Applying co-mention network analysis for event detection

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

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

Abstract: - This paper studies some characteristics and features of the economic and finance news flow. We consider a company co-mentions network as a graph in which nodes serve as the world's largest companies mentioned in financial and economic news flow. Two companies (nodes) are linked if they were mentioned in the same news item. First, we analyze the dynamics of the structural properties of the company co-mentions network over time. Then we propose new method for event detection based on the company co-mention network. The idea behind the method is that more significant news should attract more attention and lead to an increase in the intensity of the news flow. A change in the intensity of co-mentions can be interpreted as a signal or marker of unexpected phenomena that may affect a relatively narrow or a wide range of economic actors. The analysis performed in the paper suggests that the decomposition of the co-mention matrices can be used to separate news signals. News corresponding to the stable part of the graph appear more often; respectively, they carry less information. The unexpected news revealed by the method described this paper deserves special consideration when making financial and investment decisions. The proposed approach to the selection of the event part can be used in the development of algorithms for detecting new events in the financial and economic sphere.

Key-Words: - finance and economics networks; degree distribution; market graph; event detection

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 [29, 30]. 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 [33, 34]. In the company co-mention network each company is presented as a node, and news mentioning two companies establishes a link between them.

The first part of the paper is focused on the analysis of the company co-mention network dynamics over time. Note that the analysis of fluctuations of the news flow and the decomposition of the time series into components may be of interest. News agencies perform the function of aggregating disparate information from a variety of sources. Some of the information, for example, reports, balance sheets, analytical reviews, etc is published on a regular basis. Large companies make many repetitive operations in time, purposefully disclose data on their activities, which is reflected in the news. Our analysis allows us to conclude that some part of the news flow is fairly stable, since it is reproduced from period to period.

The second part of the paper proposes an approach to the detection of new events in the financial and economic sphere. In recent years, one of the most extensive research topics in social media and news analysis has been event detection [2,13,20,23,24,35,38]. It should be noted that a significant portion of the news is characterized the events that were not expected to occur. In this case, each participant in the financial infosphere in its own way assesses the relevance and significance of the newly received information, then the news is

filtered out and ranked by their importance. The publication of the news can be interpreted as the reaction of a particular agent to the occurrence of a particular episode, fact or event which, in his opinion, deserves to be reflected in financial reports. More significant news should attract more attention and lead to an increase in the intensity of the news flow. A change in the intensity of co-mentions can be interpreted as a signal or marker of unexpected phenomena that may affect a relatively narrow or a wide range of economic actors. Some of the news is of industry-specific or regional character. The more significant the event, the wider the range of companies that are mentioned in the emerging news reports. Thus, the simultaneous fluctuation of the news flow for a wide range of companies can be used as an estimation of the significance of the event that caused the news shock.

Some of our results regarding the characteristics of the news flow are published in [6]. In this article, we focus on the co-mentions in the information sources of the 500 largest companies according to the data for 2005-2010.

2 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 [33, 34].

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 provide 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 [29, 30]. In recent years news analytics tools have been developed and used in social network analysis [8, 22, 28, 32].

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 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.

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 as 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 [3, 4, 9, 14, 27,

31]. 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 [7].

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.

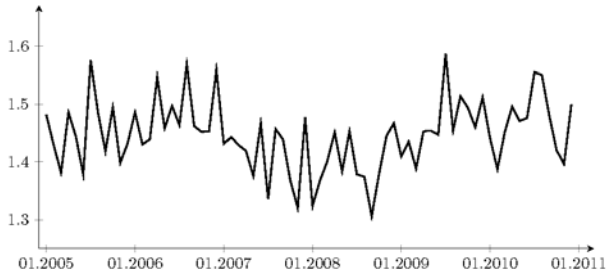

Fig. 1. Monthly dynamics of the degree exponent γ

It should be noted that the degree exponent have 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. Algorithms proposed in [1,10,15] were used to find the sizes of maximum cliques and the sizes of maximum independent sets.

