The Evolution of Degree Distribution, Maximum Cliques and Maximum Independent Sets of Company Co-Mention Network over Time

VLADIMIR A. BALASH, ALEXEY R. FAIZLIEV, ELENA V. KOROTKOVSKAYA,
SERGEI V. MIRONOV, FEDOR M. SMOLOV, SERGEI P. SIDOROV, DANIIL A. VOLKOV

Saratov State University
83, Astrakhanskaya Str., Saratov, 410012
RUSSIAN FEDERATION
sidorovsp@info.sgu.ru

Abstract: - The main subject of our research is the characteristics and features of the economic and finance news flow. In this paper we construct company co-mentions network as a graph in which nodes serve as the world's largest companies mentioned in financial and economic news flow. We link two nodes if two companies were mentioned in the same news item. We construct company co-mention networks for 72 consecutive monthly periods to analyze the dynamics of the structural properties of the company co-mentions network over time. These structural properties are examined based on different graph characteristics such as the distribution of the degrees of the vertices in this graph as well as maximum clique and maximum independent sets sizes. Some conclusions are derived with respect to the dynamics of the evolution of the company co-mentions network over time.

Key-Words: - graph properties; social networks; degree distribution; market graph

1 Introduction

The paper studies structural characteristics of news flows generated by news agencies, enterprises, organizations, social networks, etc. The news flow consists of an enormous amount of news items released in real time by a huge supply of news sources and exhibits unstructurability and high frequency (thousands of news items per second). News flow also includes SEC reports, court documents, reports of various government agencies, business resources, company reports, announcements, industrial and macroeconomic statistics.

Providers of news analytics data such as Thompson Reuters and Raven Pack collect data from different sources including news agencies and social media (blogs, social networks, etc.) and process such data in real time [23, 24]. The news analytics data enables us to study some research problems. One of such problems is the analysis of company co-mention network, which has been addressed in [27, 28]. In the company co-mention network each company is presented as a node, and news mentioning two companies establishes a link between them.

This paper focuses on the analysis of the company co-mention network dynamics over time. We construct company co-mention networks for 72 consecutive monthly periods to analyze the dynamics of the structural properties of the company co-mentions network. Degree distribution, maximum clique size and the size of largest independent set are considered as the important structural characteristics of the company co-mention network. In our opinion, these parameters are the essential graph attributes and give insight into the news flow internal structure. The examination of graph properties gives new understanding of the news flow internal structure. It is shown that the power-law structure of the co-mention graph is fairly stable. Unlike real social graphs, the company co-mention network displays power-low distribution of degrees with non-typical indicators of degree exponent. Therefore, it can be outlined that the concept of 'self-organized network' may be employed for the news flow, and the news market can be viewed as a 'self-organized' system. Another important fact is that the maximum clique increases under crisis, and it supports a widely discussed idea about global economy. It turned out that in most cases the companies which had appeared in maximum clique earlier also emerged in later periods. All maximum cliques include a lot of companies from banking sector.

2 Definitions and Notations

2.1 Cliques and Independent Sets

The graph $G = (V, E)$ is said to be connected if for any two vertices from the set V there is a path from one to another. If the graph is disconnected, it may be subdivided into several connected subgraphs (the connected components of G).

Let $G(S)$ denote the subgraph induced by S for a given subset $S \subseteq V$. Let us remind that a subset $C \subseteq V$ is called a clique if $G(C)$ is a complete graph. The maximum clique problem is the problem of finding the largest clique in a graph.

A subset $I \subseteq V$ is called an independent set if the subgraph $G(I)$ has no edges. The maximum independent set problem is to find the largest independent set in a graph and can be easily expressed as the maximum clique problem in the complementary graph $\tilde{G}(V, \tilde{E})$. Clearly, a maximum clique in $\tilde{G}$ is a maximum independent set in G, so the maximum clique and maximum independent set problems can be easily reduced to each other.

Independent sets can be considered as sets of objects that diverge from every other object in the set, and this knowledge can be important in some applications. Finding a maximum clique would give us the maximum possible size of the sets of "related" items, while discovery of a largest independent could provide us with the maximum possible size of the groups "different" objects in the network.

It is well-known that the maximum clique problem and the maximum independent set problem are NP-hard [14]. Moreover, papers [4, 16] states that these problems are difficult to approximate and therefore it makes these problems peculiarly demanding in large graphs. However, a sparse structure of the co-mention graph allows us to find the exact solution of the maximum clique problem.

