Analysis of Strong and Weak Ties in Oil & Gas Professional Community

Fedor Krasnov, Sofia Dokuka, Ilya Gorshkov, Rostislav Yavorskiy

[1] Fedor Krasnov Gazpromneft NTC, St. Petersburg, Russia
Krasnov.FV@gazprom-neft.ru
[2] Sofia Dokuka Center for Institutional Studies, Higher School of Economics, Moscow, Russia sdokuka@hse.ru
[3] Ilya Gorshkov Department of Data Analysis and Artificial Intelligence, Faculty of Computer Science, Higher School of Economics, Moscow, Russia
iagorshkov@edu.hse.ru
[4] Rostislav Yavorskiy Department of Data Analysis and Artificial Intelligence, Faculty of Computer Science, Higher School of Economics, Moscow, Russia
ryavorsky@hse.ru

Abstract. The importance of weak social ties in professional communities is well studied and widely accepted. In our paper we analyze the structure of strong ties based on the co-authorship relation and use the formal concept analysis framework to figure out weak ties. The research is motivated by fast growing need in cross-disciplinary research, which requires experts from different areas to understand the bigger picture and identify potential fellows for collaborative research projects in nearest future.

Keywords: co-authorship graph, strong ties, weak ties, professional network, professional community, research management

1 Introduction

1.1 The nature of the problem

To keep up with state of the art developments in the Oil & Gag industry the engineers have to regularly examine specialized conferences organized under the umbrella of the Society of Petroleum Engineers (SPE)[5]. The built in mechanism for expert selection of materials for these conferences is designed to provide an appropriate level of knowledge, and thus eliminates the need to waste time on publications, which are not of top quality.

All the conferences in the field are divided into regions to represent the regional development in the industry. Besides, the SPE has a rule explicitly stated in every call for papers, according to which the article may

[5] SPE is a not-for-profit professional organization for oil and natural gas exploration and production (E&P) professionals. It was founded in 1957, and today brings together more than 165,000 engineers, scientists, managers, and educators

be submitted to only one conference. Therefore the authors of cross-disciplinary articles have to choose which of the specialized conferences to apply to.

Nowadays the easily accessible hydrocarbon resources have run out, so the industry is focused on hard-to-mine resources (brownfields), which requires an integrated approach and cross-functionality. Therefore, the number of cross-disciplinary articles grows from year to year.

As a result, it may happen for a cross-disciplinary work that the article falls out of focus. On the other hand there is a common theme for all conferences, such as those associated with machine learning and big data. For those interested in topics such specialists should either keep track of all conferences at once, or use automated search engines.

1.2 The research objectives

Our goal is to develop a methodology and tools for automated analysis of a collection of research papers available at the SPE digital library. On the basis of these analyzes one should be able to:

- figure out the most important and relevant research topics,
- assess the influence of different researchers and scientific schools,
- identify strong and weak ties in the professional community,

and use all of these in daily research management process. This paper is focused on the third item in the list. It continues our study of professional communities started in [15, 13, 14, 16, 4, 5, 7].

1.3 Social network analysis

The analysis of social networks of co-authorship has a long history [12]. There are a plenty of studies examining the structure of co-authorship ties within diverse scientific fields and reveal specific collaboration patterns for the different disciplines [1, 3, 6, 19, 24]. Here we intend to uncover weak social ties in the Oil& Gas professional community. This is similar to the task of link prediction in social networks, see e.g. [26].

Weak ties within social networks is one of the key concepts. The idea of the differentiation of ties by their strength was firstly considered by sociologist Granovetter in [11], who empirically showed that weak ties (e.g. ties with not very close friends and relatives) are of a great importance in case of information propagation and knowledge diffusion. In case of Granovetter, weak ties were the source of the important information about working places and vacancies.

The identification of weak ties within a professional community has a great practical importance. Firstly, identification of people who are working on the same topic and substantial research idea is very important for information gathering and knowledge diffusion. Secondly, knowing the social environment, e.g. weak ties within the community can be important in collaboration and cooperation establishment. In this paper we aim to identify the strong and weak ties within the professionals of Oil & Gag industry based on their collaborations which can be inferred from their coauthorships. In this paper we assume that two researches have weak ties if they both work with the same objects or concepts and their research topics are very close to each other.

