Research on stock similarity and community division based on user attention sequence

Gaowei Zhang¹, Lingyu Xu^{1,*}, and Lei Wang²

¹Computer Sciences, Shanghai University, 200444, China

²East Sea Information Centre, SOA China, 200444, China

Abstract. We conduct research from the perspective of user groups and analyze the differences in the users' attention and posting order in different time periods to vectorize stocks and build relationships from the generated vectors. This provides a new perspective for the complex network construction and community division of network public opinion space. The experiment result show that we can get the community division consistent with reality using our model.

1 Introduction

In practical society, there is a mutual connection between things, this kind of connection also hides a certain regularity, the use of complex networks can be simple and effective description of the interconnection between things. Therefore, many researchers have used complex networks to describe this connection in the financial stock field, and have done a lot of research. Current researches mainly study the relationship by the stock return, which can not describe the interaction between stocks well. In addition, the group behavior of the stock is also the focus of investors. In different periods of the stock market, the performance of group behavior can help investors to better grasp the dynamics of the market. On the basis of establishing the stock network, the community detection can make a better summary and analysis of the group behavior.

Complex networks help to understand the interactions between the various elements of a complex system and how this interaction affects the overall properties of the entire complex system [1,2,3,4], which has given rise to complication. In recent years, a large number of studies aiming at the statistical properties and topological properties of networks. Through a large amount of research, most of the complex networks have been found to be different from the properties of random networks, such as the "small world" [5,6], scale-free distribution [7] and community structure [8,9,10,11], Rich club [12,13,14] and so on.

In this paper we carry on the experimental verification and analysis to the stock network, and finds out that the similarity between the stocks has the normal distribution in different periods.

2 Related work

* Corresponding author: xly@shu.edu.cn

The first application of complex networks in financial markets was the minimal spanning tree stock network model constructed by R.N.Mantegna [15]. Later, in 2004, Newman used the correlation coefficient to establish a stock network to obtain the shortest path between stocks and analyze the correlation between the non-adjacent stocks in the network. J.P.Onnela [16] developed a stock network by gradually adding edges to the network in order of strength of relevance, and studied the stocks of the New York Stock Exchange. A. Vizgunov [17] selected the data of Russian stock market from 2007 to 2011, calculated the correlation between stock prices, constructed the stock network and found that there was a strong connection between a large number of stocks and the largest faction in the Russian stock market. A. Namaki [18] studied the stocks of the Tehran Stock Exchange in Iran and found some unique topological properties of networks under different thresholds, that is, scale-free and power-law distribution. In 2010, Tes CK used the stock's closing price index to establish the correlation coefficient, used the correlation coefficient as the weight to analyze the relationship between stocks and studied the community structure [19].

Domestic use of complex networks to model the stock market started later, but in recent years still maintained a high degree of research. For example, Huang [20] take the stocks of China's SSE 180 Index and the Shenzhen Stock Exchange 100 Index as the research objects, and used the maximum filter and minimum spanning tree (MST) methods to model the stock network. From the network's perspective and viewpoint analyse the impact between stocks. Yang [21] take 284 stock daily return data of 13 industries in Shanghai and Shenzhen 300 constituent stocks as the experimental subjects, and used the stock as the node to calculate the correlation coefficient between stocks and construct the stock daily yield network model, and examined the small-world and scale-free characteristics of the network and further analyzed the statistical characteristics of the stock network structure. Chen [22] used stochastic matrix theory and correlation matrix decomposition method to analyze the network structure characteristics of stock and collective behaviors in Shanghai and Shenzhen stock markets. Wang [23] put forward a hierarchical node clustering algorithm based on influence calculation model, introduced the definition of node's activity and influence and completed the social division of stock network.

3 Stock vectorized representation model

Definitions 1 Network Opinion Space Complex networks are represented as a two-node group consisting of edges and nodes, $G = (V, E)$, where V is represented as a set of nodes, and each node corresponds to stock, the size is $|V|$. E is a set of edges representing the relationship between nodes and nodes, the number of edges in each figure is $|E|$. In this paper, we vectorize each stock as $C(stock)$ according to the stock sequences of all users in the public opinion space. Each point in the network is a stock and represented by its stock code.