The maximum clique problem arises in practical applications in numerous fields. Some of these practical problems may be directly stated as a maximum clique problem and in many cases can be reduced to a problem which requires to find a maximum clique. Real-world applications of the maximum clique problem appear in signal processing, classification theory, coding theory, computer vision, economics, finance, information retrieval, signal transmission theory, aligning DNA and protein sequences, social network analysis, and many other particular areas. Some of these applications can be found in [9, 10, 13, 15, 17, 19, 20, 29-31], among many others.

2.2 Degree Distribution

The degree of a vertex is the number of edges incident to the vertex. Let k be an integer number and let $n(k)$ be the number of vertices with the degree k. Then the probability that a vertex has the degree k is $P(k) = n(k)/n$, where $n = |V|$ is the total number of vertices in the graph G. The function $P(k)$ is called the degree distribution of the graph. The degree distribution is an important characteristic of a graph representing a dataset.

It is well-known that real graphs that appear in various areas (such as medicine, biology, economics, finance, sociology, web, telecommunications,) display the degree distribution that follows the power-law model [2, 3, 7, 11, 21, 25]. This model states that the probability that a vertex has degree k asymptotically follows

$$P(k) \propto k^{-\gamma},$$

i.e. it shows that this function has a linear dependence in the logarithmic scale:

$$\log P(k) \propto -\gamma \log k.$$

An important characteristic of this model is its so-called scale-free property. It implies that the fractal structure of a network remains consistent despite its evolution over time [5].

3 The Evolution of the Company Co-mention Network

We assume that a company is connected with the companies that were mentioned in one news item along with the company. In this type of network, the company will be the "node" or "vertex" of the graph, and the link indicates the relationship between the nodes. Thus, we treat the companies co-mention network as an undirected weighted graph. In some sense, the companies co-mention network can be viewed as a social network. Based on available data from news analytics, we have constructed an adjacency matrix that represents the relationship between companies in accordance with the approach described in [27, 28].

Thompson Reuters and Raven Pack are two well-known providers of news analytics. News analytics providers handle preliminary analysis of each news item in real time. Using AI algorithms, they calculate news-related expectations (sentiments) based on the current market situation. As a rule, providers of news analytics deliver to subscribers in real time the following attributes for each news item: time stamp, company name, company id, relevance of the news, event category, event sentiment, novelty of the news, novelty id, composite sentiment score of the news, among

others. Subscribers of news analytics data may develop and exploit quantitative models or trading strategies based on both the news analytics data and financial time series data. The survey of applications for news analytics tools can be found in books [23, 24]. In recent years news analytics tools have been developed and used in social network analysis [6, 18, 22, 26].

Thus, our analysis of the companies' co-mention network proceeds as follows.

1. We use data of the news analytics providers. Our analysis deals with all financial and economic news published during 6 years from January 1, 2005 till December 31, 2010 (72 months).

2. Then we execute the data cleansing process and we delete all news items which was released at starts and ends of exchange trading sessions, as well as news items containing analytical reports with table materials. In total, the cleansed data set contained more than 8,550,000 news items over 6-year period. The news flow intensity stayed relatively stable within the period.

3. We split 6-year interval into 72-month intervals.

4. For each time interval we calculate the number of co-mentions (weight of the link) for every couple of companies, mentioned together at least in one piece of news; if the companies were not co-mentioned in the period, the weight of the link is 0.

5. We form symmetric co-mention matrices for each time interval using these weighed calculations of the collective companies' mentions.

6. After calculating the adjacency matrix for 1500 the most co-mentioned companies it is turned out that about one third of the rows are filled with zeros (for any period of one month). Therefore, we restricted the analysis to 500 the most co-mentioned companies. Thus, we included in our analysis only data about 40 percent of co-mentions (690 thousand out of 1,790 thousand).

7. We analyze the evolution of these co-mention matrices over the time, the results are being visualized and interpreted.

Fig. 1 shows the dynamics of the degree exponent from January, 2005 to December, 2010. The evolution of the company co-mentions graph shows that the degree exponent is quite stable and the company co-mentions graph follows power-low distribution. The values of the degree exponent are always between 1.3 and 1.6. Note that for many real networks the values of the degree exponent lie between 2 and 3.