1.4 Formal concept analysis

Formal concept analysis (FCA) gives a way to analyze collections of objects and their properties. Recall some basic definitions from [**?**]. A *formal context* is a triple $K = (G, M, I)$, where G is a set of objects, M is a set of attributes, and $I \subseteq G \times M$ is a binary relation that expresses which objects have which attributes. Implication $A \to B$ for subsets A, B of the set of attributes M $(A, B \subseteq M)$ holds if $A' \subseteq B'$, i.e. every object possessing each attribute from A also has each attribute from B.

An association rule is an expression of the form $X \to Y$, where $X, Y \subseteq M$ and $X' \subseteq Y'$ may not hold. The strength of an association rule can be measured in terms of its support (denoted by *supp*) and confidence (denoted by *conf*), where

$$supp(X \to Y) = \frac{|(X \cup Y)'|}{|G|}, \quad conf(X \to Y) = \frac{|(X \cup Y)'|}{|X'|}$$

Support determines how often a rule is applicable to a given data set, while confidence determines how frequently items in Y appear in transactions that contain X. See [8] for a detailed introduction to the subject.

In this paper we utilize the FCA framework for studying the author - keyword relationship. For us

- G denotes the set of keywords.
- M stands for the set of all co-authors of the papers.
- $I \subseteq G \times M$ is a binary relation. One has $(g, m) \in I$ if m co-authors a paper for which g is among the keywords.

Then the association rules are interpreted as indicators of connectivity between different research fields, and also used to recognize weak ties between authors of different papers.

The idea to apply FCA in the context of social network analysis is not new. In [17] it was used for collective network analysis. In [25] a combination of Formal Concept Analysis and well-known matrix factorization methods were used to address computational complexity of social networks analysis and the clarity of their visualization. Bi-clustering and tri-clustering were used in [9] to analyze data collected from the Russian online social network VKontakte for extracting groups of users with similar interests, finding communities of users which belong to similar groups, and revealing users interests. FCA was extensively used for analyzing social networks based on co-references, see [18], and detecting criminal networks [22]. For other applications of FCA in social network analysis see [23]. Another rather detailed overview of FCA-based applications for Social Networks Analysis could be found in [21, 20, 2].

The rest of the paper is organized as follows. In the second section we describe the data collection procedures and provide descriptive statistics. In the third section we provide the results for empirical estimation of the data. The fourth section provides a summary of results and some conclusions.

2 Data

Our study is based on materials of annual SPE Russian Oil and Gas Conference and Exhibition 2016. The main features of this event are as follows:

- Multi-disciplinary. The conference presentations, selected on the basic directions of development Oil & Gas industry. These areas are listed below.
- Periodic. This is an annual conference.
- Regional. The majority of the participants represented mainly the Russian companies.
- High selection criteria. The conference acceptance rate is approximately 15%. The selection process is conducted by Subject Matter Experts.
- The conference program consists of four parallel sections.
- At least one co-author must attend the event and present the work.

The data we work on is retrieved from open portal of Society of Petroleum Engineers (SPE) at `http://www.onepetro.org`.

Clean up and preparation of meta information was produced using Python on hybrid cluster at Gazpromneft NTC LLC. Text analysis was done using Python NLTK library. Statistical analysis was performed using SciPy library.

2.1 Features of the collection

The collection comprises 404 articles written by 839 co-authors. It includes papers in the following areas:

1. Well construction drilling and completion.
2. Static and dynamic modeling.
3. Hard-to-recover reserves.
4. Well and formation testing.
5. Field development monitoring and control.
6. Well intervention.
7. Shelf development experience and prospects.
8. Field geophysical survey/well logging.
9. Gas condensate and oil gas condensate field development.
10. Brownfields.
11. Geomechanics.
12. Oil and gas production - equipment and technologies.
13. Cores recovery, examination and analysis

2.2 Structure of the data

In the retrieved data each publication record includes the following information:

- title and abstract of the article;
- the list of authors and their affiliations;
- year of publication.

The most time-consuming step was to prepare the data and make the data set clean and useful. Unfortunately, the portal does not have a directory for authors. As a result sometimes we had up to 6 different spellings of the same name in different articles.

