International Journal of Industrial Engineering & Management Science, Vol. 7, Issue 1, (2020) 100-112

University of Hormozgan

International Journal of Industrial Engineering & Management Science

Journal homepage: www.ijiems.com

Application of community detection in stock market analysis

Farideh Rahimnezhad[a], Seyed Kamal Chaharsooghi [a,*], Mehdi Seyfi Sariqaya[a], Mohammad Sheikhi[b]

[a] *Tarbiat Modares University*
[b] *Amirkabir University of Technology*

Keywords:

[1] Community detection
[2] Complex network
[3] Modularity
[4] Stock market
[5] Stock return correlation

A B S T R A C T

Nowadays, in accordance with the recent research, the stock market has a lot of data that is constantly fluctuating. Data analyzing and separating them into distinct groups is time-consuming and difficult for investors and managers. In this study, techniques of community detection are used to simplify stock market analysis. For this purpose, Tehran stock exchange has been selected and after collecting the latest price of transaction data, a correlation network of stock return has been developed. Then, using community detection, cohort companies were identified and it became clear that the stock of the same industry belonged to a common group. Therefore, it can be said that the correlation between stock prices of companies largely depends on the industry in which these companies are active. Hence, it can be concluded that the community detection technique works completely logically and its application will facilitate and accelerate the analysis of stock market data.

1. Introduction

Complex networks are very simple and useful tools for displaying the characteristics of many actual complex systems (Chaharsooghi & Rahimnezhad, 2016). In fact, the theory of complex networks essentially comes back to discrete mathematics and graph theory. The difference between networks and the graphs is in the amount of more available information that they transmit to the nodes and edges of a network. Each network has a graph called the network's graph, but the vice versa of that does not exist, it means that each graph does not necessarily belong to a network in the real world (Kito & Ueda, 2014). In a study in 2015, a network of 1065 stock returns of Hong Kong Stock Market from 2011 to 2015 was designed in which stocks were considered as nodes and cross-correlations between stock returns as edges

* Corresponding author, skch@modares.ac.ir

and different threshold values for comparison of the network, characteristics were considered. In Figure 1, a view of the network is shown with a threshold value of 0.3 (Huang et al., 2015).

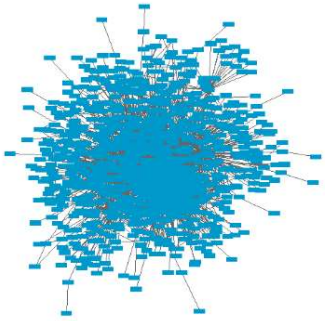

Fig.1. Hong Kong stock network with 0.3 threshold (Huang et al., 2015)

Boginski and his coauthors introduced complete graphs and independent collections in stock networks and provided a technical method for stock clustering (Boginski, Butenko & Pardalos, 2005). Tumminello and his colleagues had a statistical evaluation on the networks and determined the clusters of investors in the financial market. They reviewed specific data on the business activities of Nokia's stockholders and found that the identified clusters of investors in the period of their decision making, had a high degree of symmetry, which the issue has various reasons, including adopting similar strategies by investors, exchanging information among investors, etc. (Tumminello et al., 2012). The resulting network from the design of the relationship between the investors and the final clustering is shown in Figures 2 and 3.