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.

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

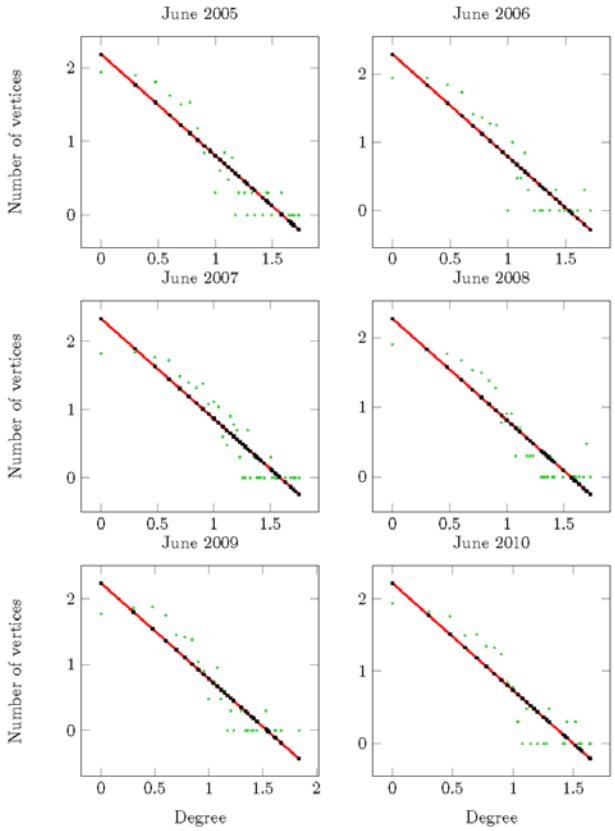

Fig. 2. Degree distribution of the co-mention graph for some periods

3 Applying co-mention network analysis for event detection

We denote $c_{ij}(t)$ the number of co-mentions of companies i and j in the period *t.* The elements of such a matrix can be considered as an adjacency matrix of an undirected graph, the vertices of which correspond to the companies, and the edges correspond to the connections between the vertices. The number of co-mentions determines the weight of the edge in the graph. The number of edges emanated from the vertex and the sum of the weights of these edges as of December 2010 are given in Table 2.

Table 2. Companies with the highest number of co-mentions in December 2010

Company	The number of co-mentioned companies	Company	The number of co-mentions
US/JPM	151	US/BAC	1913
CH/UBSN	146	US/JPM	1759
US/GS	138	US/MS	1505
US/BAC	131	US/MHP	1492
US/MS	128	US/C	1406
US/MHP	120	CH/UBSN	1389
US/C	117	US/GS	1345
GB/BARC	114	HK/0388	1277
CH/CSGN	111	US/CME	1184
DE/DBK	105	GB/RIO	1161

Choosing a time step Δt, we can obtain a sequence of adjacency matrices $C(t) = \{c_{ij}(t)\}, y = 1,2,\dots,T$ and corresponding graphs for a series of periods, which can also be interpreted as a multidimensional time series. Note that the weights of the edges in the adjacency graphs corresponding to two consecutive periods are fairly stable over time. Figure 3 shows the values of the Spearman's rank correlation coefficient between elements of adjacency matrices in two consecutive months. The correlation coefficient ranges from 0.35 to 0.48. But, if we limit the analysis to a hundred of the largest companies, the correlations are manifested significantly stronger.

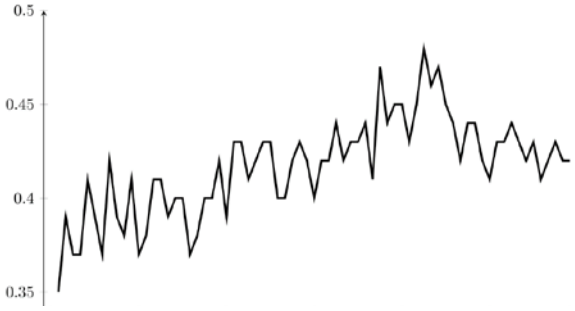

Fig. 3. The values of the correlation coefficient between the matrices of co-mentions in two consecutive months from January 2005 to December 2010

Assume that it is possible to decompose $c(t)$ into separate elements

$$C(t) = C^P(\mathrm{t}) + C^E(t) + U(t),$$

where $C^p(t)$ is reproducible part from period to period (stable) of the co-mention graph; $C^E(t)$ is component, reflecting the impact of events of the period t; $U(t)$ is random component.