2.3 Strong ties

Almost every paper in the collection is written jointly by a few authors. It usually takes at least several months to write a good paper, so in the context of professional community each publication could be considered as a proof of strong ties between the co-authors.
The descriptive statistics for the co-authorship network is given below.

- Number of nodes: 839
- Number of strong ties: 2315
- Number of connected components: 127
- Size of the largest connected component: 198
- Size of the second largest connected component: 20

2.4 Network visualization

The visualization of co-authorship networks is presented in Fig. 1. This and the other graphs in this paper are produced with yEd Graph Editor [27].
An inspection of the largest connected component shows that it mostly consists of participants of the well established collaborative program between Gazprom subsidiaries and Schlumberger. Otherwise the picture is very typical for a large industrial research conference, where the audience consists of big number of small cliques, which hardly communicate with each other. It shows that authors prefer working within their small community and it is difficult for them to establish new links with other colleagues.
As it was already mentioned above the goal of our work is to help the members of a professional community identify participants with similar interests and then convert weak ties into strong ones by establishing mutually beneficial collaborative research projects.

3 Identification of weak ties

3.1 Heuristics for identifying weak ties

The importance of weak ties is well studied in the literature, see [11, 10]. In this paper we assume that two researchers have weak ties if they both work with the same objects or concepts. We believe that if two persons work on the same substantial problem (e.g. they share same narrow research topic), they should at least know each others' works. We assume these social ties are weak, because they are very much likely to know each other and even communicate, but the intensity of their interactions and communications is very much likely to be low, because they are not involved in joint projects.
The heuristics is implemented in the following way. First, we start from extracting keywords for each paper in order to create a formal context, i.a. object-attribute relation in which objects are words, attributes are authors, and the relation is *"a keyword w is used by an author a"*. Second,