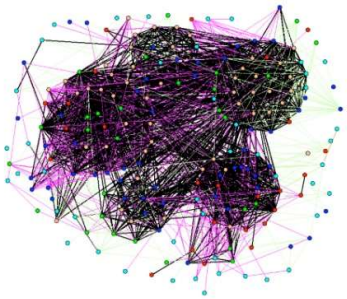

Fig.2. Investors network **(**Tomminello et al., 2012**)**

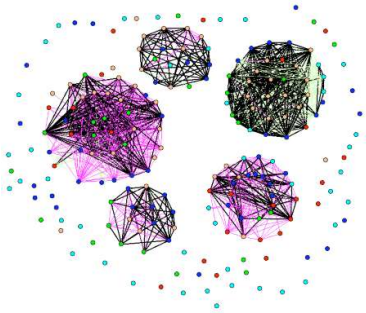

Fig.3. Investors network clustering **(**Tomminello et al., 2012**)**

1.2. Community detecting

Community detection is one of the most popular methods for grouping nodes in the network according to the network structure (network communication), which has many real-world uses from the discovery of communities in bio-networks to community detection on Internet sites, scientific and research (He et al., 2018**)**. In a network, nodes in a community have common characteristics (Gao et al., 2018). Large network grouping by identifying smaller communities, makes it possible to use a lower cost and energy to studied, monitored, and managed each community separately with more detail or the entire community at a higher level with less detail. for example, in terms of a particular criterion (Bhatia & Rani, 2019). Usually, real-world networks have a hierarchical structure. Small communities are in the midst of large communities and large communities themselves are in the midst of larger communities Fortunato, S. (2010). Community detection also has other indirect benefits. For example, after identifying the community and their boundaries, they can be used to determine the position of nodes in each community and consequently, to classify them according to their place. For example, nodes which located within the boundaries of communities (or nodes belonging to several communities) can make a large contribution to inter-community exchanges (Veldt, Gleich & Wirth, 2018).

1.2.1. Categorization of community detection algorithms

One of the first presented categorizations for communities detecting approaches in Porter et al research is seen (Porter, Onnela & Mucha, 2009). Although other categories are also provided (Papadopoulos et al., 2009). One simple way of determining communities in a network is to identify and remove the edges that connect the nodes of different communities. This technique is the main philosophy of divisive algorithms. One of the most popular divisives algorithms is presented by Girvan and Newman (Newman & Girvan, 2004). Dynamic algorithms are also a set of community detection methods in which a repeatable convergence procedure takes place, in which each part of the answer varies. Among the major models in this field, we can mention spin models (Son, Jeong & Noh, 2006) and random walk (Hu et al., 2008).

2. Research model

In the stock market, stock price fluctuations are not independent of each other, and there is a complex relationship between them. Calculation of the correlation between the stock prices can be a way of showing these connections. Subsequently, the model for constructing a stock return correlation network will be described. Then, the model will be implemented for extracting and analyzing the communities of the Tehran Stock Exchange.

2.1. Stock return correlation network model

The correlation network of stock return is constructed as follows:

If $P_i(t)$ denotes the price of the stock I on day t, then, the return of the stock price i in a one-day period from t-1 to t will be in accordance with Equation (1) (Shariatmadari, 2014):

(1)

$$R_i(t) = \ln\left(\frac{P_i(t)}{P_i(t-1)}\right)$$

The correlation coefficient between the price of companies i and j is obtained from Equation (2) according to Pearson's correlation coefficient formula (Shariatmadari, 2014):

$$C_{ij} = \frac{\langle R_i R_j \rangle - \langle R_i \rangle \langle R_j \rangle}{\sqrt{\left\langle R_i^2 - \langle R_i \rangle^2 \right\rangle \left\langle R_j^2 - \langle R_j \rangle^2 \right\rangle}} \quad (2)$$

The element of $\langle R_i \rangle$ denotes the average return of stock i during N days and is calculated according to Equation (3):

$$\langle R_i \rangle = (1/N)\sum_{t=1}^{N} R_i(t) \quad (3)$$

The correlation coefficient determines the direction and intensity of the correlation between the two random variables. The value of this coefficient varies between -1 and 1, meaning that 1 is a positive total correlation, 0 is no correlation and -1 is a complete negative correlation (Namaki, 2011).

Now, we use threshold constraints to draw the edge. In this way, we assign 0 to correlations smaller than the threshold and assign 1 to correlations larger than the threshold. Therefore, the adjacency matrix A is defined for the considered network in Formula (4):

$$A_{ij} = \begin{cases} 1 & c_{ij} \geq \theta, i \neq j \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

Then, we assume in the obtained network that the edge with C_{ij} weight is plotted between the two nodes if and only if the absolute value of the correlation coefficient is greater than the threshold θ.

The positive correlation coefficient means that the two companies have the same behavior over time and the value of θ determines the degree of this similarity. In fact, if the correlation coefficient is positive, the price increase of the two companies will be directly correlated and, if it is negative, the price increase will be inversely correlated.