In this paper, we use the representation model based on the word vector to model the user's attention sequence. The model structure is shown in figure 1. Assuming the attention sequence of the user u_q in the network's opinion space, represented as $S_q = \{stock_{t_1}^1, stock_{t_2}^2, ..., stock_{t_n}^m\}$. Then the different users' attention sequence constitutes a group of interest series S , the attention of the stock in the set S constitutes a non-overlapping set of stocks $V_STOCK = \{stock^1, stock^2, ..., stock^k\}$, k is the number of non-overlapping stocks. The stocks in each user's attention sequence over a period of time are in some relation to their contextual concern stocks, the probability of the j -th stock

$stock_{t_j}^j$ in the attention sequence S_q of the q-th user u_q can be inferred from the context in which the stock appears, $P(stock = stock_{t_j}^j | context)$, where $context$ is a context-aware stock within a m-sized window of $stock_{t_j}^j$, then we can turn the attention sequence representation problem into a maximum likelihood probability.,

$$P(Z_1 = stock_{t_1}^1, ..., Z_n = stock_t^n) =$$
$$\prod_i P(Z_i = stock_{t_i}^i | g_i(Z_{i+1-m-1/2} = stock_{t_{i+1-m-1/2}}^{i+1-m-1/2},$$
$$Z_{i+2-m-1/2} = stock_{t_{i+2-m-1/2}}^{i+2-m-1/2}, ..., Z_{i-1} = stock_{t_{i-1}}^{i-1})$$

(1)

It can be further transformed into the following formula,

$$\frac{1}{n} \sum_{i=1}^n \sum_{-q \leq m \leq q, m \neq 0} \log P(stock_{t_i}^i | stock_{t_{i+m}}^{m+i})$$

(2)

where m is the size of the context window, and the larger the value of m is, the more data will be entered into the model to make the model more accurate, but the training time will also increase greatly.

Fig. 1. Model structure where represents the network model and represents the characteristic representation of the i-th stock.

4 Experiment

In the period of 2014/10/01 to 2015/05/31, 200 non-stop stocks in different sectors of the SSE Composite Index were selected, they can be found in table 1.

Table 1. Stock Used.

Stock code	Stock code	Stock code	Stock code	Stock code	Stock code	Stock code	Stock code
000002	002033	600005	600642	000651	002484	600235	600980
000004	002041	600006	600648	000652	002500	600248	600993
000011	002053	600015	600680	000702	002533	600259	600999
000017	002083	600016	600684	000717	002544	600265	601005
000025	002128	600022	600690	000750	300024	600267	601098
000027	002142	600050	600697	000758	300029	600316	601166
000029	002185	600051	600708	000759	300033	600322	601168
000043	002186	600059	600713	000767	300039	600323	601169
000049	002237	600064	600742	000789	300042	600327	601288
000088	002243	600072	600768	000816	300046	600348	601318
000159	002262	600082	600784	000823	300061	600350	601328
000400	002265	600088	600785	000825	300067	600362	601369

000402	002267	600100	600802	000836	300080	600371	601398
000404	002287	600108	600811	000869	300092	600386	601600
000421	002297	600123	600831	000876	300102	600391	601601
000524	002304	600127	600834	000885	300103	600397	601628
000529	002313	600133	600836	000888	300105	600409	601668
000550	002338	600138	600854	000898	300107	600449	601678
000554	002362	600150	600855	000900	300120	600459	601727
000561	002391	600192	600859	000910	300127	600460	601801
000563	002401	600198	600873	000916	300138	600493	601808
000568	002419	600199	600883	000921	300176	600506	601818
000606	002420	600210	600893	000937	300185	600551	601898
000613	002454	600220	600969	000983	600000	600573	601919
002017	601998	600608	601988	001896	600004	600581	601939

We crawl the user group concerns including posting and reply sequence on network public opinion space during 2014.7 to 2015.7. We get the attention vector of the 200 stocks that have not been suspended in the above time period.

$$sim(stock_i, stock_j) = \frac{\sum_p C(stock_i)^p \times C(stock_j)^p}{\sqrt{\sum_p [C(stock_i)^p]^2} \sqrt{\sum_p [C(stock_j)^p]^2}} \quad (3)$$

We use the formula (3) to calculate the similarity of each stock at different time periods and obtain the time series of similarity. We make contour heat maps for the experimental results. The results are as shown in figure 2.