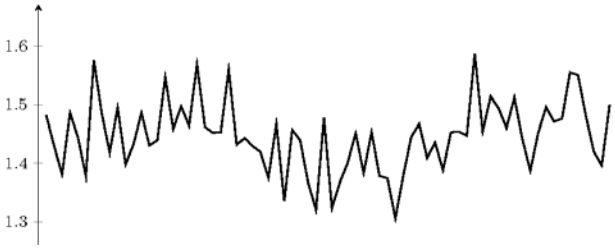

Fig. 1. Monthly dynamics of the degree exponent γ

It should be noted that the degree exponent has their lowest value at the beginning of the financial crisis of 2007–2008. The distribution of the degree of the graph represents the general characteristics of the news flow. The results presented in Table 1 lead us to the conclusion that the global news structure is fairly stable over time.

Table 1. Characteristics of the company co-mentions graph, 2005-2010

Period	Edge density, %	Size of max. clique	Degree exponent γ	Size of max. indep. sets
2005, Jan-Jun	0.202	5	1.43	319
2005, Jul-Dec	0.208	6	1.47	308
2006, Jan-Jun	0.208	5	1.48	306
2006, Jul-Dec	0.207	6	1.49	304
2007, Jan-Jun	0.27	7	1.43	283
2007, Jul-Dec	0.269	7	1.4	283
2008, Jan-Jun	0.267	7	1.4	287
2008, Jul-Dec	0.232	7	1.39	299
2009, Jan-Jun	0.212	6	1.43	300
2009, Jul-Dec	0.194	6	1.5	307
2010, Jan-Jun	0.193	6	1.45	313
2010, Jul-Dec	0.178	6	1.48	320

Results show that the degree distribution is stable over all considered time periods, and it follows a power law. Fig. 2 presents the degree distributions (in the log-log scale) for some instances of the co-mention graph corresponding to different time periods. It can be seen that these plots can be well approximated by straight lines, which means that they represent the power-law distribution.

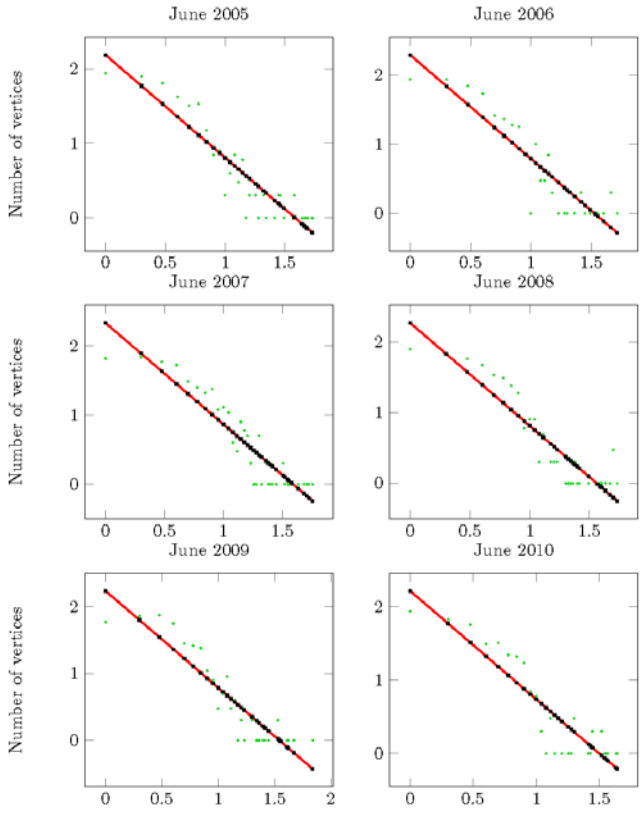

Fig. 2. Degree distributions of the co-mention graph for some periods

The second problem we address in the paper is the analysis of the evolution of the size of the maximum clique of graphs during 6 years from January 1, 2005 till December 31, 2010 (72 months). It follows from the definition of the clique that it is a set of fully interconnected vertices, and therefore each stock that refers to the clique is firmly connected with all the other stocks in this clique. Thus, a stock is associated with a given group if and only if it is co-mentioned with all other stocks in this group. Stocks in one clique are more inclined to have common characteristics of economic and financial activities, and the fact that they are mentioned in joint news is a consequence of their financial and economic associations in the real world. These economic and financial relationships between companies can appear due to different reasons, for example,

- if the log returns of company assets are correlated;
- if there are common members of the board of directors in both companies;
- if one of the companies makes investments in another;
- if one of the companies consumes services or goods of another company;
- if one of the companies supplies goods or services to another company.

