Block & Comovement Effect of Stock Market in Financial Complex Network

Chongwei Du[1], Xiong Wang[2], and Liyin Qiu[3]

[1] Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, China
visucl.dcw@gmail.com
[2] Department of Mathematics, Shanghai Jiao Tong University, Shanghai 200240, China
wangxiong8686@gmail.com
[3] Department of Information Security, Shanghai Jiao Tong University, Shanghai 200240, China
qiuliyin@gmail.com

Abstract. In the work, we present a method to analyze block & comovement effect of stock market by finding out the community structure in the financial complex network. We choose the stocks from Shanghai and Shenzhen 300 Index as data source and convert them into the complex network in matrix format which is based on the measurements of correlation we proposed in this paper. The classical GN algorithm and the NetDraw tool are applied to obtain the modularity and draw all the community structures. The results of our work can offer not only the internal information about the capital flows in the stock market but also the prediction of variation and trend line of some stocks with delay-correlation.

Keywords: block & comovement effect, complex network, community structure, correlative coefficient, Girvan-Newman algorithm, faction, delay-correlation, prediction.

1 Introduction

Generally, financial market is think of as complex system model [1]. It is the system composed of considerable agents which are interacted by high non-linear method. The study of complex systems has recently been recognized as a new discipline, not only between physics, chemistry and biology, but also between social sciences, including economics, sociology, and psychology, even if interdisciplinary field. The research on the complex network is developed dramatically and it is regarded as a strongly powerful tool for describing and analyzing complex system. With the increasing availability of computer power, scientists can search for regularities and patterns from huge amount of data and empirical observation more efficiently and conveniently. Community structure is a vital characteristic in many real complex network. Searching and analyzing community structure are rich in significance for the study of structure and feature in

J. Zhou (Ed.): Complex 2009, Part II, LNICST 5, pp. 1248–1260, 2009.

network. In the last decades, one issue that has received a considerable amount of attention is the detection and characterization of community structure in networks such as clustering algorithm, divisive algorithm and so on [2] [3] [4].

Financial system is chosen in our work because it is a data rich system and is accumulated with a wealth of high-frequency data supporting the study of complex network theory and empirical analysis. Comovement effect is a typical price phenomenon in security market, which refers to some obvious correlations existing in return-volatility of different securities. In our work, we propose a method to build up the complex network with correlative coefficient matrix in the stock market and apply the algorithm of community structure to analysis the comovement effect in the financial market. The classical GN algorithm is introduced and used in our work for finding the biggest modularity value Q in order to obtain the optimal community structure. Then, we utilize NetDraw tool to draw a detailed community relationship. Finally, we explain and analyze the cause and characteristics of Chinese block and comovement effect. In addition, we discuss the investment strategy and feasibility according to our empirical results which makes theoretical and practical sense.

In Sec.2, we first give a methodology of block & comovement effect in data collection, measurements of correlation we propose and community structure algorithm. We then discuss our empirical results with analysis of comovement effect and application in Sec.3. In Sec.4, we present our conclusions.

2 Methodology of Block & Comovement Effect

2.1 Data Collection

Shanghai and Shenzhen 300 Index is composed of 300 A-shares selected from the Shanghai and Shenzhen stock markets. It covering the sample of about 60 percent of the market value has a strong representation of the Chinese stock market. It is able to reflect effectively the Shanghai and Shenzhen A-share market's overall tendency of market conditions and fluctuations in the all characteristics. Therefore, we choose all 300 A-shares as our source data for further process in the work. All statistical data of 300 stocks of Shanghai and Shenzhen 300 Index we consider are from Jan.4, 2007 to May.19, 2008, during which it passed through the Bull market and the Bear market in China, which helps us analyze the potential comovement of stock in different phases.

2.2 Measurements of Comovement Correlation

In the first place, let us consider the classic linear correlation coefficient. It is defined as:

$$\rho_{XY} = \frac{Cov(X,Y)}{\sqrt{D(X)D(Y)}} \tag{1}$$

where X and Y are two arrays of random variables (To stock market, array consists of a sequence of the price of one certain stock). When this method is used to quantify the correlation of two stocks, it is easy to make it into practice

but not so accurate since it is quite possible that the correlation between two really related stocks are not linear. Hence, we propose an improved measurement in our work.