(a) 2014.7.2-2014.8.12 (b) 2014.11.20-2015 (c) 2015.3.11 - 2015.4.2 (d) 2015.6.16-2015.7.28

Fig. 2. Schematic diagram of stock similarity at each stage.

Among them, the darker the red, the higher the similarity. The darker the blue, the lower the similarity is. Each row in the chart represents a stock, each column represents a day. It can be seen that in the figures a, b and c, the warm part is increasing, that is to say, in the stationary period, the start-up period, and then during the high tide period, the similarity of stocks gradually increases. During the collapse in figure d, the relationship between stocks is relatively more consistent.

Using the formula (3), the statistical analysis of the similarity results is shown in figure 3. It can be seen that the similarity of stocks in different time periods basically accords with the normal distribution. Similarly, we set the proportion of similarities between the shares of stock to the sum of the similarities of the attention vectors of all the 200 stocks as

$$\mathcal{G}, \mathcal{G} = sim_i / \sum_{i=1}^n sim_i .$$

Fig. 3. Network public opinion space stock attention vector similarity distribution.

From figure 3, we can see that the first 10% nodes with the highest similarity between nodes form our complex network. So we choose $\mathcal{G} = 0.1$ as the threshold to construct our

network. The results of the complex network in the network public opinion space are shown in figure 4.

Fig. 4. Network public opinion space complex network construction results in different periods.

Then we use AP algorithm to cluster stock similarity. Figure 5 shows the result.

Fig. 5. Community classification results.

Table 2. Community Center and the section where it is located.

(a) 2014(7.2-8.12)			(b)2014(11.20-1.7)			(c) 2015(3.11 - 4.27)			(d)2015(6.16-7.28)		
stock	section	no	stock	section	no	stock	section	no	stock	section	no
000823	electronic	15	002243	Light industry	22	000898	Steel	14	600000	Bank	20
000888	entertainment	37	002419	Trade	31	000529	Synthesis	25	600016	Bank	132
000983	mining	18	002544	Telecom	66	000554	Chemical	86	600350	Transport	32
002243	Light industry	27	600108	Animal husbandry	36	600834	Transport	35	600648	Estate	16
002267	Utilities	10	600210	Light industry	4	600386	Media	20			
002362	computer	14	601369	Equipment	41	600064	Estate	11			
600391	National defense industry	22				300029	Electrical equipment	9			
601628	Non-bank financial	24									
601668	Building decoration	33									

Table 2 shows the statistical result of the community division. Taking the time period from 2015.6.16 to 2015.7.28 as an example, there are 132 stocks with 600016 as the center of the community, among which 152 stocks have the strongest relationship with bank shares, accounting for 75% of the stock. We find that most of the stocks in the stock market during the period of collapse are the banking sector, which means that there is a strong relationship between bank stocks and other stocks. And the actual situation is also the same, in the 2015.6 during the collapse, most of the stocks are plummeting energy outflow status, and

large-cap stocks such as bank shares and real estate have repeatedly led the entire stock market during this period. In other time periods, it can be seen that (a) the distribution of nodes in the time period is more even, while the distribution of (b), (c) is more concentrated, indicating that in the stationary period, the relationship between stocks, and during the start-up and soaring period, it is more concentrated and more in line with the actual situation.

5 Conclusion

We construct the complex network of the stock market from the perspective of the group's awareness activity on the network using the embedding based model. It can be seen from the experimental results that the social division result is in line with the people's division of stock in the real society. Network public opinion information is not only in the East Money Post Bar considered in this article. In future research, it may also need to consider the information contained in Weibo and WeChat. We may need to include a large range of other transaction-related data and be subject to a number of realistic trading rules.

