Analysis and Prediction of Fluctuations for Sector Price Indices with Cross Correlation and Association Based Networks: Tehran Stock Exchange Case

Arash Negahdari Kia¹, Saman Haratizadeh² and Zainabolhoda Heshmati³

Network science has become an ever-increasing and interesting field of research in the recent decade by focusing on finding hidden knowledge in complex networks. This study of complex relationships in network structures has also gained a lot of interest in the world of finance and stock markets. This study focuses on Tehran Stock Exchange (TSE), looking into the market price indices data of different market sectors and their fluctuations over time. Four different network structures have been extracted from the TSE market data, two with association rules mining and two with Pearson cross correlation. Using the correlation with different threshold cuts, different networks have been created and importance of market sectors has been analyzed using different centrality measurements. After that, by using Apriori algorithm to find association rules in fluctuations of the price indices, many patterns are extracted for building different directed networks. The networks created by these patterns are used in assessing current market dynamics as well as predicting future market price fluctuations that is tested through an evaluation method.

1. Introduction

Network science has shown enormous applications in many interdisciplinary fields in the recent decade. Any structure or natural phenomena that can be modeled as a network of nodes and edges can be studied by the means of graph theory and network science. This study is about using networks in the field of finance and economy, in particular, analysis of Tehran Stock Exchange (TSE) market and fluctuation prediction of different market sectors price-indices. In this study, association rules are used in an extra-ordinary way to study fluctuation patterns of the market.

¹ Faculty of New Sciences and Technologies University of Tehran. E-mail: nkia.arash@ut.ac.ir
² Faculty of New Sciences and Technologies University of Tehran. E-mail: haratizadeh@ut.ac.ir
³ Faculty of New Sciences and Technologies University of Tehran. E-mail: zheshmati@ut.ac.ir
There are many previous studies done mostly in recent decade that focus on analysis and prediction of financial markets with complex networks. The most cited research that was done in 2003, by Giovanni et al. was about studying New York stock market with help of correlation to make networks of stock prices time series and finding minimal spanning trees to have a better understanding of the topology of stock market [1]. Complex networks analysis is a new tool for understanding many different aspects of financial markets that could not be fully understood before and makes an important role in new studies of financial markets. The importance of this tool was discussed in a research by Gatti et al. in 2010 [2].

In a study in 2004, various applications of network science in finance were presented by Caldarelli et al. [3]. This study showed some applications of graph theory methods that could be useful in finance and economy. With the spread of using complex networks in finance and economy, many studies tried to focus on different aspects of markets for building different kinds of networks. By studying these networks, the researchers found out a better knowledge of the financial markets in many aspects. Some studies are presented in this section as examples of how the researchers made networks out of financial data and what they found out with the help of network science.

In a research in 2005 by Garlaschelli et al., a network description of large market investments was proposed where stocks and shareholders were vertices and the edges of the network were weighted and corresponded to shareholdings [4].

In 2007, another study by Naylor et al. used two hierarchical methods, to develop a topological influence map for some currencies from a distance matrix. They used minimal spanning trees and ultra-metric hierarchical trees to understand the topology of complex networks for foreign exchange market and discussed the scale-free structures found out in the networks made [5]. Correlation matrices of stock returns over time in New York Stock Exchange were analyzed using spectral and network methods in a research in by Heimo et al. [5]. In a study of Hank Seng stock market of Hong Kong, Li and Wang extracted the hidden fluctuation patterns of the stock index from a directed network topology. They used betweenness and inverse participation ratio of the nodes of the network to analyze the fluctuations of the stocks [6]. A review of the
literature on small-world networks used in management and social science was done by Uzzi et al. where they showed different interdisciplinary applications of small-world networks as previously discussed by Milgram in other fields of science [7].

In 2008, Yang & Yang presented a reliable procedure to build networks from correlation matrix of different time series. They used the correlations between time series to build adjacency matrix based on different thresholds [8]. Kwon and Yang used transfer entropy to show direction and strength of information flow between stock indices time series [9]. It was a new way of building directed networks out of time series. As it will be seen later in this study, using statistical correlation can only make undirected networks. In this study, a new method of building directed networks out of financial data has been used by finding association rules between the fluctuation patterns to extract a directed network.

Yang et al. did a research in 2009 to investigate six exchange rate time series by means of a visibility graph [10]. A visibility graph is a network made by all of time series data as nodes and edges between any two nodes that can be seen with a direct line in the time series graph. By this mean they constructed a network for every financial time series and analyzed the power law degree distribution and scale-free topology of exchange rate time series. Huang et al. studied Chinese stock market in another work and represented the stock market data as a complex network [11]. They also studied the scale-freeness of the network and centrality measurements of the nodes and cliques in the topology.