2.1.1. Modularity criterion

The Modularity criterion is a quantitative criterion used to compare the different proposed partitions for the network. Using this definition, the community detecting issue can be turned into an optimization problem with the aim of maximizing the Modularity criterion (Lastusilta, Papageorgiou & Westerlund, 2011). Of course, modularity provides a satisfactory answer for a particular group of graphs. For example, in a graph of several clicks, communities may not

be properly recognized (Xu, Tsoka & Papageorgiou, 2007). This criterion was first proposed in 2004 by Newman and Girvan in an innovative algorithm as a quality evaluation criterion. Girvan and Newman, have stated that social networks have not random structure. In other words, from the point of view of these two scientists, the characteristics of a partition of the network are far from features of the randomly distributed network, and that partition is closer to the concept of the community. Accordingly, he defined a target criterion called the modularity with the Q symbol, which measures this difference of characteristics quantitatively (Newman & Girvan, 2004). The following describes the mathematical relations of the modularity.

Consider a network with N nodes and L edges, which is divided into n_c communities. Each community contains Nc nodes and Lc edges. If Lc is larger than the expected value of the edges between the nodes of this community, then we can say that the correct community is selected. Therefore, the difference between the real network diagram (Aij) and the expected value of the edges between i and j in a random network (pij), the modularity is defined as the equation (5):

$$M_c = \frac{1}{2L} \sum_{(i,j)\in C_c} (A_{ij} - p_{ij}) \tag{5}$$

In equation (5), P_{ij} is calculated as (6):

$$p_{ij} = \frac{k_i k_j}{2L} \tag{6}$$

Therefore, modularity can also be written in the form of equation (7) in which kc is the degree of total nodes in the community:

$$M = \sum_{c=1}^{n_c} \left[\frac{L_c}{L} - \left(\frac{k_c}{2L} \right)^2 \right] \tag{7}$$

The M is larger, the better the clustering on the network. If we consider the whole network as a community, the value of M is zero. If we separate each node in a community, Lc is obtained zero, and consequently, M is negative.

3. Network construction

In order to build the initial graph, the necessary data must first be collected and the necessary pre-processing is done.

3.1. Data collection

In order to collect data, an extracted dataset from the rahavard novin software, including the closing price of 465 companies, was used during the period from 2015-03-25 to 2017-03-25, equivalent to 486 working days. Based on the 2010 research findings (Chi, Liu & Lau, 2010),

a two-year period is enough to achieve a significant correlation coefficient in stock market data, and the network is strong and significant. Therefore, this research selected the closing price data in Tehran Stock Exchange over two years. Pre-processing was also carried out on these initial data, so that they were prepared for applying in network analysis software, programming environments and numerical computing, and a list of nodes and edges in the Excel became ready to import to the Gephi software.

3.2. Building the initial graph

In Figures 4, 5 and 6, the created networks in the 0.9.1 version of Gephi software are observed for different values of θ.

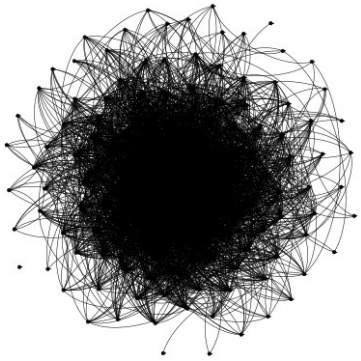

Fig. 4. The correlation network of stock returns based on 0.15

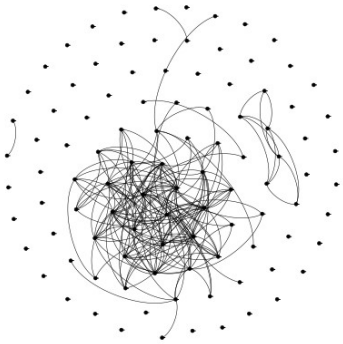

Fig. 5. The correlation network of stock returns based on $\theta = 0.25$

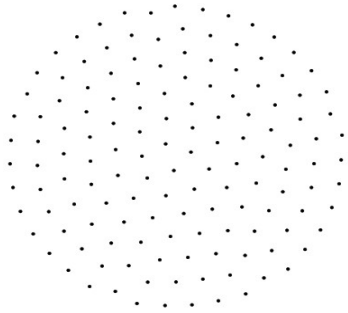

Fig. 6. The correlation network of stock returns based on $\theta = 0.50$

Considering the shape of the networks, the network with a threshold of 0.15 is very dense, in such a way that the average correlation of each node with 46.61 of other nodes is upper than15%. On the other hand, networks with thresholds of 0.25 and 0.50 have very few edges. Hence, the analysis of the relationship between companies on the network is carried out with a threshold of 0.15.