Given price time series of certain two stocks, the dp_i and dq_i denote the derivative of the price of two stocks at the certain t moment. We can get the correlation coefficient as follows:

$$\int_{t1}^{t2} p_i' q_i' dt \tag{2}$$

where the product of two derivative will be positive if the price of two stocks increase simultaneously, contrarily negative while one increases and the other decreases. Therefore, the integration will be huge if there exists strong correlation between two stock during the corresponding period and it trends to 0 by offset the positive and the negative with no correlation. Actually, with regard to discrete sequence of price, we set interval of derivative t as 1 day, so the increase of everyday closing price is what we need in our work. Considering the different stock price varies in a large range, relative increase is used in order to make the measurements more accurate and comparable. Then, we have the complete definition of our correlative coefficient r_{ij}:

$$r_{ij} = \frac{1}{T} \sum_{t=1}^{T-1} [\frac{p_i(t+1) - p_i(t)}{p_i(t)} \cdot \frac{p_j(t+1) - p_j(t)}{p_j(t)}] \tag{3}$$

where $p_i(t)$ denotes the closing price of stock i at $No.t$ day. r_{ij} is normalized by the sum of product divided by T. So the correlative coefficient r_{ij} is independent of the length of time interval during which the products of increments of closing price of the two stocks are accumulated. In addition, we substitute the everyday closing price with values obtained by 3-day moving average of all stocks. The moving average of 3-day is chosen because the trend line will be relative smooth and efficient to treat the datasets, i.e., $p_i = (p_{i-1} + p_i + p_{i+1})/3$. For the convenience of our data process, we multiply all data with the constant 10^3. In this case, coincidently, almost all the correlative coefficient r_{ij} ranges from 0 to 1. Figure 1 shows the statistical distribution of coefficient (Compared to the number of r_{ij} between 0.25 and 0.3, those bigger than 0.6 are too few to be shown clearly in the following histogram):

We build up a matrix to represent the correlation between stocks. The elements a_{ij} means the correlative coefficient between stock $No.i$ and stock $No.j$. Due to every 2 stocks have 1 correlative coefficient, there are totally $C_{300}^2 = 44850$ coefficients in the $300 * 300$ symmetrical correlation matrix (The elements on diagonal are 0). Consequently, we set a threshold value to determine whether two stocks are correlative, i.e. via threshold value, we judge whether there is an edge between 2 vertices in the stock network. If the threshold is too small, the number of edges between vertices will be too many to detect community well. Similarly, too large threshold will give rise to too little information and edges to analyze. Based on the analysis of the number of edges between vertices in the stock network via empirical test, 0.45 is chosen to be the threshold in order to obtain