Acknowledgement

This paper is supported by Project supported by the National Research Program of China (Grant No. 2016YFC1401902).

References

1. Albert R, Barabasi A L. Statistical mechanics of complex networks [J]. Rev. Mod. Phys.(2002)
2. Boccaletti S, Latora V, Moreno Y, et al. Complex networks: Structure and dynamics[J].Phys. Rep.(2006)
3. Newman M E J. The structure and function of complex networks [J]. SIAM Rev, **45**,167-256(2003)
4. Dorogovtsev S N, Mendes J F F_ Evolution of Networks: From Biological Nets to the Internet and the WWW[M]. Oxford: Oxford University Press (2003)
5. de Vries, L.,Gensler, S., Leeftang, P.S.H.: Popularity of brand posts on Brand Fan pages: an investigation of the effects of social media marketing. J. Interact. Mark. **26**, 83–91 (2012)
6. Watts D J, Strogatz S H. Collective dynamics in ‘small-world , networks[J]. Nature, **393**, 440-442(1998)
7. Asur, S., Huberman, B.A.: Predicting the future with social media. In: The 2010 IEEE/WIC/ACM International Conference on Web Intelligence and IntelligentAgent Theory, Toronto, Canada. doi:10. 1109/WIIAT.2010.63 (2010)
8. Bollen, J., Mao, H. & Zeng, X. Twitter mood predicts the stock market. J. Comp.Sci. **2**, 1–8 (2011)
9. Newman M E J. Coauthorship networks and patterns of scientific collahoration[J]. Proc.Natl. Acad. Sci. U.S.A., **101**,5200-5205(2004)
10. Krause A E, Frank K A , Mason D M, el al. Compartments revealed in food-web struc-lurefJl. Nature, **426**,282-285 (2003)
11. Brody, D., Meister, B., Parry, M.: Informational inefficiency in financial markets. Math. Fin. Econ **6**, 249–259 (2012)

12. Jiang Z Q, Zhou W X. Statistical significance of the rich-club phenomenon in complex networks[J]. New J. Phys. **10**,043002(2008)
13. Colizza V, Flammini A, Serrano M A, et al. Detecting rich-club ordering in complex networks[J]. Nat. Phys. **2**,110-115 (2006)
14. Zhou S, Mondragon R J. The rich-club phenomenon in the internet topology[J]. IEEE Comm. Lett. **8**,180-182 (2004)
15. Mantegna R N. Hierarchical structure in financial markets[J]. The European Physical Journal B, **11**,193-197 (1999)
16. Onnela J P, Kaski K, Kertész J. Clustering and information in correlation based financial networks[J]. The European Physical Journal B, **38**,353-362 (2004)
17. Vizgunov A, Goldengorin B, Kalyagin V, et al. Network approach for the Russian stock market[J]. Computational Management Science, **11**,45-55 (2014)
18. Namaki A, Shirazi A H, Raei R, et al. Network analysis of a financial market based on genuine correlation and threshold method[J]. Physica A Statistical Mechanics & Its Applications, **390**,3835-3841(2011)
19. Chi K T, Liu J, Lau F C M, et al. A network perspective of the stock market[J]. Journal of Empirical Finance, **17**,659-667 (2010)
20. Huang Weiqiang, Zhuang Xintian, Yao Shuang. Topological Properties and Clustering Structure of China Stock Correlation Networks [J]. Management science, **21**,94-103(2008)
21. Yang Zhihui, Jia Hanmei. The Network Structure Analysis of Stock Returns Correlation [C]. Chinese Society of Automation Control Theory Committee B. 5732-5736 (2011)
22. Chen H, Mai Y, Li S P. Analysis of network clustering behavior of the Chinese stock market[J]. Physica A Statistical Mechanics & Its Applications, **414**,360-367 (2014)
23. Wang Hao, Li Guo Huan, Yao Hongliang, Li Junzhao. Stock Network Association Method Based on Influence Calculation Model.[J]. Computer Research and Development,**1**,2137-214 (2014)