In 2010, Tse et al. used complex networks to study correlations between prices of all US stock markets [12]. In their networks, the nodes represented stocks. In another research, Materassi and Innocenti, tried to solve the problem of reconstructing the tree-like structure of a network for linear dynamic systems [13]. They used a distance function to calculate the closeness between processes. Zhang et al. did a study to analyze the time series of Shanghai stock index with the use of complex network theory. They showed that the network of the main series is fitted with a power law, and the network extracted from the return series is fitted by an exponential curve [14]. Tabak et al. investigated the Brazilian stock market sectors with building the minimum spanning tree [15]. By network measurement tools, they showed that energy, finance,
and material sectors of the market were the most important ones. A part of this study has used the sector indices of Tehran stock market to build a network for understanding the relationship between different sectors and finding the most important sectors of the Tehran stock market. In another research, Jiang and Zhou, investigated complex network of stock trading data among investors of Shenzen Development Bank [16]. The nodes were stock traders and the links presented the trade with the weight as the volume of it. They made a network for each trading day and showed that networks present a power law degree distribution.

In 2011, Namaki et al. did a study on financial markets using random matrix theory [17]. They analyzed the clustering coefficients and component numbers of the networks. The data of Dow Jones Industrial Average (DJIA) and Tehran Stock Exchange (TSE) were used in their study. Ma et al. did a study and established networks of cross-shareholding for some companies in China in a period of time and analyzed the networks [18]. They studied the topology of cross-shareholding networks in an 8 year period and discussed the differences of the networks before and after the financial crisis in 2008. In another research by Sun et al. a full transaction records of more than hundred stocks were used to build trading networks where nodes represented the investors and links connected sellers to buyers [19]. They showed that degree distribution of these networks obeyed the power law and manipulated stocks can be distinguished from non-manipulated ones by a high lower band of the power law tail and high average degree. In another study by Allali et al., directed network of ten important world’s financial markets were made by use of partially directed coherence [20]. Partially directed coherence was used before in neuroscience studies to find out the causality between different processes.

In 2012, a research was done by Wang and Wang, to study the visibility graph network of four macro-economic time series of China [21]. Similar to their previous study, by the means of visibility graph they tried to capture new features from these time series and study the differences of the network structure before and after some economic policies of China’s government. They studied the small-network effect in the visibility graph of these time series. Chunxia et al. studied the relation between the variations of the structure and fluctuations of the Shanghai stock market [22]. They used a moving window to scan through the stock prices time series for a period of time. Caraiani studied
the properties of returns of the stock markets from Europe with the help of complex networks. He analyzed the properties of the networks extracted by means of the centrality measurement parameters [23]. Ko et al. studied the correlation network between two important stock markets of Korea and compared the topology of the networks constructed before and after the financial crisis in 2008 [24]. Another research by Farmer et al. emphasized the importance of studying economy and financial markets with the help of network science [25]. They talked about importance of interdisciplinary studies in complex networks, economy, and finance and the effects of these studies on science, technology and society.

In 2013, Caetano and Yoneyama studied the sudden changes in direction of stock market index [26]. They made a new indicator using wavelet decomposition and used the correlation network with nodes as stocks and links as correlations. They used a combinational method of eigenvalues of adjacency matrix and their indicator to find out the points where the stock index changes direction. Roy and Sarkar, did a research on 93 different stock markets by making the correlation network between them before and after the collapse of Lehman-Brothers [27]. They studied the minimal spanning tree of the networks and analyzed them with an index called turbulence of the market that was calculated by eigenvector centrality measurement of the nodes in the networks. Sensoy et al. also studied the correlation networks of different stock markets across the world before and after the financial crisis of 2008 [28]. They tried to find out which markets were more important in the networks by the help of centrality measurements. Liao and Chou, used association rules and K-means clustering to make a good portfolio of stocks between the China, Taiwan, and Honk Kong stock markets [29]. They didn’t use the network approach in their study but their work was important for this research due to using association rules to understand the relationship between different stocks. Hu et al. divided the China into 31 including Hong Kong and constructed a correlation network with respect to GDP of these regions [30]. They showed that the location and distance of the regions to each other plays an important role in connection between the nodes in the network. Park and Shin, tried to predict the fluctuations of the market using a semi-supervised algorithm on a network of different stock markets, exchange rates, oil price and some other financial time series [31]. They claimed that their method could see interactions and cyclic effects of markets on each other.
In 2014, Lim et al. did a research on relationship between credit market and stock market before and after subprime crisis [32]. They used the network topology and random matrix theory and compared the eigenvalue of the network matrixes. It was found out that the eigenvalue of credit market became bigger than the one for stock market right before the crisis happened and after the crisis, the correlations between two markets became stronger than before. Another research by Castren and Rancan, tried to construct networks of important enterprises and firms in euro region for both local regions and the whole euro region [33]. They studied the propagation of shocks from a node to other nodes in these networks with the help of entropy matrix made from adjacency matrix. Diebold and Yilmaz focused on Lehman-Brothers collapse time period and made networks of different firms by their stock price time series with the help of variance decomposition [34]. They constructed directed networks and analyzed the changes of their topology in the crisis time period. In another study by Yang et al. co-integration coefficient was used to make directed network of stock markets before and after financial crisis and collapse of Lehman-Brothers [35]. They found out that the impact of US stock market on other markets has reduced after the crisis and Chinas markets impact has increased. Similar to a part of the work done in this research, Mai et al. constructed a correlation network of Chinas market sectors and showed that the degree distribution of the network obeyed power law with little exponent [36]. They found out the scale-free topology of the network and said that Industry sector had more impact on other sectors. In this research we show what sector has more impact on others in TSE market in both undirected correlation networks and directed Apriori networks made for understanding the fluctuations of the indices.