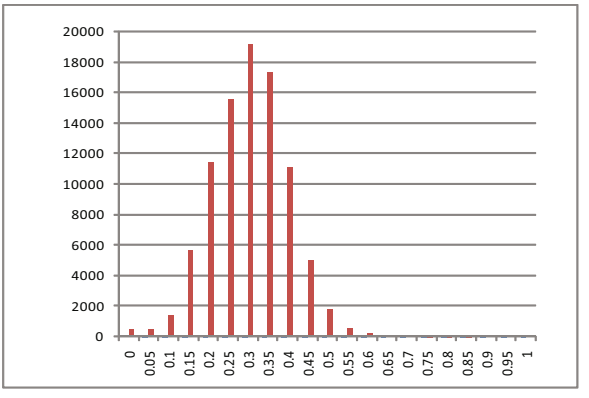


Fig. 1. The histogram of correlative coefficient

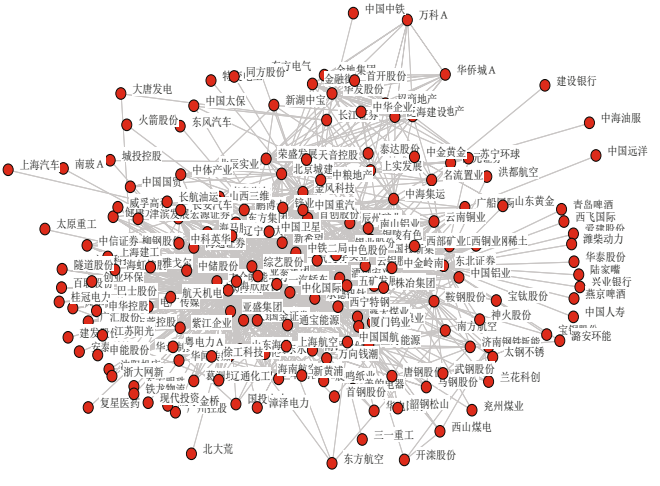


Fig. 2. The graph of the correlation of 300 stocks

better effect of analysis. Then, two stock are regarded as correlative (existing an edge) only when their coefficient of the edge is larger than 0.45. Now a new graph has been initialized with 300 vertices and 1321 edges. Figure 2 [5] is built up according to the correlative coefficient matrix (Each vertex represents one stock.).

2.3 Description of Community Structure Algorithm

In this part, we introduce the method and the algorithm we utilize to find the community structure in stock complex network. M.E.J. Newman and M.Girvan [2] proposed an algorithm to extract the community structure from complex networks which is a kind of divisive method. We implement this algorithm by programming and try to find out some relevant information which indicates the

blocks and community structure as our target. In general, the GN Algorithm has the following basic processes:

1. Calculate the betweenness score for each of the edges in the network
2. Find the edge with the highest score and remove it from network
3. Go back to step 2 until the system breaks up into N non-connected vertices

The betweenness is defined as the number of shortest paths passing through every edge in the network. Obviously, if there are many shortest paths going through one edge, this edge is crucial to the network and it is quite possible that if we remove this edge, the whole network will split into two parts (communities) with strong structural characteristic. To put this algorithm into implementation, something detailed should be taken into consideration. The one is how to search and measure the betweenness score for each vertex, and the other is how to judge a good and effective community in the network. There are many methods to calculate all the betweenness, such as defining the betweenness like the random walker betweenness or the current-flow betweenness which are not be mentioned here [2]. The way we use in our work is to use breath-first search(BFS) algorithm. Before implement BFS, another concern is that for a certain pair of vertex, there could have several shortest paths with equal length. We use the process proposed by M.E.J. Newman and M.Girvan [2]. If there are N shortest paths for one pair of vertex, we add $1/N$ to the betweenness of each edge in the N paths. Thus, we could make the total betweenness a constant. The detailed steps of calculating the betweenness of each vertex based on BFS:

1. For the source vertex s, we define the distance $d_s = 0$, the power $w_s = 1$, and the degree $deg_s = 0$;
2. Search the graph by BFS. Every time, a vertex j is found from the source vertex i, considering:
   (a) If this vertex is found first time, we set the $d_j = d_i + 1$, $w_j = w_i$, $deg_i = deg_i + 1$;
   (b) If this distance of vertex has been assigned to $d_j = d_i + 1$, we add the power w_j by w_i;
   (c) If the distance also been assigned, but $d_j < d_i + 1$, this vertex is ignored.
   (d) Repeat step c until the end of search.
3. After the calculation of distance and power, we now use following method to figure out the betweenness:
   (a) We use the sequence of topological sort to process each vertex (degree of vertex has been calculated). For each vertex j, every time we got a leaf-node I if there exists an edge between I to j. Then we add the betweenness of this edge by $w_j/w_i * (sum_i + 1)$, where sum_i is the total weight of vertices which lie under node I in topological order;
   (b) Repeat the process until all the vertices are considered;
   (c) For the betweenness of the network, we need to set each vertex as source and do the calculation above.

