asset tree.
Figure 3 Single-step survival ratio as a function of time (thicker curve for the asset graph)
Source: [6]

Figure 4 Multi-step survival ratio as a function of time
Source: [6]
Micciche et al. compared the characteristics of the MST of daily price returns and the MST of volatility which they define by the relation

$$\sigma_i(t) = 2[\max\{P_i(t)\} - \min\{P_i(t)\}] / [\max\{P_i(t)\} + \min\{P_i(t)\}], \quad (7)$$

where $\max\{P_i(t)\}$ ($\min\{P_i(t)\}$) are the highest (lowest) prices for the given trading day. For the construction of the volatility MST Micciche et al. used the Spearman rank-order correlation coefficient that is more stable and more appropriate (when comparing to Pearson correlation coefficient) with respect to the dynamics of the stock degrees. The dataset contained prices of 93 most capitalized stocks traded on the US market over a 12-year period. The time series were divided into overlapping time windows of length $T$. The analysis was based on comparing the vertex degree time series of the MST of price returns and the MST of volatility. From the vertex degree time series, graph of autocorrelation functions for both MST obtained from price returns and volatility were constructed leading to a conclusion that the “the stability of the degree of MST is lower for volatility time series than for price return time series” (see Figure 5).
3 THRESHOLD CORRELATION METHOD

Huang et al. researched the interconnectedness between stocks traded in the Shanghai and Shenzhen stock markets in China. 1080 stocks over the period 2003-2007 have been considered. The approach is similar to abovementioned approaches but differs in considering the links between stocks when creating stock network. Edges between stocks in the stock network were created only when the appropriate correlation coefficient was greater than a given threshold value $\theta; \rho > \theta; -1 \leq \theta \leq 1$. 

Figure 5 Autocorrelation function of price return MST and volatility MST for time windows with different length T (115, 227, 357 days)

Source: [4]
With respect to different threshold values four topological characteristics of the network were studied. The first one was vertex degree distribution and if it followed the power law. The second one was clustering coefficient that can be defined for vertex $i$ as a ratio

$$\frac{m_i}{k_i(k_i-1)/2}$$

(8)

where $m_i$ is a number of edges between $k_i$ nearest neighbours. Number of connected components and maximum component size (number of vertex in the component) was the third topological characteristic. The last characteristic was the maximum clique size that is the largest clique in the graph (there is no clique with more vertices included in it) where “a subset $C \subseteq V(G)^3$ is a clique if $G(C)$ is a complete graph, i.e. it has all possible edges.” (Huang et al.) A very important question in social or financial network analysis is how stable the connections between analyzed objects are. Huang et al. used vertex attack method to test topological stability of correlation based stock network. Vertices were removed by stochastic removal method (vertices are removed randomly) or by selective removal method (in certain order – the first is removed vertex with the highest degree). The stability of the network is then expressed by the ratio of maximum component size of the new network (after some vertex removal) to the original network (before any vertex removal) $R_{|CO|}$ ($0 < R_{|CO|} < 1$). The topological stability of the network expressed by aforementioned ratio with respect to proportion of removed vertices $f$ is depicted in Figure 6.

---

$V(G)$ is a set of vertices of graph $G$. 
Tse et al. criticized the MST approach when creating financial network for stocks. He rather used the second approach - considering links between stocks only when the cross-correlation coefficient between stocks is larger than a given threshold value $\theta$. This approach was then applied on 19,807 stock closing prices, volumes and price returns over the time period from July 2005 to August 2007. Further work included the comparison of network characteristics such as the number of nodes, number of connections, average clustering coefficient, average vertex degree within created networks with respect to different threshold value $\theta$. Most connected stocks in the network have various properties of central nodes and thus their prices and returns are reflected in other stocks in the network that are directly or indirectly connected to them. From all stocks Tse et al. identified 10% most connected stocks and based on these stocks an index was created that can be formulated as

$$Index = \frac{\sum_i (price \times number \ of \ shares_i)}{total \ market \ value \ of \ stocks \ during \ base \ period} \quad (9)$$
The development of newly created index was then compared to the development of S&P 500, Dow Jones and Nasdaq Composite Index over the period of July 2005 to August 2007.

The approach of Namaki et al. to creating the network structure of stocks traded in Tehran Stock Exchange is similar to Tse et al., Huang et al. and is based on the threshold correlation of stock price returns. Correlation matrix contains not only authentic but also spurious correlations between stocks. To separate the spurious part of correlation the largest eigenvector of correlation matrix was used as a collective evolution of a large group of stocks. In this way the evolution of a stock index can be determined. Namaki et al. considered a one-factor model when filtering the market effect (similar to Bonanno et al.)

\[
 r_i(t) = \alpha_i + \beta_i M(t) + \epsilon_i(t),
\]

where \(\alpha_i\) and \(\beta_i\) are coefficients of the model, \(\epsilon_i\) are residuals and \(M(t)\) is market effect that can be approximated by

\[
 M(t) = \sum_{i=1}^{N} u_i \eta_i(t),
\]

where \(u_i\) is the largest eigenvector of the correlation matrix. For construction of the financial network of stocks, correlation matrix of estimated residuals \(\epsilon_i\) from the one factor model is used. Namaki et al. then characterized the network using characteristics similar to the ones used by the previous authors, like the vertex degree distribution, clustering coefficient or component structure.

4 CONCLUSION

The paper presented a review of research performed in the field of applying graph and network methodology in financial markets. The first paper presented written by Bonanno et al. presented the basis for all the following articles. Here, authors introduced the method of applying the Minimum Spanning Tree structure formulation in an effort to identify potential networks formed within financial markets. The study was performed using 1071 stocks from the NYSE. The methodology presented was then adopted by the following authors: Onnela et al., Micciche et al., Tabak et al. and Roy et al. Individual contributions of these authors were presented in the paper. A modified approach of forming the asset graph using a correlation threshold method was introduced by Huang et al. along with the introduction of various network characteristics used to compare individual networks. This approach was to a various degree adopted by authors Tse et al. and Namaki et al. whose studies are also summarized. The overall aim of this paper was to provide an overview of research performed in the given field and to serve as a basis for further research.
REFERENCES