In the next section of the paper, we introduce our methodology of extracting directed and undirected networks out of stock exchange sector indices data. We analyze the networks by centrality measurement parameters and show the most important sectors that have more impact on others in TSE. Then we introduce a way of predicting fluctuations of the sectors stock indices by finding the paths in the directed network extracted from the association rules by Apriori algorithm. Section three is about the data gathering and preparation phase of this research. Section four talks about the evaluation methodology used to test our approach of predicting fluctuations of the indices time series. In section five the results are presented and discussed in detail and the last section
concludes with some suggestions for further researches. The abbreviations of the sector names in TSE market that we used are presented in a table in appendix 1 and the pseudo-code of the evaluation methodology algorithm for our fluctuation prediction test is presented in appendix 2.

2. Methodology

The methodology used in this paper consists of extracting networks of stock sector price indices out of TSE dataset, in order to find the important sectors from the networks in different aspects. It also provides a way to predict rising fluctuations of the price indices over time from the directed networks made by association rules and Apriori algorithm. At the end of this section a diagram of the whole research methodology is presented and explained.

2.1. Network Extraction Methods

In this study, first a correlation matrix between the sector indices time series is extracted as the adjacency matrix of the networks. By this method, and choosing a threshold it can be assumed that there is a link between sectors that have cross correlations bigger than the threshold. Two different undirected networks with two different thresholds for the correlation values between the sector indices data of TSE are extracted as an example to show how the proposed methods work in the TSE market. The correlation formula is shown in Eq. (2.1) [27].

\[
r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}
\]

(2.1)

In the correlation equation \(x\) and \(y\) are the two different sector indices time series and \(r\) is the correlation between them. The correlation, \(r\), is a digit between -1 and +1. Two different thresholds of 0.7 and 0.9 have been used to check if \(r\) is greater than the thresholds or not. Any other thresholds could be used to make different networks. In this study using these thresholds caused more than half of the sectors appear in the networks (In the threshold of 0.7, 36 out of 38 sectors appeared in the network). Due to one of the aims of this study to find out important
sectors with more positive correlations with other sectors, the disappearance of some sectors in the network did not affect the results. It is obvious that any other threshold cuts could be used to make other networks. This study had to choose some threshold cuts to show numerical results of its methodology used in practice. It is clear that by using the threshold of -1 a complete graph is constructed and no knowledge of the market can be extracted out of it (The case of using a complete weighted correlation graph is different and is discussed in other researches explained in the introduction of this study). If the value of $r$ is more than the threshold, then the corresponding element in the adjacency matrix of the network will be 1 otherwise 0. This means there will be a link between the nodes of two stock sector indices in the network. As mentioned, with lower thresholds the network will have more links and a link between two nodes shows positive correlation between the indices higher than the threshold. In other word, when a sector index rises, the other sector indices linked to it in the network will rise with more probability. Once the two different correlation networks corresponding to two thresholds have been extracted, various centrality measurements are calculated on the networks in order to find the most important sectors in the TSE Market.

In the other phase of the research, a directed network is built out of the stock sector indices with the help of association rules and Apriori algorithm. First, each series are converted to binary series that only consists of zeros and ones. A rise in the value of the stock sector index from the previous value in the time series is represented by a 1, and a 0 indicates a fall in the value. After converting all the indices time series into these new series that show fluctuations, all these series are put as the columns of a new matrix. Each column of this matrix is a binary fluctuation time series of a different sector index and each row is a day in the stock market. Figure 1 presents a sample of converting the data for using in Apriori algorithm as described before for a 4 day period of time. It is obvious that after the conversion the new dataset will have one day less than the original dataset.
By using an algorithm of finding association rules in this matrix and looking at each row like a buying transaction in a store, all the couple indices that rise together are found. Rules like \( A \rightarrow B \) that indicate if sector A, rises with a specific confidence and support value, the sector index B also rises are extracted. The Apriori algorithm is used to find out rules like this in the TSE market data. A brief description of the Apriori algorithm that is used in this research is presented in section 2.3.