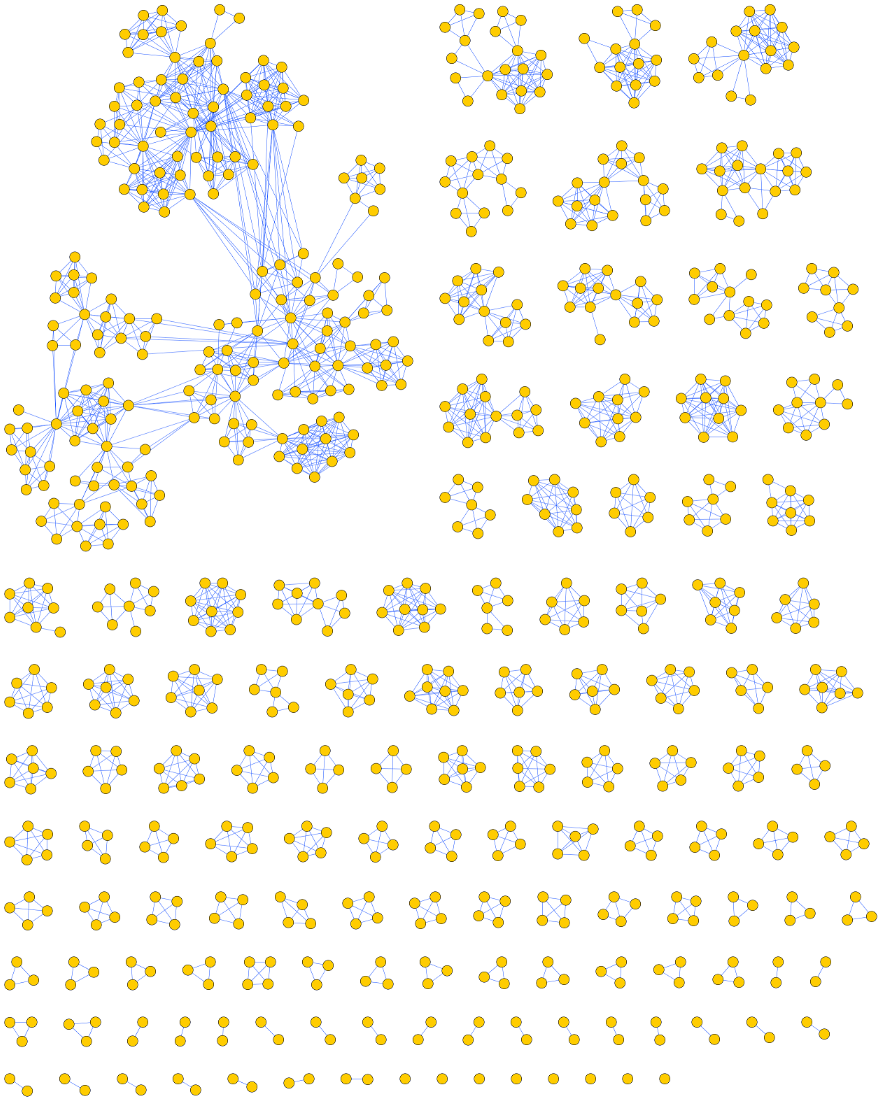

Fig. 1. Visualization of the strong ties in Oil&Gas professional community. Nodes are authors, links correspond to the co-authorship relation. The graph has 839 nodes, 2315 ties, 127 components.

the association rules with high characteristics of support and confidentiality are computed using Concept Explorer tool, see [28, 29].
Finally, for every association rule of the form

$$a_1, \dots, a_m \Rightarrow b_1, \dots, b_k, \tag{1}$$

where $a_1, \dots, a_m, b_1, \dots, b_k$ are author IDs we assume that all members of the joint group $\{a_1, \dots, a_m, b_1, \dots, b_k\}$ are weakly connected.

3.2 Keywords extraction

As it was mentioned above our data set stores titles and abstracts of papers. As these texts are rather small we initially consider all words as equally important.
After the clean up the object-property table has 729 objects (keywords) and 839 attributes (authors).

3.3 Association rules

Table 1 presents several examples of association rules. Each rule has two parts, antecedent and consequent, which are sets of attributes. Support indicates the number of objects, which share these attributes. In our case, support is the number of keywords common for all authors in the set.

Table 1. Examples of the computed association rules. Attributes are authors' IDs, support is the number of common keywords for these authors.

Support	Antecedent attributes	Confidence	Support	Consequent attributes
17	564;825	= 94% ⇒	16	133
16	335;636	= 94% ⇒	15	226;131;542;552
15	131;335	= 100% ⇒	15	226;542;552;636
16	101;436	= 88% ⇒	14	132
15	801;357;510	= 93% ⇒	14	8
15	333;754	= 93% ⇒	14	42;133
6	108;233	= 83% ⇒	5	754

3.4 Identifying weak ties from the association rules

The main idea here is to interpret each association rule as an evidence of common interests for the involved authors. For example, from rule