Since the size of the maximum clique represents the largest possible group of similar objects, it can be considered as an important characteristic of the co-mention graph.

In this paper one of the variants of the Bron–Kerbosh algorithm proposed in [8] was employed to find the explicit maximum clique. The Bron–Kerbosh algorithm recursively solves sub-problems detailed by three sets of vertices:

- the vertices that must be embodied in the given clique,
- the vertices that should be excluded from the clique,
- and some remaining vertices whose status remains unknown.

This algorithm is proved to be competent and powerful for appropriately sparse networks. Note that our networks of co-mentions are highly sparse. An accurate description of the algorithm is presented in the paper [12].

On the other hand, the problem of detection of the maximum independent sets for our sparse co-mention networks is a much more computationally complex task, in view of high density of the complimentary graph for the company co-mention network. By this reason, to settle this problem we employ the semi-external greedy randomized adaptive search procedure proposed in [1]. It is known that the algorithm works quickly enough but it not necessary may obtain exact solutions.

Table 1 presents the values of index edge density during the period under consideration. The results show that company co-mention networks are sparse. On the contrary, the sizes of independent sets at each period of time are large and are almost constant over time (Table 1). It implies that interconnections of a large amount of companies are not presented in the news flow.

The dynamics of the size of the maximum clique for the co-mention network are shown in Fig. 3. It can be seen that the size of the maximum clique attained its largest value at the climax of the financial crisis (September, 2008). It should be noted that at each period the algorithm finds several different cliques with the common maximum size. Table 2 presents the cliques with a largest value of the degree centrality among all cliques having the same maximum size, for 12 (of 72) different periods of time. It is not surprise that some companies listed in the maximum cliques in distinct periods of time. The cliques presented in Table 2 display that maximum cliques include a significant amount of companies from the financial sector such as

JPMorgan Chase & Co. (JPM), UBS Group AG (UBSN), Credit Suisse Group AG (CSGN), The Goldman Sachs Group, Inc. (GS), Deutsche Bank Aktiengesellschaft (DBK), Morgan Stanley (MS), Citigroup Inc. (C). We can conclude that the important finance and banking companies are regularly presented among the groups of the most co-mentioned companies during the considered periods of time.

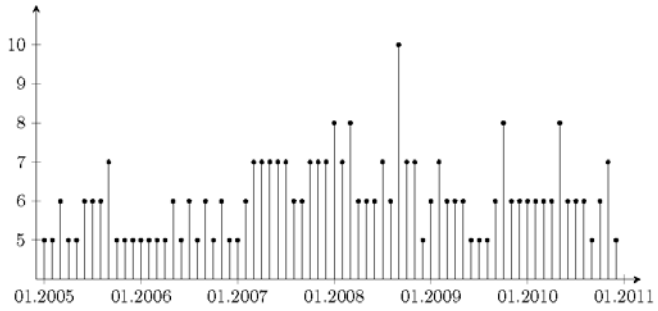

Fig. 3. Monthly dynamics of the size of maximum cliques

4 Conclusion

In this paper we transform news analytics data into the company co-mentions graph. The examination of the graph properties gives new understanding of the news flow internal structure. We investigated the dynamics and changes of the company co-mentions graph structural properties over time. As a result, several interesting conclusions were made. It was shown that the power-law structure of the co-mention graph is fairly stable. Moreover, unlike real social graphs, the company co-mention network displays power-low distribution of degrees with non-typical coefficients of degree exponent. Therefore it can be outlined that the concept of 'self-organized network' may be employed for the news flow, and the news market can be viewed as a 'self-organized' system. Another important fact is that the maximum clique increases under crisis and it supports a widely discussed idea about global economy. It turned out that in most cases the companies, which had appeared in clique earlier, emerged again in later periods. All maximum cliques include a lot of companies from banking sector.

Acknowledgments.
This work was supported by the Russian Fund for Basic Research, project 18-37-00060.