The complexity of this algorithm is $O(mn)$,where m denotes the number of edges and n denotes the number of vertices in all. Since our network is a sparse

network, the complexity comes to $O(n^2)$. For the whole process, we need to recalculate the betweenness matrix every time we remove an edge from the network. So the complexity for this algorithm is $O(m^2n)$, and $O(n^3)$ for sparse network.

For the problem how to define a good community and which step we terminate the split, we use Modularity Q [2] to quantify how well our community is split. We define a $k * k$ symmetric matrix E whose elements e_{ij} is the fraction of all edges in the network that link vertices in community i to vertices in community j [6]. Here all edges in the origin network should be considered even if it is removed during the process. Based on denoting $Tre = \sum_i e_{ii}$ as the sum of all elements on the diagonal of the matrix and $a_i = \sum_j e_{ij}$ as the sum of all elements in $No.i$ row (or column), the following expression is used to define the criteria of modularity:

$$Q = \sum_i (e_{ii} - a_i^2) = Tre - \parallel e^2 \parallel \tag{4}$$

where $\parallel x \parallel$ denotes the sum of all elements in the matrix. Its physical significance is that the ratio of the edges connecting two same kind vertices (the edges in community) subtract the expectation of ratio of the edges, arbitrarily, connecting these two vertices under the same community structure, i.e., if the ratio of the edges in community is no larger than the expectation by arbitrarily connecting, then $Q = 0$. The upper limit of Q is 1, and the closer Q is to this value, the more obvious the community structure is. Another key point is that each edge could only exists once in this matrix. To achieve this, we have two methods. The first one [2] is to split the edge e_{ij} half by half with equivalence of e_{ij} and e_{ji} elements. The second one we propose is to use the expected value of the proportion of the number of the edges. For example, we define the value e_{ij} as $e'_{ij} * c_i/(c_i + c_j)$, where e'_{ij} is the previous value of e_{ij}, c_i denotes the number of vertices in community i.

We use the above algorithm to process the stock data to try to find any relationship of blocks. We draw a graph of modularity Q during the process shown in Figure 3:

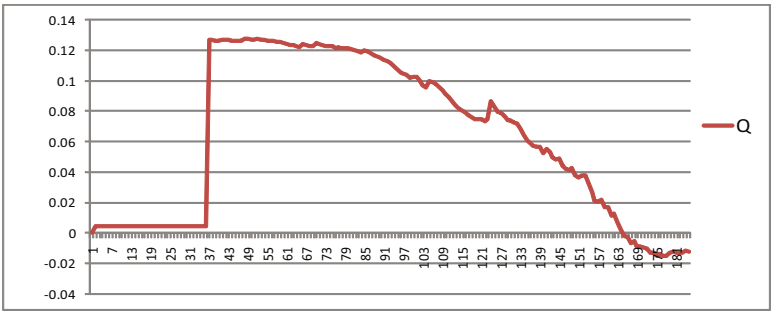


Fig. 3. The relationship curve between Modularity Q and the number of communities

In the origin network, we have 116 isolated vertices which have no edge with other vertices. We will disregard the 116 isolated vertices. In the graph of Q, we find the maximum Q value 0.127 appears when 38 communities are formed. In order to illustrate the communities' relationship better, we use the Faction algorithm [7] based on Tabu search [8]. The Faction is defined as a group whose members are connected more tightly to each other than to the members of other group, which is similar to the self-contained GN algorithm [9]. The NetDraw tool is used to draw the community relationship graph by the Faction algorithm. This algorithm is only different from the classical GN algorithm by the criteria of good communities. In this algorithm, two levels of communities are defined:

1. Definition of a community in a strong sense:
   The subgraph V is a community in a strong sense if $k_i^{in}(V) > k_i^{out}(V)$, $\forall i \in V$. And each node has more connections within the community than the rest of the network.
2. Definition of a community in a weak sense:
   The subgraph V is a community in a weak sense if $\sum_{i \in V} k_i^{in}(V) > \sum_{i \in V} k_i^{out}(V)$. And the sum of all degrees within V is larger than the sum of all degrees toward the rest of the network.