Every association rule like \( A \rightarrow B \) in the converted matrix means a directed link between the nodes A and B. Two different confidence values of 0.60 and 0.65 and a support of 0.1 are used to make two different directed networks for the TSE market. Again it should be explained that any other confidence and support can be used to make different directed networks. Using these thresholds for confidence and support made two networks with 36% (for 0.65 confidence level) and 84% (for 0.60 confidence level) of the sectors appear in the directed network. Using lesser confidence levels reveals lesser knowledge from the topology of the network and by using a confidence of zero a complete graph would have appeared. The centrality measures are calculated again for the two new networks and are also presented and discussed later.

2.2. Market Sector Analysis

Finding the most important sectors of the TSE market is done by centrality measure analysis of the nodes in the extracted networks. Centrality measurement is a good tool to find out which nodes are more important in the network [37]. In this work different centrality measures
are used that are calculated for all the nodes of the networks that correspond to the market sectors.

Degree centrality calculates all the links that start or end from a node. If the links coming to a node are counted, it is called in-degree and if the links going out of the node are counted, it is called out-degree. The closeness centrality measures the mean distance from a node to other nodes. Another centrality measure used in this study is called betweenness. It measures the extent to which a node lies on paths between other couple of nodes. Eigenvector centrality and page rank are extensions to degree centrality. Not all the nodes have the same importance and having link to some nodes are more important than others. Eigenvector centrality and page rank are increased if having links to other important nodes [37, 38].

Network constraint is another measure that shows the extent to which a node links to other nodes that are already linked to each other. Betweenness and network constraint, both try to find bridges in the network topology. Lower network constraint and higher betweenness indicate bridging [37, 38]. This means that in the results section, the nodes with the lowest network constraint or with the highest betweenness or degree are presented as important nodes (sectors in our case).

2.3. Association Rules and Apriori Algorithm

Association rules, present the relationship between different item sets in terms of occurrence. This means, if some items appear in a transaction (a record, or a row of matrix in our case), it can be assumed that some other items will also appear in the same transaction. Apriori is the name of a famous algorithm to find association rules in a dataset (The fluctuation matrix that was described in section 2.1. in our case).

Apriori algorithm was presented by Agrawal and Srikant in 1994. They provided an algorithm for finding association rules in large database of sales transactions. The name of the algorithm comes from the fact that it uses prior knowledge to find frequent item sets in the database. Any given sequence of the items in the database is called an item set. The algorithm creates some candidate item sets with \( k \) items, which is shown as \( C_k \). Those candidate itemsets that have repeated more than a
proportion called support are frequent item sets. Frequent item sets with
the length k are shown as $L_k$ [39].
In the first step of Apriori algorithm, the $C_1$ item sets are gathered and
$L_1$ is obtained from all $C_1$ item sets which are frequent (are repeated
more than a predetermined number called support). Other steps of the
algorithm are presented below:

- $L_{k-1}$ is obtained.
- $C_k$ is obtained from Cartesian product of $L_{k-1} \times L_{k-1}$.
- All $C_k$ that have sub-itemsets which are not frequent can not be
frequent themselves.
- $L_k$ is obtained.

In this paper, after finding the $L_2$ itemsets, our work is finished with the
Apriori algorithm. This is due to the nature of the rules that are going to
be used in the proposed network extraction method (As mentioned
before, rules like $A \rightarrow B$ are used that only consist of two items A and
B). After finding $L_2$ item sets, all of $A \rightarrow B$ rules that have confidence
level more than a predetermined threshold are extracted to use in
directed network building process. The confidence and support in the
Apriori algorithm are defined in Eq. 2.3.1 and 2.3.2.

$$support(A) = \text{proportion of itemset } A \text{ in the dataset} \quad (2.3.1)$$

$$Confidence(A \rightarrow B) = \frac{\text{Support}(A \cup B)}{\text{Support}(A)} \quad (2.3.2)$$

2.4. Fluctuation Prediction with Directed Apriori Networks

After making the adjacency matrix of directed networks (having one in
element $(i, j)$ of the matrix if there is a rule $i \rightarrow j$, otherwise zero), the
paths of length n can be found, by multiplying the matrix to itself, n
times. If the adjacency matrix is called A, the element $(i, j)$ of matrix $A^n$
presents how many paths of length n are from i, to j.