Table 2. Maximum central cliques for different time periods

Period	Company included into maximum central clique
Sep. 2005	C-Corporation, JPMorgan Chase, Urban Select Capital Corporation, Goldman Sachs Group, Deutsche Bank Aktiengesellschaft, Mouser Electronics, Der Deutsche Richterbund
Sep. 2006	C-Corporation, Urban Select Capital Corporation, Goldman Sachs Group, Morgan Stanley Bank, Deutsche Bank Aktiengesellschaft, Telstra Corporation Limited
Mar. 2007	Modern Home Products, C-Corporation, JPMorgan Chase, Urban Select Capital Corporation, Credit Suisse Group AG , Bank of America Corporation, Deutsche Bank Aktiengesellschaft
Oct. 2007	Bank of America Corporation, NOKIA, Alphabet Inc., Eli Lilly and Company, C-CorporationOF, Gilead Sciences, Inc., The Hershey Company
Jan. 2008	US/MOT, Delta Air Lines, eBay Inc., United Technologies Corporation, C-CorporationOP, General Dynamics Corporation, Southwest Airlines Co., C-CorporationOF
Mar. 2008	Federal National Mortgage Association, Boeing Co, US/MOT, Wells Fargo Advantage Funds, Eli Lilly and Company, MBIA Inc., Amgen Inc., Northrop Grumman Corp.
Jul. 2008	Modern Home Products, C-Corporation, JPMorgan Chase, Moran Environmental Recovery, Bank of America Corporation, Federal National Mortgage Association, US/FRE
Sep. 2008	Modern Home Products, C-Corporation, JPMorgan Chase, Urban Select Capital Corporation, Credit Suisse Group AG , Moran Environmental Recovery, Morgan Stanley Bank, Lehman Brothers, American International Group, Weibo Corporation
Feb. 2009	Modern Home Products, Urban Select Capital Corporation, Credit Suisse Group AG , Morgan Stanley Bank, Lehman Brothers, ND Software Co Ltd, Wells Fargo & Company
Oct. 2009	Modern Home Products, C-Corporation, JPMorgan Chase, Urban Select Capital Corporation, Credit Suisse Group AG , Bank of America Corporation, ND Software Co Ltd, Wells Fargo & Company
May 2010	C-Corporation, JPMorgan Chase, Urban Select Capital Corporation, Credit Suisse Group AG , Morgan Stanley Bank, Bank of America Corporation, Deutsche Bank Aktiengesellschaft, Barclays Group
Nov. 2010	Modern Home Products, JPMorgan Chase, Urban Select Capital Corporation, Credit Suisse Group AG , Goldman Sachs Group, Morgan Stanley Bank, Bank of America Corporation

References:

[1] Abello, J., Pardalos, P.M., Resende, M.G.C.: On maximum clique problems in very large graphs. In: *External Memory Algorithms*. pp. 119–130. American Mathematical Society (1999)

[2] Albert, R., Barabasi, A.L.: Statistical mechanics of complex networks. *Reviews of Modern Physics* 74, 47–97 (2002)

[3] Albert, R.: Scale-free networks in cell biology. *Journal of Cell Science* 118, 4947–4957 (2005)

[4] Arora, S., Safra, S.: Approximating clique is NP-complete. In: *Proceedings of the 33rd IEEE symposium on foundations on computer science*. pp. 2–13 (1992)

[5] Barabási, A.L., Albert, R.: Emergence of scaling in random networks. *Science* 286(5439), 509–512 (1999)

[6] Batrinca, B., Treleaven, P.C.: Social media analytics: a survey of techniques, tools and platforms. *AI & SOCIETY* 30(1), 89–116 (Feb 2015)

[7] Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., Hwang, D. U.: Complex networks: Structure and dynamics. *Physics Reports* 424, 175–308 (2006)

[8] Bron, C., Kerbosch, J.: Algorithm 457: Finding all cliques of an undirected graph. *Commun. ACM* 16(9) (Sep 1973)

[9] Brown, M.L., Donovan, T.M., Mickey, R.M., Warrington, G.S., Schwenk, W.S., Theobald, D.M.: Predicting effects of future development on a territorial forest songbird: methodology matters. *Landscape Ecology* 33(1), 93–108 (2018)

[10] Daron, A., Kostas, B., Asuman, O.: Dynamics of information exchange in endogenous social networks. *Theoretical Economics* 9(1), 41–97 (2014)

[11] Dorogovtsev, S.N., Mendes, J.F.F.: Evolution of networks. Adv. Phys 51, 1079 (2002)

[12] Eppstein, D., Löffler, M., Strash, D.: Listing all maximal cliques in sparse graphs in near-optimal time. *CoRR* abs/1006.5440 (2010)