In the NetDraw tool, the Faction algorithm is similar to the GN algorithm[7]. When this Faction algorithm is used, we need to set some parameters, and one of them is the number of communities. We assign the number as 38, which comes from the result with maximum Q of classical GN algorithm.

3 Discussion of Empirical Results

3.1 Analysis of Block and Comovement Effect

Through the GN [2] and Faction algorithm we describe last section, we get the block map of 38 community structures in clustering form, which is shown in Figure 4:

Usually, in a stock market, the nominal sectors are divided according to the industry, geographical location and also the concept. For example, industry sectors include energy block, steel block, etc; geographical location sectors include Beijing block, Shanghai block, etc; concept sectors include Olympic block, High-Tech block, etc. The traditional method of division has its defect because not all the shares in a block have a strong correlation. The objective of our work is to find out the accurate blocks in which all the shares are strongly correlative.

In 38 clustering communities we obtain, according to the size and characteristic of community, respectively, we classify them into big community whose size is more than 8 and corresponding small community; the community whose members are belonged to the same nominal sector and the community whose members build up new block with strong correlation. We find stronger correlation(similar Trend Line) in small community and relatively loose structure in big community. However, we mainly focus on the second method of category, because the new community belonged to origin nominal sector owns stronger correlation and more precise

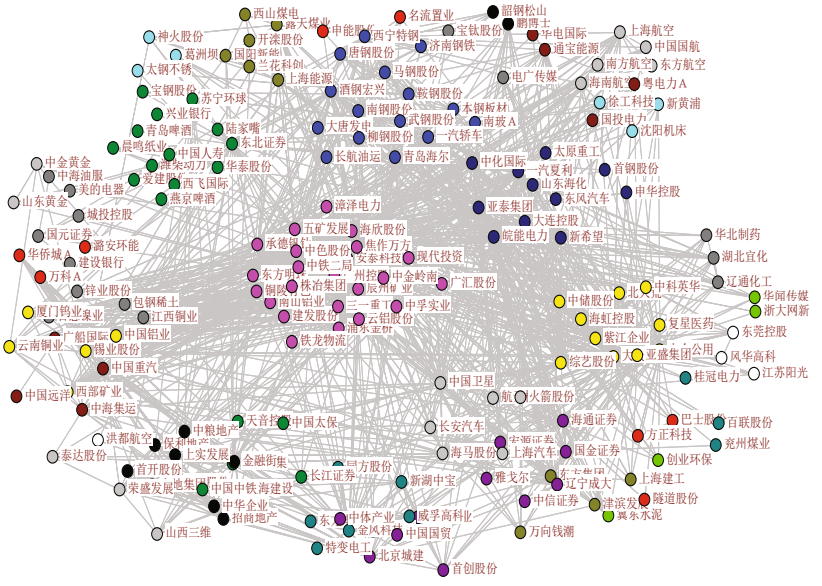


Fig. 4. The clustering structural correlation of 300 stocks

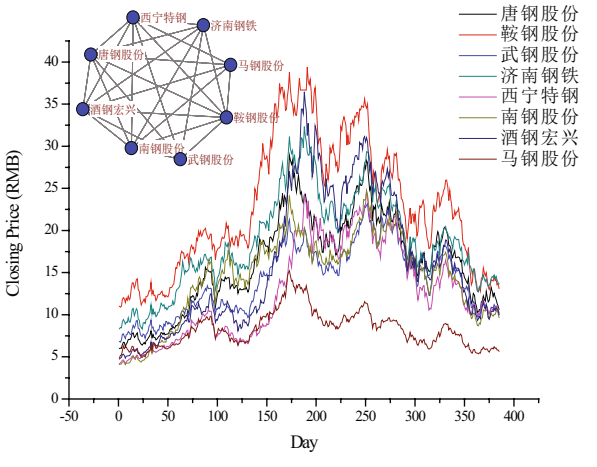


Fig. 5. The Community Structure and Trend Line of new community belonging to Steel Block

scope when we make the decision of investment; and the stocks composing new community will offer valuable information about capital flow and portfolio of institute of investment and internal potential relationship between several stocks under nominal sector.