$$15 \mid 333; 754 \;\; = 93\% \Rightarrow \;\; 14 \mid 42; 133$$

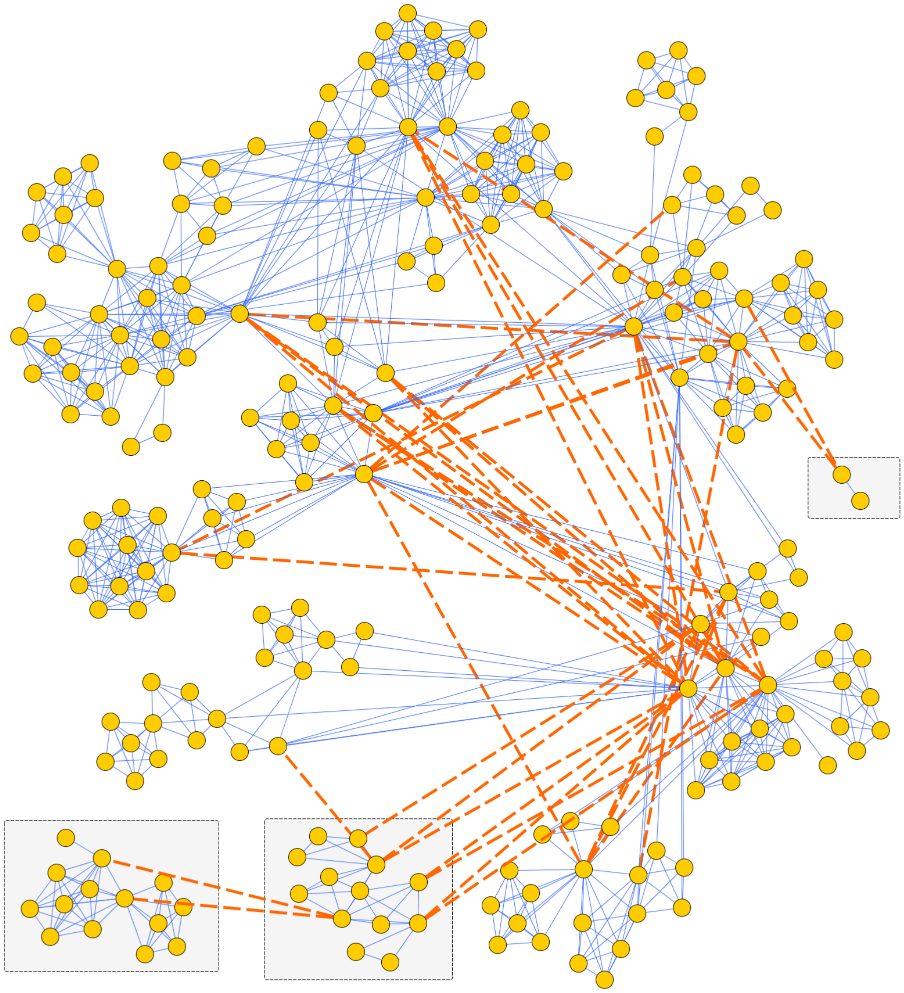

Fig. 2. Visualization of the largest connected component with the weak ties. Nodes are authors, co-authorship relation is represented by blue solid links, the dashed red edges correspond to the weak ties. Grey boxes set out previously disconnected fragments, which get bridged with the weak ties.

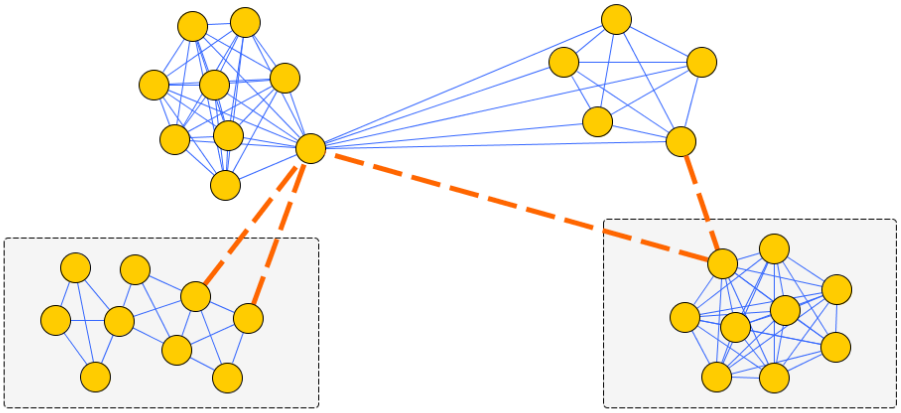

Fig. 3. Second largest connected component with the weak ties (dashed red). Grey boxes are used to set out previously disconnected fragments, which get bridged with the weak ties.

we conclude that members with IDs 333, 754, 42, and 133 work on the close subjects as they use 14 common keywords, so each two of them are considered weakly tied.
In general, each rule of a form

$$s \mid a_1; \ldots ; a_n \ = c\% \Rightarrow \ s' \mid b_1; \ldots ; b_m$$

produces C^2_{n+m} pairwise weak ties within the union set $\{a_1, \ldots, a_n, b_1, \ldots, b_m\}$.

4 Results

For the data set of SPE papers the suggested procedure yielded the following. First, we have got 216 association rules with confidence greater than 80% and support at least 5 objects (keywords). Some of them are listed on Table 1. That resulted in 436 weak links out of which 149 were unique. Finally it turned out that the bigger part of them duplicates some of the existing strong ties and only 46 out of 149 suggest new connections. The network graph with the added weak ties is presented in Fig. 2 and in Fig. 3.
Briefly, most of the isolated islands are not affected and remain isolated.Three cliques got connected to the largest component, see Fig. 2. Another two joined the second largest component, see Fig. 3.
The fact that out of 149 identified weak ties 103 are duplicates of the already established strong ties shows that the suggested heuristic is rather conservative, two thirds of the found connections are certainly relevant. For the remaining new links we rely on expert opinion. To this end visualization in Fig. 4 was used together with the respective table of suggested candidate pairs for collaboration.