We say that some edge is included in the stable part of the graph $C^P(t)$, if, firstly, the probability of its implementation in the period t exceeds the specified level δ, $(0 < \delta < 1)$:

$$P\big(c_{ij}(t) > 0\big) > \delta,$$

and, secondly, the observed of link weight lies within the confidence interval for the average over T the periods of the level:

$$P\big(|c_{ij}(t) - \bar{c}_{ij}| < \Delta\big) = \gamma,$$

where Δ denotes the width of the confidence interval corresponding to the selected level of reliability γ.

In the calculation results below, we used the values $\delta = 0.8$; $\gamma = 0.9$.

In the stable part of the graph, significantly fewer edges are presented, but these edges correspond to high weights. At clipping level δ=0.8 the stable part of the graph reflects approximately two thirds of the total number of news. Table 3 provide an illustration of the stable part of the graph highlighted in this way.

Table 3. The number of edges in the stable part of the co-mention graph for the 20 largest companies

100 percent periods		50 percent periods	
US/MER	15	US/C	118
US/JPM	14	US/JPM	109
DE/DBK	13	US/GS	99
CH/UBSN	12	CH/UBSN	88
US/C	11	US/MS	78
US/GS	10	CH/CSGN	69
US/MS	10	GB/HSBA	68
CH/CSGN	9	US/BAC	68
GB/BARC	9	DE/DBK	66
US/BAC	9	US/MER	66
AU/RIO	8	US/GM	59
GB/HSBA	8	US/MSFT	58
GB/RIO	8	US/MHP	53
AU/BHP	7	AU/RIO	44
US/MHP	7	GB/BARC	40
AU/CBA	5	AU/MBL	39
FR/13110	5	FR/13110	38

Some of the edges of the graph are present only in certain periods. The greater the frequency of mentioning the same number of companies in a limited period of time, the higher the significance of the event or process reflected in the news, which occurs within a limited time period. The problem of detecting significant events (News Event Detection) is widely discussed in the literature of the subject [2,13,20,23,24,35,38]. Usually, a newly received text message is analyzed for similarities with already classified documents and related to certain events. With sufficient similarity, the text corresponds to a particular event, otherwise it is considered as a reflection of a new event. Wherein the time lag between the message and the event can be counted. The more time has passed since the event was highlighted and the message, the less similarity there is. For the analysis and clustering of messages, various variations of cluster analysis methods are used. Next, the task is to identify keywords and a common set of terms characterizing the news, the intensity of the discussion and the lifetime of the event.

However, unlike the analysis of mass media and messages in social networks, we focus not on the description of terms characterizing the event but on the range of companies that are affected by the event. To highlight the components of the graph $C^E(t)$, which can be interpreted as a reaction of the news to some events $E(t)$, we used the following rule. The graph edge belonged to the subset $C^E(t)$, if the edge weight exceeded the value qth quantile for empirical distribution $c_{ij}(t)$ for all periods

$$c_{ij}(t) > c_{ij}^{q}.$$

Subgraphs selected in this way respond well to meaningful interpretation. In quiet periods, the event part contains a small number of vertices and edges. The most significant event part corresponds to the 2008 financial crisis. During this period, the anomalous number of co-mentions was recorded for more than a hundred and fifty of the largest companies. Example of subgraphs for quiet and anomalous periods are shown in Figure 4.