In the next step, we will show and analyze some of new community we find out(Because of limitations of space, we can not list all the results here. The

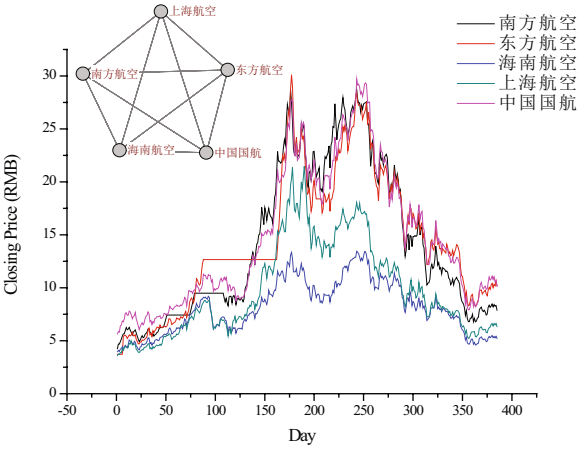


Fig. 6. The Community Structure and Trend Line of new community belonging to Airline Block

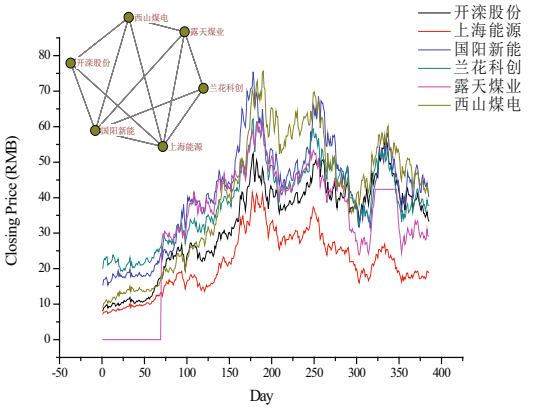


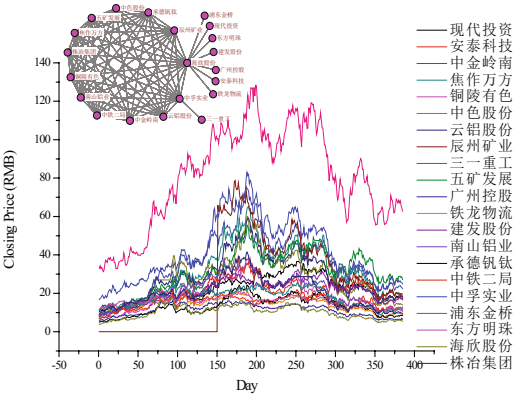
Fig. 7. The Community Structure and Trend Line of new community belonging to Energy Block

horizontal line and step in part of trend line mean that this stock is suspension in that period of time). Figure 5 shows the example of new accurate community of Steel Block.

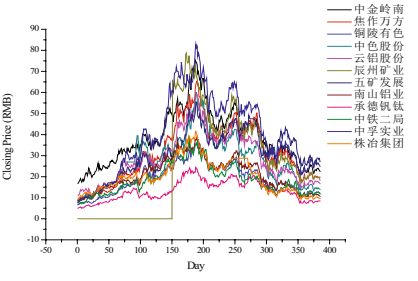
Figure 6 shows the example of new accurate community of Airline Block:

Figure 7 shows the example of new accurate community of Energy Block: We can easily find that the stocks in the same new community own similar Trend Line in above figure.

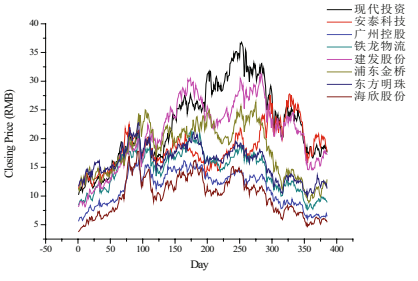
Following figures include the stocks coming from different nominal sectors and composing the new potential comprehensive concept block. Figure 8(a) shows the example of new community with the combination of the Nonferrous Metal Block and the High Technology Block. Figure 8(b): shows the Trend Line of left