3.3. Network visualization

In the following, for the network visualization, the Fruchterman Reingold layout algorithm was first used in the Gephi software, because this algorithm makes it easy to observe the correlation pattern of stocks across the market (Sun, Tian & Yang, 2015). Then, the Noverlap and Label Adjust layout algorithm were used to create a beautiful view without overlapping the nodes and with their labels being legible. Figure 7 shows the initial graph after applying the layout's algorithms.

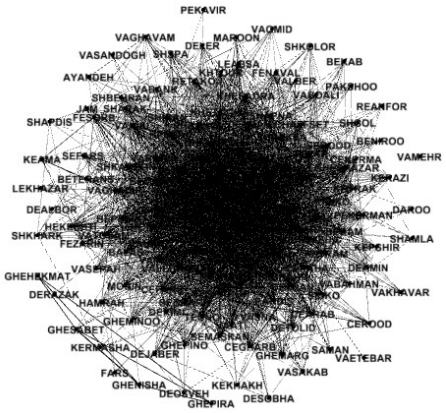

Fig. 7. The primary graph by applying Fruchterman Reingold, Noverlap and Label Adjust algorithms in Gephi

4. Results

In this research, for extracting the existing communities in the stock return correlation network, Louvain algorithm was used in Gephi software, which is a greedy algorithm for optimizing the modularity criterion (De Meo et al., 2011).

Louvain algorithm is an innovative greedy algorithm that was invented in 2008 and is used to extract communities in large networks. This algorithm optimizes the modularity criterion. In the Louvain method, first small communities are found by optimizing the modularity and then, each small community is grouped together with one node and again the first stage is repeated. For more information, see (Blondel et al., 2008).

4.2. Community extraction

In order to detect communities in this study, different colors in the network are used in terms of the modularity class, and a color view of the network is displayed in which the same group nodes have the same color. An intermediate centrality criterion has been used to distinguish the size of the nodes. Thus, the larger nodes have larger intermediate centralities and vice versa. Also, the edges are color-coded to suit their weight, so that the edges with more weight have a darker color and vice versa. After running the Louvain algorithm in the Gephi software, a network with 12 communities (12 different colors) as shown in Fig. 8 was obtained.