As discussed before in section 2.1, a directed link from node i to node j
shows that by a confidence threshold, if the market sector index i rises,
the market sector index j rises in the same day (day or any other time
unit that is used). Now if there is a directed path of length 2, from i to j,
and then from j to k, the fluctuation of i propagates to k through the node
Suppose this hypothesis is going to be tested that the probability of market sector index $k$ rising in the next day would be bigger than fifty percent, when there is a path of length 2, from $i$ to $k$. The reason of having this hypothesis is that, without any prior knowledge there is equal chance for rising or falling of the market sector index of $k$. But with the prior knowledge of having a path from $i$ to $j$ and $j$ to $k$, it can be assumed that $k$ rising in the next day may be more probable. The evaluation process in this study shows the truth of this hypothesis in case of directed sector index market of TSE. For paths of length more than 2, the test methodology and evaluation process is extended and the results are also presented.

### 2.5. The General Diagram of the Research

The overall research process diagram is presented in figure 2. All the boxes (sub-processes) have been described in detail in this paper. The boxes (sub-processes) in the diagram are labeled and also a brief description is given in this section to give a general idea of the whole process. These sub-processes are described in details in their corresponding section.

**Figure 2: Diagram of the Research**

Here is a brief description of the sub-processes in order of their box number:

1. Historical data collection for stock market sectors (described in section 3).
2. Normalizing the data (described in section 3).
3. Calculating the correlation between stock sector indices (described in section 2.1).
4. Extracting correlation networks from connecting the sectors with correlation higher than a threshold (described in section 2.1).
5. Converting the stock sector indices into fluctuation time series (described in section 2.1 and in figure 1).
6. Applying Apriori algorithm on the rows of the fluctuation series matrix (having each day as a transaction like figure 1) for finding rules like A→B (described in section 2.3).
7. Extracting directed networks by connecting items (sectors) of left and right side of the rules (described in section 2.1).
8. Calculating network parameters and centrality measures for market sectors (Nodes of the Networks) for each network and sorting the results to find the most important nodes or sectors in different aspects in the market (described in section 2.2 and results presented in section 5).
9. Comparing and discussing centrality measurement of market sectors for correlation and Apriori networks (presented in section 5).
10. Testing fluctuation prediction for \( n \) days later of market sector indices with finding paths of length \( n+1 \) in the directed Apriori network (described in section 4 and results presented in section 5).

In this phase, the hypothesis of ability to predict fluctuations with directed network is tested in three different ways:

- Using all the paths of length \( n \) between two nodes as the input of testing algorithm to evaluate our hypothesis.
- Using all the paths of length \( n \) between two nodes apart from those that there is also a path of length one between the source and destination nodes.
- Using 80 percent of the data to build an Apriori network, and using the other 20 percent for testing the hypothesis.

The evaluation process or box number 10 is explained in more details in section 4.
3. Data Preparation

Data for market sector price indices are gathered from the official site of TSE [40]. Price indices for every industrial, commercial, or service sector are calculated every day in TSE. Data gathered are for the period of 11th October 2009 to 11th May 2013. Considering the missing data (national holidays), 860 daily stock sector index data were gathered for 38 market sectors out of 43 market sectors. Five market sectors did not have enough data in the official site of TSE in the specified time period.

One can consider the dataset as a matrix with 860 rows for each day that data was available and 38 columns for each market sector index. Because of the different value region of the indices, this dataset was normalized with formula (3.1).

\[ \text{Normalized Value} = \frac{\text{Original Value} - \text{Minimum Value in the Index Time Series}}{\text{Maximum Value in the Index Time Series} - \text{Minimum Value in the Index Time Series}} \] (3.1)

Due to the long names of market sectors in TSE, abbreviations were used for each of the official names of the sectors. These abbreviations are presented in appendix 1.

4. Evaluation Method

Considering the directed networks made by the Apriori algorithm, a hypothesis is proposed along with an evaluation method to test it. The evaluation method is described as below:

- Hypothesis: If there is a path of length \( n \) between nodes A and B, in the directed networks, this means that if the value of index A rises, the value of index B rises \( n-1 \) days later. It is clear that \( n > 1 \).
- Evaluation Method Steps:

  I. Two variables of \( T \) (True) and \( F \) (False) are defined and set to zero. These variables count the number of true and false predictions of the fluctuation by our hypothesis in the directed network.

  II. All the directed paths of length \( n \) in the network are found. For this purpose, the adjacency matrix \( A \) will be multiplied to itself, \( n \) times.
Any matrix element that is 1 or more than 1 shows there is 1 or more than 1 path of length $n$ between the two nodes.

III. Wherever there is a path of length $n$ between $A$ to $B$, the fluctuation dataset matrix (figure 1) receives our attention. Anytime corresponding column for $A$ has the value 1 (meaning that $A$ is rising), column $B$, $n-1$ rows ahead value is monitored. If this value is also 1, the variable $T$ is increased one unit for a true prediction of fluctuation. If the value of the $n-1$ rows ahead of column $B$ is 0 (meaning that despite the value of $A$ that was rising $n-1$ days before, the value of $B$ does not rise), the variable $F$ is increased one unit as a false prediction.