(a) The graph of whole Community Structure and Trend Line



(b) The Trend Line of Left Part of the community



(c) The Trend Line of Right Part of the community

Fig. 8. The new community belonging to the Nonferrous Metal Block and the High Technology Block

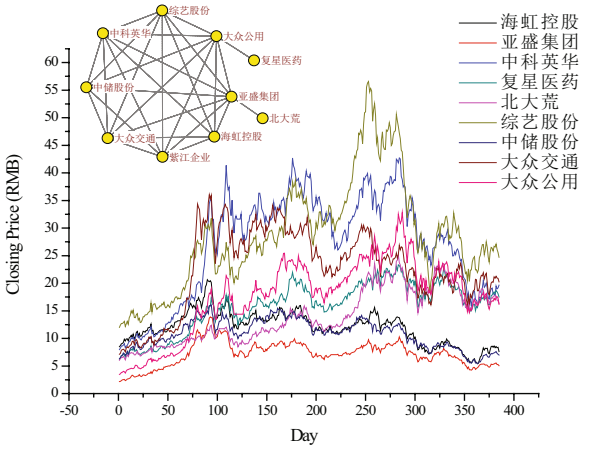


Fig. 9. The new community belonging to Comprehensive Block

part of the community, the Nonferrous Metal Block, and Figure 8(c) shows the Trend Line of right part of the community, the High Technology Block:

Figure 9 shows the correlative stocks involved in the Transportation, Material, Agriculture and Medicine fields:

By above 2 figures and analysis, we can easily find that the stocks in the same new community almost change simultaneously which meets empirical situation, even if they come from various origin nominal sectors.

3.2 Synchronous and Asynchronous Variation and Further Application

If certain stocks are invested by some capital flows simultaneously, in the community, the stocks will change synchronously as shown in Figure 10(a), 10(b), 10(c) and 10(d) (The price of the stocks increase or decrease almost in the same time):

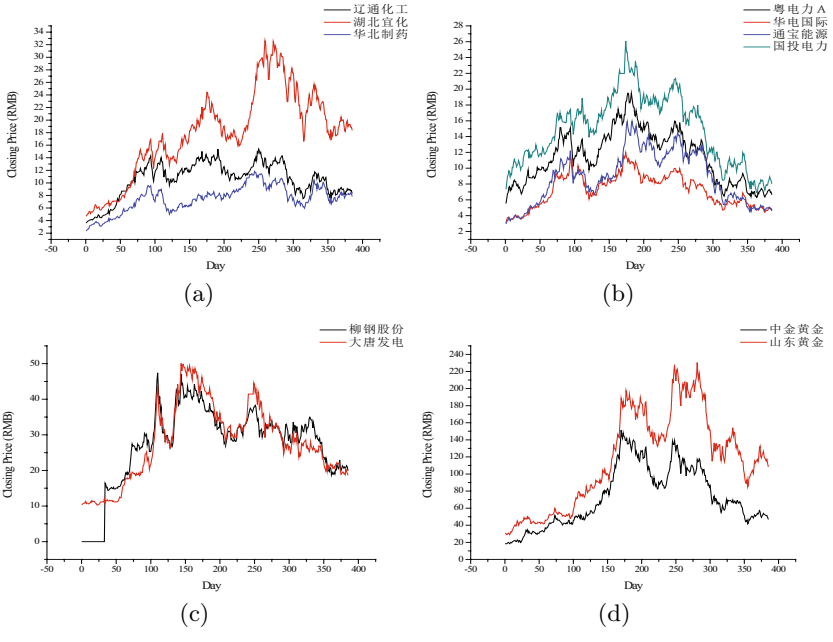


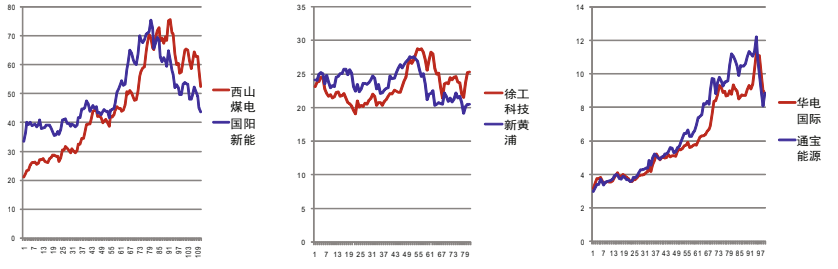
Fig. 10. The synchronous variation in the community