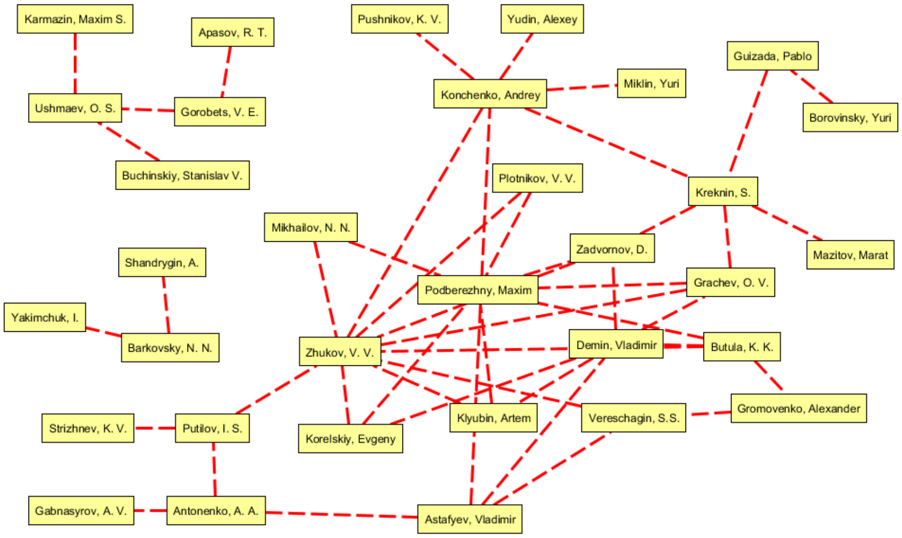

Fig. 4. Graph of the new identified weak ties.

4.1 Conclusion

In this paper we have used Formal Concept Analysis for the identification of weak ties in a social network of co-authorship. This task has a lot of applications, for example identifying colleagues with similar academic and professional interests and aims. The identification of people with similar interests can also significantly improve the mechanism of academic and professional recruiting.
We also believe that the current methodological approach can be reframed for the case of dynamic social networks and identification of weak ties formation and dissolution in a professional community.

5 Acknowledgments

The article was prepared within the framework of the Basic Research Program at the National Research University Higher School of Economics (HSE) and supported within the framework of a subsidy by the Russian Academic Excellence Project '5-100'. We are also thankful to the anonymous referees for their comments and suggestions, which significantly improved the text of this paper.

References

1. Acedo, F.J., Barroso, C., Casanueva, C., Galán, J.L.: Co-authorship in management and organizational studies: An empirical and network analysis*. Journal of Management Studies 43(5), 957–983 (2006)