IV. Finally, for each path of length $n$ from any node $A$ to the any node $B$, the hypothesis is tested, and the results of the tests for every path will be put in a test vector (the vector length will be the number of paths with length $n$). If more than 50 percent of the results in the test vector are 1, this means that directed network and our hypothesis for the fluctuation prediction can be used and is better than using a random classifier. From the test vector, the value of $T/(T+F)$ is calculated. This value presents the recall of the prediction model. Recall formula is presented in Eq. 4.1.

The pseudo-code of the evaluation methodology is presented in appendix 2.

\[
\text{Results from test vectors for each path length} = \frac{T}{T+F} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{4.1}
\]

5. Results and Discussion

In this study, two open-source applications are used. Graphviz (Graph Visualization Software) is used to depict the networks, and SNAP (Stanford Network Analysis Platform) is used to calculate various network parameters such as centrality and clustering. Correlation networks were extracted with two different correlation thresholds of 0.7
Analysis and Prediction of Fluctuations for Sector Price Indices with Cross Correlation and Association Based Networks: Tehran Stock Exchange Case

and 0.9. The network parameters of these correlation networks are presented in table 1.

Table 1: Network Parameters for Correlation Networks Made by Two Different Thresholds.

<table>
<thead>
<tr>
<th>Model</th>
<th>Number of Nodes</th>
<th>Number of Links</th>
</tr>
</thead>
<tbody>
<tr>
<td>Threshold = 0.7</td>
<td>36</td>
<td>251</td>
</tr>
<tr>
<td>Threshold = 0.9</td>
<td>21</td>
<td>41</td>
</tr>
</tbody>
</table>

As it can be seen in table 1, with reducing the correlation threshold, more nodes (market sectors) will have the privilege of being in the network and the number of the links increases. It is obvious that with a threshold of -1, the network becomes a complete graph. Figure 3 and 4 present the visualization of the correlation networks with thresholds of 0.7 and 0.9.

Figure 3: Correlation network of Tehran Stock Exchange Market Sectors with 70% Threshold.
Figure 4: Correlation Network of Tehran Stock Exchange Market Sectors with 90% Threshold.

As explained before, the Apriori networks are extracted with support level of 0.10 and confidence levels of 0.60 and 0.65. The Apriori network parameters are presented in table 2 and the visualization of the networks are shown in figures 5 and 6.

Table 2: Network Parameters for Apriori Networks Made by two Different Confidence Levels.

<table>
<thead>
<tr>
<th>Model</th>
<th>Number of Nodes</th>
<th>Number of Links</th>
<th>Nodes with Zero In-Degree</th>
<th>Nodes with Zero Out-Degree</th>
<th>Bi-Directional Links</th>
</tr>
</thead>
<tbody>
<tr>
<td>Confidence ≥ 0.60</td>
<td>32</td>
<td>189</td>
<td>14</td>
<td>1</td>
<td>80</td>
</tr>
<tr>
<td>Confidence ≥ 0.65</td>
<td>14</td>
<td>46</td>
<td>3</td>
<td>1</td>
<td>18</td>
</tr>
</tbody>
</table>
Analysis and Prediction of Fluctuations for Sector Price Indices with Cross Correlation and Association Based Networks: Tehran Stock Exchange Case

**Figure 5:** Apriori Network of Tehran Stock Exchange Market Sectors with Confidence Level of 60% or Higher.

**Figure 6:** Apriori Network of Tehran Stock Exchange Market Sectors with Confidence Level of 65% or Higher.
In the networks extracted with Apriori algorithm, the more the confidence level is raised, the less are the links and nodes. This is due to extracting less association rules with higher confidences. The results of centrality measurement data sorted and calculated for two correlation networks and two Apriori networks are presented in table 3. These data are extracted after sorting the nodes according their importance in the centrality and the most important node (market sector) is presented in table 3. The concept of importance in different centrality measurement methods has been discussed in section 2.2. The results are presented in abbreviation form and the full name of the market sectors in TSE are presented in the appendix 1.