On the other hand, if certain stocks are invested by different capital flows, perhaps, there exists some delay-correlation in the Trend Line which we possibly find by improved method as follows:

1. Move each stock via parallel shift 3, 7, 15 and 30 days, respectively, to other stocks;

2. According to the Trend Line, select the corresponding period of time, such as from 50 to 80 days, and build up the matrix of correlation with the same measurement;
3. Find positive strong correlation by community structure algorithm;

After these steps, we find out some delay-correlation and list 3 most representative results: in Figure 11(a), Xishan Coal and Electricity leads over International New Energy about 7 days; in Figure 11(b), Xu Gong Technology leads over New Huang Pu about 5 days; in Figure 11(c), Huadian Electricity leads over Tongbao Energy about 3 days in the latter part.



(a) Delay: 5-10 days (b) Delay: 4-6 days (c) Delay: 2-4 days

Fig. 11. Delay-correlation in the market by community structure search

With regard to the results, because the scale of capital flows is relative smaller to all huge stock market, it is impossible that we can find many obvious community delay-effect. In addition, only short-term delay are discovered other than obvious long-term delay effect for large capital stock in Shanghai and Shenzhen 300 Index. Therefore, via our work, we can obtain the following applications:

1. If the stocks of new community are belonged to one origin nominal sector, we can locate the most correlative stocks accurately from the large range;
2. If the stocks of new community are belonged to different origin nominal sectors, we can find the internal potential correlation from different nominal sectors;
3. If we learn the variation of one stock in the community, at this moment, to gain more returns, we can invest other stocks in that community which are at the bound of community with small transaction volume and uncertain tendency;
4. With regard to the stock with delay-effect, so-called leading shares, we can predict the variation of other stocks according to it and make the decision to invest them in advance.

4 Conclusion

In this paper, the matrix of correlation from Shanghai and Shenzhen 300 Index has been proposed and built up first. After the process by community structure

algorithm and modularity Q of the network, we find the stocks having strong correlation and obtain the accurate blocks in the Chinese stock market. Based on the comovement effect in synchronous variation and delay-correlation, we offer some investment strategy and suggestion coming from empirical analysis as our further applications.

Acknowledgments. The authors wish to thank Zhongxing Ye, Haibo Hu and Junjun Tang for their valuable suggestions and discussions. Thanks also to Xiaofan Wang for providing the reference. This work was supported by National Basic Research Program of China(973 Program No. 2007CB814903) and National Natural Science Foundation of China (No. 70671069 and 10801097).

References

1. Bonanno, G., Lillo, F., Mantegna, R.N.: Levels of complexity in financial markets, arXiv: cond-mat/0104369 (2001)
2. Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. Phys. Rev. E 69, 026113 (2004)
3. Pujol, J.M., Bejar, J., Delgado, J.: Clustering algorithm for determining community structure in large networks. Phys. Rev. E 74, 016107 (2006)
4. Hu, Y., Li, M., Zhang, P., Fan, Y., Di, Z.: Community detection by signaling on complex networks. Phys. Rev. E 78, 016115 (2008)
5. Due to no formal translated English name for the stock in China, we use its origin Chinese name in Figure but translate it in our analysis and discussion
6. Newman, M.E.J.: Mixing patterns in networks. Phys. Rev. E 67, 026126 (2003)
7. Hanneman, R.R.: Introduction to social network methods (2005)
8. Glover, F.: Tabu Search Part 2. ORSA Journal on Computing 2, 4–32 (1990)
9. Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., Parisi, D.: Defining and identifying communities in networks. Proc. Natl. Acad. Sci. 101, 2658–2663 (2004)