2. Aufaure, M.A., Le Grand, B.: Advances in fca-based applications for social networks analysis. International Journal of Conceptual Structures and Smart Applications (IJCSSA) 1(1), 73–89 (2013)
3. Barabási, A.L., Jeong, H., Néda, Z., Ravasz, E., Schubert, A., Vicsek, T.: Evolution of the social network of scientific collaborations. Physica A: Statistical mechanics and its applications 311(3), 590–614 (2002)
4. Barysheva, A., Golubtsova, A., Yavorskiy, R.: Profiling less active users in online communities. In: SNAFCA@ ICFCA (2015)
5. Barysheva, A., Petrov, M., Yavorskiy, R.: Building profiles of blog users based on comment graph analysis: The habrahabr. ru case. In: International Conference on Analysis of Images, Social Networks and Texts. pp. 257–262. Springer (2015)
6. Ding, Y.: Scientific collaboration and endorsement: Network analysis of coauthorship and citation networks. Journal of informetrics 5(1), 187–203 (2011)
7. Dokuka, S., Yavorskiy, R., Krasnov, F.: The structure of organization: the coauthorship network case. In: Analysis of Images, Social Networks and Texts. 5th International Conference, AIST 2016, Yekaterinburg, Russia, April 7-9, 2016, Revised Selected Papers. Communications in Computer and Information Science. pp. 93–101. Springer International Publishing (2017)
8. Ganter, B., Stumme, G., Wille, R.: Formal Concept Analysis: foundations and applications, vol. 3626. Springer (2005)
9. Gnatyshak, D., Ignatov, D.I., Semenov, A., Poelmans, J.: Gaining insight in social networks with biclustering and triclustering. In: International Conference on Business Informatics Research. pp. 162–171. Springer (2012)
10. Granovetter, M.: The strength of weak ties: A network theory revisited. Sociological theory 1(1), 201–233 (1983)
11. Granovetter, M.S.: The strength of weak ties. American journal of sociology pp. 1360–1380 (1973)
12. Hou, H., Kretschmer, H., Liu, Z.: The structure of scientific collaboration networks in scientometrics. Scientometrics 75(2), 189–202 (2008)
13. Krasnov, F., Ustalov, D., Yavorskiy, R.: Comparison of online communities on the base of lexical analysis of the news feed. In: Proceedings of 2-nd conference on Analysis of Images, Networks and Texts, Yekaterinburg. pp. 254–257 (2013)
14. Krasnov, F., Vlasova, E., Yavorskiy, R.: Connectivity analysis of computer science centers based on scientific publications datafor major russian cities. Procedia Computer Science 31, 892–899 (2014)
15. Krasnov, F., Yavorskiy, R.: Measurement of maturity level of a professional community. Business Informatics 23(1) (2013)
16. Krasnov, F., Yavorskiy, R.E., Vlasova, E.: Indicators of connectivity for urban scientific communities in russian cities. In: Analysis of Images, Social Networks and Texts, pp. 111–120. Springer (2014)
17. Kurtz, C.F.: Collective network analysis (white paper available at www.cfkurtz.com.) (2009)

18. Kuznetsov, S., Obiedkov, S., Roth, C.: Reducing the representation complexity of lattice-based taxonomies. In: U. Priss, S. Polovina, R. Hill, Eds., Proc. 15th International Conference on Conceptual Structures (ICCS 2007), Lecture Notes in Artificial Intelligence (Springer), Vol. 4604. pp. 241–254. Springer (2007)
19. Newman, M.E.: The structure of scientific collaboration networks. Proceedings of the National Academy of Sciences 98(2), 404–409 (2001)
20. Obiedkov, S., Roth, C.: Social Network Analysis and Conceptual Structures: Exploring Opportunities: Proceedings, Clermont-Ferrand, France, February 2007. Université Blaise Pascal, Laboratoire Limos (2007)
21. Pensa, R.G., Boulicaut, J.F.: Towards fault-tolerant formal concept analysis. In: Congress of the Italian Association for Artificial Intelligence. pp. 212–223. Springer (2005)
22. Poelmans, J., Elzinga, P., Ignatov, D.I., Kuznetsov, S.O.: Semi-automated knowledge discovery: identifying and profiling human trafficking. International Journal of General Systems 41(8), 774–804 (2012)
23. Poelmans, J., Ignatov, D.I., Kuznetsov, S.O., Dedene, G.: Formal concept analysis in knowledge processing: A survey on applications. Expert systems with applications 40(16), 6538–6560 (2013)
24. Rodriguez, M.A., Pepe, A.: On the relationship between the structural and socioacademic communities of a coauthorship network. Journal of Informetrics 2(3), 195–201 (2008)
25. Snasel, V., Horák, Z., Kocibova, J., Abraham, A.: Analyzing social networks using fca: complexity aspects. In: Web Intelligence and Intelligent Agent Technologies, 2009. WI-IAT'09. IEEE/WIC/ACM International Joint Conferences on. vol. 3, pp. 38–41. IET (2009)
26. Wang, P., Xu, B., Wu, Y., Zhou, X.: Link prediction in social networks: the state-of-the-art. Science China Information Sciences 58(1), 1–38 (2015)
27. Wiese, R., Eiglsperger, M., Kaufmann, M.: Yfiles – visualization and automatic layout of graphs. In: Graph Drawing Software, pp. 173–191. Springer (2004)
28. Yevtushenko, S., Tane, J., Kaiser, T.B., Obiedkov, S., Hereth, J., Reppe, H.: Conexp-the concept explorer (2006)
29. Yevtushenko, S.A.: System of data analysis concept explorer. In: Proceedings of the 7th national conference on Artificial Intelligence KII. vol. 2000 (2000)