Table 3: Important Sectors with Consideration of Network Centrality Measures

<table>
<thead>
<tr>
<th>Model</th>
<th>Degree</th>
<th>Closeness</th>
<th>Betweenness</th>
<th>Eigen Vector</th>
<th>Network Constraint</th>
<th>Clustering Coefficient</th>
<th>Page Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correlation network with threshold = 0.7</td>
<td>Industries</td>
<td>Industries</td>
<td>Finance</td>
<td>Industries</td>
<td>RealState</td>
<td>Car-Electrical Medical</td>
<td>Petrochemistry</td>
</tr>
<tr>
<td>Correlation network with threshold = 0.9</td>
<td>Industries</td>
<td>Industries</td>
<td>Industries</td>
<td>Cement</td>
<td>RealState</td>
<td>IT</td>
<td>Metal</td>
</tr>
<tr>
<td>Apriori network Confidence ≥ 0.60</td>
<td>Car</td>
<td>Car</td>
<td>Car</td>
<td>Car</td>
<td>Car</td>
<td>ManufactICT Paper-Ceramic</td>
<td>EngServices</td>
</tr>
<tr>
<td>Apriori network Confidence ≥ 0.65</td>
<td>Car</td>
<td>Car</td>
<td>Car</td>
<td>Car</td>
<td>Car</td>
<td>Industries</td>
<td>Investors</td>
</tr>
</tbody>
</table>

By looking at the table 3 it can be seen that in correlation networks the industrial companies (see appendix 1) sector shows more correlation with other sector indices. This results that if some could only see one sector to analyze the fluctuations of the Tehran Stock Exchange Index
(TEPIX), the industrial companies’ sector index showed a better behavior of the TEPIX than other sector indices.

The results for the models made by Apriori algorithm on fluctuation time series presents a fact about the vehicle and parts manufacturing (abbreviated as car in our tables and figures) sector. This sector is affected by many of other sectors, meaning where they rise in value, the vehicle and parts manufacturing rises also in the same day. From Apriori networks the paths can also be found out from a sector to another sector to understand the effects of a rise in a sector price index to other indices.

In table 4, the results of testing the fluctuation prediction hypothesis are presented. The values in the brackets show the second type of testing by eliminating all paths from \( A \) to \( B \) that also had a single link from \( A \) to \( B \). In table 4 results show a high success except for paths of length 3 for networks of confidence level 60% (prediction of 2 days later fluctuation).

**Table 4:** Results of Testing the Prediction Hypothesis for all the Dataset with and Without Eliminating Paths that Have a Single Link Between Two Nodes. (Those Results Which Come from Elimination are in Brackets).

<table>
<thead>
<tr>
<th>Model</th>
<th>Model Number</th>
<th>Number of Tests for Paths with length 2</th>
<th>Percentage of Successful Predictions</th>
<th>Number of Tests for Paths with length 3</th>
<th>Percentage of Successful Predictions</th>
<th>Number of Tests for Paths with length 4</th>
<th>Percentage of Successful Predictions</th>
<th>Number of Tests for Paths with length 5</th>
<th>Percentage of Successful Predictions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apriori network Confidence ≥ 0.65</td>
<td>85 (46)</td>
<td>98.8% (97.8)%</td>
<td>110 (66)</td>
<td>50% (48.5)%</td>
<td>117 (73)</td>
<td>60.4% (58.9)%</td>
<td>117 (73)</td>
<td>75.2% (71.2)%</td>
<td></td>
</tr>
<tr>
<td>Apriori network Confidence ≥ 0.60</td>
<td>359 (188)</td>
<td>89.8% (87.7)%</td>
<td>448 (266)</td>
<td>45.8% (49.6)%</td>
<td>459 (277)</td>
<td>60.4% (51.3)%</td>
<td>459 (277)</td>
<td>61.2% (56.3)%</td>
<td></td>
</tr>
</tbody>
</table>

6. Conclusion and Further Research

In this study concepts of network science have been used to come up with a clearer understanding of the TSE market and relationship between its industrial and service sectors. The researchers also presented a network model using Apriori algorithm to predict the fluctuations of sector markets. It is found out that the industrial companies sector has
the highest correlation with other sectors and the correlational structure of the TSE was extracted. By directed networks extracted from association rules, it became clear that the vehicle and parts manufacturing sector was the most influenced sector by other sectors price fluctuation in Tehran Stock Market.

Finally, the fluctuation prediction model was tested and it was shown that this model can be used with directed Apriori networks built upon high confidences to predict the fluctuations and it works better on short-time predictions.

For fluctuation prediction out of the fluctuation dataset, other association rule mining algorithms and sequential pattern mining algorithms can be studied to see how they result. Also, other datasets of other markets can be the focus of future researches in predicting and analyzing the financial markets with our methodology of network construction.
### Appendix 1: Table of Abbreviations Used for the Market Sector Names Used in the Results