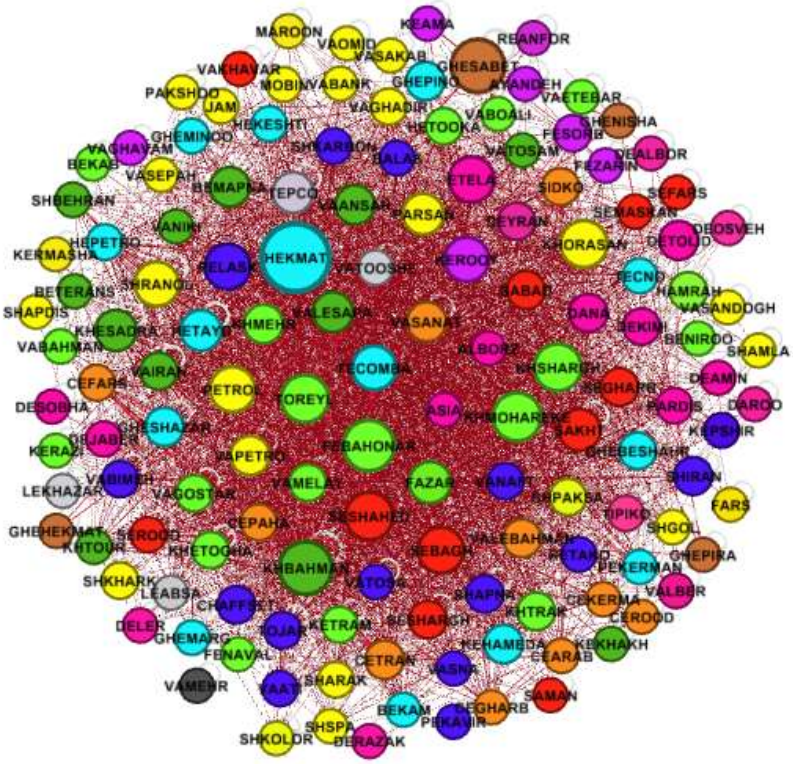

Fig. 8. The network visualization with 0.082 modularity and 12 communities

Table 1 shows network characteristics.

Table 1. Characteristics of the network

Characteristics	Value
Number of nodes	141
Number of edges	3286
Number of communities	12

Layout	Fruchterman Reingold
Modularity	0.082
Average Degree	46.61
Network Diameter	4
Graph Density	0.333
Average Clustering Coefficient	0.714
Average Path Length	1.773

4.3. Community analysis

In general, in the final network after detecting community process by using the Louvain algorithm in Gephi software, a variety of communities with 12 different colors were distinguished from each other. Subsequently, the stocks of the 6 major communities and their analysis are presented.

4.3.1 Dark gray community

In the dark gray community, only VAMEHR is located, which is a symbol of the Mehr bank. The company itself has formed one community, it means that it does not have any correlation with any other company.

4.3.2 Light gray community

In the light gray community, there are only companies of machinery and equipment such as Pars Khazar, Absal and so on. Some of the symbols in this group are: LEKHAZAR, VATOOSHE, TEPCO and LEABSA.

4.3.3 Brown community

In the brown community, there are symbols of GHENISHA, GHEPIRA, GHESABET and GHEHEKMAT which all belong to the sugar group.

4.3.4 Red community

In the red society, all the companies in the group are mass production and properties such as SEROOD, SESHAHED, SEMASKAN, SEBAGH, SEGHARB and etc.

4.3.5 Pink community

In the pink community, there are often companies in the group of pharmaceuticals and materials such as Amin Pharmacy, Sobhan Group, etc. There are also some insurance and investment companies. It could be argued that if more communities had been obtained for the

network, presumably, the insurance companies would have been completely separate from the drug group and the separate community.

4.3.6 Yellow community

In the yellow community, most of the companies in the group are chemical products such as Khorasan Petrochemical Company, Pardis Petrochemical Company and etc. There are also a number of companies related to petroleum products which show a high correlation of petrochemicals, petroleum products and nuclear fuel.

5. Conclusion

Today, according to recent researches, complex networks based on graph theory are used as a new tool for understanding and analyzing a huge body of data. Since the stock market has large data that is constantly changing and fluctuating, it is difficult to analyze these data and manage the stock marketing for investors and corporate executives, so that complex networks are used in stock market analysis. In this study, complex networks and detecting community techniques are used. So the Tehran stock exchange market was considered and after building the stock return correlation network, by using community detection, the stock of the same group was analyzed and it was determined that the co-industry stocks were grouped or if a company is in a group beyond its own industry, instead of that company, a stockholder of the same companies is placed in that group. Therefore, it can be said that the correlation between stock prices of companies largely depends on the industry in which these companies are active. Therefore, it can be concluded that the calculation of correlations by industry is more significant, and the community building of the Louvain algorithm, which is based on the modularity optimization, acts entirely logically and scientifically. On the other hand, large-scale grouping makes it possible, by identifying smaller communities, to be able to use a lower cost and energy to separate each community in greater detail or the entire community at a higher level with less detail to be analyzed, monitored and managed and the process of macro data analysis is facilitated and accelerated.

References

Afrouzy, Z. A., Paydar, M. M., Nasseri, S. H., & Mahdavi, I. (2018). A meta-heuristic approach supported by NSGA-II for the design and plan of supply chain networks considering new product development. Journal of Industrial Engineering International, 14(1), 95-109.