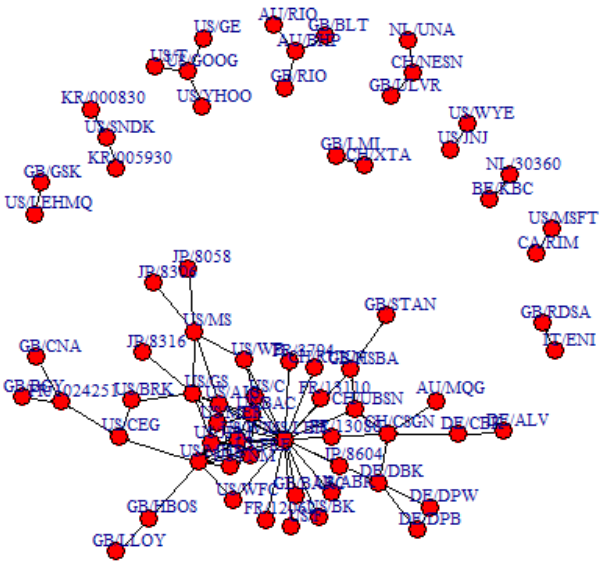

Fig. 4. The event part of the co-mention graph as of September 2008

3 Conclusion

In this paper we transform news analytics data into the company co-mentions graph. The examination of 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.

The analysis performed suggests that the decomposition of the co-mention matrices can be used to separate news signals. News corresponding to the stable part of the graph appear more often; respectively, they carry less information. The unexpected news revealed by the method described this paper deserves special consideration when making financial and investment decisions. The proposed approach to the selection of the event part can be used in the development of algorithms for detecting new events in the financial and economic sphere.

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

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] Aggarwal, C.C.: Mining text and social streams: A review. *SIGKDD Explor. Newsl.* 15(2), 9-19 (Jun 2014)

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

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

[5] 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)

[6] Balash V.A., Chekmareva A, Faizliev A.R., Sidorov S.P., Mironov S.V. and Volkov D.: Analysis of news flow dynamics based on the company co-mention network characteristics. *Lecture Notes in Engineering Science*. In Press

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

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

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

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

[11] 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)

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

[13] Dong, X., Mavroeidis, D., Calabrese, F., Frossard, P.: Multiscale event detection in social media. *Data Min. Knowl. Discov.* 29(5), 1374-1405 (Sep 2015)

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

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

[16] 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)

[17] 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)

[18] 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)

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

[20] Huang, Y., Li, Y., Shan, J.: Spatial-temporal event detection from geo-tagged tweets. *ISPRS International Journal of Geo-Information* 7(4) (2018)

[21] 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)

[22] 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)

[23] Kleinberg, J.: Bursty and hierarchical structure in streams. In: *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. pp.

91-101. KDD '02, ACM, New York, NY, USA (2002)

[24] Kolchyna, O., Souza, T.T.P., Treleaven, P.C., Aste, T.: A framework for twitter events detection, differentiation and its application for retail brands. In: *2016 Future Technologies Conference (FTC)*. pp. 323-331 (Dec 2016)

[25] 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)

[26] 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)

[27] 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

[28] 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)

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

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

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

[32] 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)

[33] 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* (2018), in press

[34] 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)

[35] Sidorov S.P., Faizliev A.R., Levshunov M., Chekmareva A., Gudkov A., Korobov E.: Graph-Based clustering approach for economic and financial event detection using news analytics data. In: Staab S., Koltsova O., Ignatov D. (eds) *Social Informatics. SocInfo* 2018. *Lecture Notes in Computer Science*, vol 11186. Springer, Cham: 271-280

[36] 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)

[37] Wu, Q., Hao, J.K.: Solving the winner determination problem via a weighted maximum clique heuristic. *Expert Syst. Appl.* 42(1), 355–365 (2015)

[38] Yang, Y., Pierce, T., Carbonell, J.: A study of retrospective and on-line event detection. In: *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. pp. 28-36. SIGIR '98, ACM, New York, NY, USA (1998)

[39] 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)