<table>
<thead>
<tr>
<th>Full Name of the Market Sector</th>
<th>Abbreviation</th>
<th>Full Name of the Market Sector</th>
<th>Abbreviation</th>
<th>Full Name of the Market Sector</th>
<th>Abbreviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Insurance and Pension Funds</td>
<td>Insurance</td>
<td>Chemical Products</td>
<td>Chemicals</td>
<td>IT and Related Activities</td>
<td>IT</td>
</tr>
<tr>
<td>except Social Security</td>
<td>Oil</td>
<td>Pharma</td>
<td>Pharma</td>
<td>Medical, Optical</td>
<td>Medical</td>
</tr>
<tr>
<td>Oil and Gas Extraction and</td>
<td>ICT</td>
<td>Tanning, Polishing Leather</td>
<td>Leather</td>
<td>and Measuring Instruments</td>
<td>Agriculture</td>
</tr>
<tr>
<td>Ancillary Services Except</td>
<td>Ceramic</td>
<td>and Footwear Manufacturing</td>
<td>RealEstate</td>
<td>Manufacturing</td>
<td>Furniture</td>
</tr>
<tr>
<td>Exploration</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Information and Communication</td>
<td>Investors</td>
<td>Mass Construction and Real</td>
<td>Industries</td>
<td>Furniture Manufacturing</td>
<td>Construction</td>
</tr>
<tr>
<td>Ceramic and Tile</td>
<td>Cement</td>
<td>Companies</td>
<td>EngServices</td>
<td>Industrial Contractors</td>
<td>MetalExtract</td>
</tr>
<tr>
<td>Investors</td>
<td>NonMetal</td>
<td>Engineering Services</td>
<td>Coal</td>
<td>Metal Extraction</td>
<td>Finance</td>
</tr>
<tr>
<td>Cement, Lime and Gypsum</td>
<td>Food</td>
<td>Mining of Coal</td>
<td>Transport</td>
<td>Financial Intermediation</td>
<td>Textile</td>
</tr>
<tr>
<td>Other non-Metallic Mineral</td>
<td>Brokers</td>
<td>Transportation and Storage</td>
<td>Plastic</td>
<td>and Monetary</td>
<td>ManufactICT</td>
</tr>
<tr>
<td>Products</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Food and Drink Products except</td>
<td>Sugar</td>
<td>Rubber and Plastic</td>
<td>Wood</td>
<td>Textile Industry</td>
<td>Mine</td>
</tr>
<tr>
<td>Sugar</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other Financial Brokerages</td>
<td>Paper</td>
<td>Wood Industries</td>
<td>Metal</td>
<td>ICT Manufacturing</td>
<td>MetalManufact</td>
</tr>
<tr>
<td>Sugar</td>
<td>Car</td>
<td>Manufacture of Basic Metals</td>
<td>Publish</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Paper Products</td>
<td>Machinery</td>
<td>Publishing and Printing</td>
<td>Petrochemistry</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Appendix 2: Fluctuation Prediction Hypothesis Evaluation Algorithm in Pseudo-Code

\( A \) = Read data of the adjacency matrix for Apriori network

\( \text{dir} = \) Read fluctuation dataset (the data set which made in a process explained in section 2.1 and figure 1).

\( \text{Table}_{-}\_\text{of}_{-}\_\text{Results} = \text{cell}(4, 1) \)

// n is the length of the paths, in our pseudo code n will be between 2 and 5, for prediction of 1 day to 4 days later

for \( n = 2 \) to 5

//Make the matrix of paths by length n

\( \text{Adj}_{-}n = A^n \)

//number of predictions in the matrix of paths of length two

\( \text{num}_{-}\_\text{of}_{-}\_\text{preds} = 0 \)

\( \text{TEST} = 0; \% \text{result of the Test} \)

//how many days after today you think the price rises

//\( \text{lead} = n - 1 \), due to the testing paths with length n

\( \text{lead} = n - 1 \)

for \( i = 1 \) to length(\( \text{Adj}_{-}n(:, 1) \))

for \( j = 1 \) to length(\( \text{Adj}_{-}n(1, :) \))

if \( \text{Adj}_{-}n(i, j) \geq 1 \)

\( \text{num}_{-}\_\text{of}_{-}\_\text{preds} = \text{num}_{-}\_\text{of}_{-}\_\text{preds} + 1 \)

//number of true & false predictions for a path with length \( n \)

\( T = 0, F = 0 \)

for day=1 to length(\( \text{dir}(, 1) \))- lead

if (\( \text{dir}(\text{day}, i) == 1 \)) AND (\( \text{dir}(\text{day} + \text{lead}, j) == 1 \))

\( T = T + 1 \)

elseif (\( \text{dir}(\text{day}, i) == 1 \)) AND (\( \text{dir}(\text{day} + \text{lead}, j) == 0 \))

\( F = F + 1 \)

end

end

if \( \frac{T}{T+F} > 0.5 \)

\( \text{hitrate} = 1 \) //percentage of right predictions
else  
    hitrate = 0
end

TEST(num_of_preds) = hitrate
end
end

Table_of_Results[n - 1, 1] = TEST
end
References