Chaharsooghi, K. and Rahimnezhad, F. (2016). Application of complex networks in stock market analysis. In Proceedings of Innovation in IS/IT Management with BI Approuch conference, Tehran, Iran.

Kito, T. and Ueda, K. (2014). The implications of automobile parts supply network structures: A complex network approach. *CIRP Annals-Manufacturing Technology*, *63*(1), pp.393-396.

Huang, S., Chow, S.C., Xu, R. and Wong, W.K. (2015). Analyzing the Hong Kong Stock Market Structure: A Complex Network Approach. *Available at SSRN 2633433*.

Boginski, V., Butenko, S. and Pardalos, P.M. (2005). Statistical analysis of financial networks. Computational statistics & data analysis, 48(2), pp.431-443.

Tumminello, M., Lillo, F., Piilo, J. and Mantegna, R.N. (2012). Identification of clusters of investors from their real trading activity in a financial market. *New Journal of Physics*, *14*(1), p.013041.

He, K., Li, Y., Soundarajan, S. and Hopcroft, J.E. (2018). Hidden community detection in social networks. *Information Sciences*, *425*, pp.92-106.

Gao, C., Ma, Z., Zhang, A.Y. and Zhou, H.H. (2018). Community detection in degree-corrected block models. *The Annals of Statistics*, *46*(5), pp.2153-2185.

Bhatia, V. and Rani, R. (2019). A distributed overlapping community detection model for large graphs using autoencoder. *Future Generation Computer Systems*, *94*, pp.16-26.

Fortunato, S. (2010). Community detection in graphs. *Physics reports*, *486*(3), pp.75-174.

Veldt, N., Gleich, D.F. and Wirth, A. (2018). A correlation clustering framework for community detection. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web* (pp. 439-448). International World Wide Web Conferences Steering Committee.

Porter, M.A., Onnela, J.P. and Mucha, P.J. (2009). Communities in networks. *Notices of the AMS*, *56*(9), pp.1082-1097.

Papadopoulos, S., Skusa, A., Vakali, A., Kompatsiaris, Y. and Wagner, N. (2009). Bridge bounding: A local approach for efficient community discovery in complex networks. *arXiv preprint arXiv:0902.0871*.

Newman, M.E. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical review E*, *69*(2), p.026113.

Son, S.W., Jeong, H. and Noh, J.D. (2006). Random field Ising model and community structure in complex networks. *The European Physical Journal B-Condensed Matter and Complex Systems*, *50*(3), pp.431-437.

Hu, Y., Li, M., Zhang, P., Fan, Y. and Di, Z. (2008). Community detection by signaling on complex networks. *Physical Review E*, *78*(1), p.016115.

Shariatmadari, F. (2014). Pridicting of stock price fluctuations in Tehran Stock Exchange using social network analysis. Thesis of Master of Science. Faculty of Computer Engineering. Tehran University.

Namaki, A. (2010). Analysis of Tehran Stock Exchange using complex networks and randomized matrix theory. Thesis of Master of Science. Faculty of management. Tehran University.

Lastusilta, T., Papageorgiou, L.G. and Westerlund, T. (2011). A Comparative Study of Solving the Problem of Module Identification in a Complex Network. *CHEMICAL ENGINEERING*, *24*, p.319.

Xu, G., Tsoka, S. and Papageorgiou, L.G. (2007). Finding community structures in complex networks using mixed integer optimisation. *The European Physical Journal B-Condensed Matter and Complex.*

Chi, K.T., Liu, J. and Lau, F.C. (2010). A network perspective of the stock market. *Journal of Empirical Finance*, *17*(4), pp.659-667.

Sun, W., Tian, C., & Yang, G. (2015). Network Analysis of the Stock Market.

De Meo, P., Ferrara, E., Fiumara, G. and Provetti, A. (2011). Generalized louvain method for community detection in large networks. In Intelligent Systems Design and Applications (ISDA), 2011 11th International Conference on (pp. 88-93). IEEE.

Blondel, V.D., Guillaume, J.L., Lambiotte, R. and Lefebvre, E. (2008). Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment, 2008(10), p. P10008.