[13] Eppstein, D., Löffler, M., Strash, D.: Listing all maximal cliques in large sparse real-world graphs. *J. Exp. Algorithmics* 18, 3.1:3.1–3.1:3.21 (2013)

[14] Garey, M.R., Johnson, D.S.: *Computers and Intractability; A Guide to the Theory of NP-Completeness*. W. H. Freeman & Co., New York, NY, USA (1990)

[15] Gendreau, M., Picard, J.C., Zubieta, L.: An efficient implicit enumeration algorithm for the maximum clique problem. In: Eiselt, H.A., Pederzoli, G. (eds.) *Advances in Optimization and Control*. pp. 79–91. Springer Berlin Heidelberg, Berlin, Heidelberg (1988)

[16] Hástad, J.: Clique is hard to approximate within $n^{(1-\varepsilon)}$. In: *Acta Mathematica*. pp. 627–636 (1996)

[17] Kalyagin, V., Koldanov, A., Koldanov, P., Pardalos, P., Zamaraev, V.: Measures of uncertainty in market network analysis. *Physica A: Statistical Mechanics and its Applications* 413, 59–70 (2014)

[18] Khan, W., Daud, A., Nasir, J.A., Amjad, T.: A survey on the state-of-the-art machine learning models in the context of nlp. *Kuwait Journal of Science* 43(4), 95–113 (2016)

[19] Kremnyov, O., Kalyagin, V.A.: Identification of cliques and independent sets in pearson and fechner correlations networks. In: Kalyagin, V.A., Koldanov, P.A., Pardalos, P.M. (eds.) *Models, Algorithms and Technologies for Network Analysis*. pp. 165–173. Springer International Publishing, Cham (2016)

[20] Latyshev, A., Koldanov, P.: Investigation of connections between pearson and fechner correlations in market network: Experimental study. In: Kalyagin, V.A., Koldanov, P.A., Pardalos, P.M. (eds.) *Models, Algorithms and Technologies for Network Analysis*. pp. 175–182. Springer International Publishing, Cham (2016)

[21] Lofdahl, C., Stickgold, E., Skarin, B., Stewart, I.: Extending generative models of large scale networks. Procedia Manufacturing 3(Supplement C), 3868 – 3875, *6th International Conference on Applied Human Factors and Ergonomics (AHFE 2015) and the Affiliated Conferences*, AHFE 2015

[22] Manaman, H.S., Jamali, S., AleAhmad, A.: Online reputation measurement of companies based on user-generated content in online social networks. *Computers in Human Behavior 54(Supplement C)*, 94 – 100 (2016)

[23] Mitra, G., Mitra, L. (eds.): *The Handbook of News Analytics in Finance*. John Wiley & Sons (2011)

[24] Mitra, G., Yu, X. (eds.): *Handbook of Sentiment Analysis in Finance* (2016)

[25] Newman, M.E.J.: The structure and function of complex networks. *Siam Review* 45, 167–256 (2003)

[26] Schuller, B., Mousa, A.E., Vryniotis, V.: Sentiment analysis and opinion mining: on optimal parameters and performances. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 5(5), 255–263 (2015)

[27] Sidorov, S.P., Faizliev, A.R., Balash, V.A., Gudkov, A.A., Chekmareva, A.Z., Anikin, P.K.: Company co-mention network analysis. *Springer Proceedings in Mathematics and Statistics* 247, 341-354 (2018)
[28] Sidorov, S.P., Faizliev, A.R., Balash, V.A., Gudkov, A.A., Chekmareva, A.Z., Levshunov, M., Mironov, S.V.: QAP analysis of company co-mention network. In: Bonato, A., Prałat, P., Raigorodskii, A. (eds.) *Algorithms and Models for the Web Graph.* pp. 83–98. Springer International Publishing, Cham (2018)
[29] Vizgunov, A., Goldengorin, B., Kalyagin, V., Koldanov, A., Koldanov, P., Pardalos, P.M.: Network approach for the russian stock market. *Computational Management Science* 11(1), 45–55 (2014)
[30] Wu, Q., Hao, J.K.: Solving the winner determination problem via a weighted maximum clique heuristic. *Expert Syst. Appl.* 42(1), 355–365 (2015)
[31] Zhai, J., Cao, Y., Yao, Y., Ding, X., Li, Y.: Coarse and fine identification of collusive clique in financial market. *Expert Systems with Applications* 69, 225–238 (2017)