Network Studies in Russia: From Articles to the Structure of a Research Community

Daria Maltseva and Ilia Karpov

Abstract Our research focuses on the structure of a research community of Russian scientists involved in network studies, which is studied by means of analysis of articles published in Russian-language journals. The direction of network studies in Russia is quite new form of research methodology—however, in recent years we can observe the growing number of scientists working at this direction and institutionalized forms of their cooperation. Studying the structure of these researchers' community is important for the fields development. This paper is the first report on the research, that is why it focuses on methodological issues. It covers the description of method of citation (reference) analysis that we use and the process of data collection from eLibrary.ru resource, as well as presents some brief overview of collected data (based on analysis of 8 000 papers). It is concluded by representation of future steps of the research.

1 Introduction

The development of a certain science discipline in many respects depends not only on institutional context—the official approvement of discipline and presence of organizations engaged in certain type of research—but on the structure of informal (implicit) social and communicational structures of researchers as well. Such system of relations between researchers was developed in the sociology of science by Diana Crane in 1972 building on Derek de Solla Price's work on citation networks and called "invisible college", meaning the informal (implicit) social and communicational structures of researchers, who refer to each other in their publications without being linked by formal organizational ties. The usage of this notion made it possible to find out the presence of some new fields and disciplines.

D. Maltseva (✉) · I. Karpov
International Laboratory for Applied Network Research,
National Research University Higher School of Economics, Moscow, Russia
e-mail: d_malceva@mail.ru

I. Karpov
e-mail: karpovilia@gmail.com

V.A. Kalyagin et al. (eds.), *Models, Algorithms, and Technologies for Network Analysis*, Springer Proceedings in Mathematics & Statistics 197,
DOI 10.1007/978-3-319-56829-4_19

Network study as a form of research methodology, which operates with the notion of networks for studying different social phenomena, was one of the fields, which was recognized as a separate discipline in 1970s in Western sociology. The discipline called "Social network analysis" (SNA) was characterized both by institutionalized forms—journals, conferences, knowledge transfer centers, and educational programs—and the presence of its own professional community (informal "invisible college") [15].

The exclusion of Russia from the context of social sciences, which was typical for the Soviet period, has further led to certain lags in some areas, including network studies. However, during recent years we can observe the growing interest to this new form of a research methodology—the usage of social network analysis technics becomes evident and "fashionable" in Russian scientific space (which can be seen by the increase of journal publications), the appearance of scientists who nominate themselves as "network researchers" and the development of institutionalized forms of their cooperation (e.g., research sections at universities and organizations, laboratories).

However, there is no information on the characteristics of the community of network scientists in Russian-language space, language yet—who are the main drivers of the field's development, if they consider themselves as cooperators, representing "invisible college", or see each other as competitors, if they interact with each other or mostly prefer some "significant others". The literature review of the works of Russian scientists involved into the field of network studies in the sociology shows that different authors regard different—but foreign—scientists as "founding fathers", whose works are important for the filed establishment and development, such as B. Wellman, S. Wasserman, K. Faust, L. Freeman [10], B. Wellman, S. Berkowitz, M. Granovetter, H. White [26], R. Emerson, K. Cook, J. Coleman [18], R. Emerson, K. Cook, L. Molm, M. Emirbayer, anthropologists [7], B. Latour, J. Law, M. Callon [17, 31].

This situation makes it important for the current state of field's development to study the structure of a research community of scientists involved into network studies in Russia—who are these main drivers, how they relate to each other, and at what research teams—Russian or foreign—they are mainly focused. We propose to build the structure of this research community basing on the quantitative method of citation (reference) analysis of articles on "network" topics published in Russian journals in different disciplines, which are provided by the largest electronic library of scientific periodicals in Russian eLibrary.ru.

As many articles presenting the results of citation analysis often do not provide enough information on their methodology to reproduce the study or rationale for methodological decisions [14], in the present article we would like to cover some methodological issues concerning the proceeding study and present the overview of the method of citation (reference) analysis as a tool for studying scientific fields and describe the process of data collection in detail. We propose that such description can be interesting for the scientists who do not have an experience of working with Russian-speaking platform of scientific measuring. Providing the brief

information on already collected data on articles on network topics, we conclude with the description of future steps of the study.

2 Citation Analysis as a Tool for Studying Science

In the first section we present some general information on the method of citation analysis as a special tool used for studying scientific fields. As we are mostly interested in sociology and social network analysis, we provide some examples of the previous studies done in these disciplines using observed tool.

2.1 Citation Analysis as a Method

Citation analysis is a method used in the field of informetrics (more precisely, its subareas bibliometrics and scientometrics) for the study of different forms of social interaction networks, including authors' citation networks, co-citation networks, collaboration structures, and other (look [2] for a detailed review). Citation analysis was established as an instrument of managerial control of modern science, which was institutionalized in the middle of twentieth century and changed the practice of references from concrete names to precisely dated texts [21]. The first well-known usage of a citation analysis is associated with the name of the chemist Eugene Garfield, who developed the first and revolutionary citation index Science Citation Index, SCI, in 1955, as a representative of the Institute for Scientific Information (now Thomson Reuters) [6, 22].

Even though Garfield's works were innovative for the filed, there were some other authors who could be assumed as pioneers of this methodological approach. The first paper that can be considered as a citation analysis was published in 1927 by Gross and Gross, who studied the references found during 1 year's issues of the Journal of the American Chemical Society. According to Casey and McMillan [4], even though the term bibliometrics is dating only to the late 1960s, the field itself has roots reaching back at least 80 years to the work of Lotka published in 1926. Garfield himself, in his highly cited article [9], enumerates names of others citation analysis pioneers such as Bradford, Allen, Cross and Woodford, Hooker, Henkle, Fussler and Brown. De Bells [6] names John Desmond Bernal and Derek John De Solla Price as "philosophical founders of bibliometrics". Other important names, due to de Bells, are sociologist Robert Merton and chemist and historian of science, leading researcher with Garfield's former company Henry Small. From its beginning, now citation analysis has grown into a developed field. During the recent decades, there was a substantial literature on citations [30]. In 1980, Hjerrpe published a review with more than 2 000 items of research on this and related topics [22]. It is appropriate to note that in 2015 in Web of Science data base there were more than 2 500 publications found by the query of citation analysis, starting from the Garfields 1972 work as the most cited article [9].

Citation itself can be understood as a complex phenomenon, which considers interaction between networks of authors and texts, that is why it indicates not only cognitive, but also social contexts of a knowledge claim [21]. However, it is important to clarify the differences between citation and *reference*. Although in practice many researchers do not divide these terms and use them interchangeably, each of them represents a different entity in the citing or cited perspective. The reference is made within a citing document and represents an acknowledgement of another study (and can be measured by the number of items in its bibliography as endnotes, footnotes, etc.), while a citation represents the acknowledgement received by the cited document from other publications (and can be seen in citation index) [2, 28].

Talking about citations in meta-level, citations can be viewed as explanans (something explaining something else) and *explanandum* (something to be explained). While a lot of interest is usually given to the first notion when citations explain research impact and value, the question of what is a citation should be brought up for the discussion in order to better understand what is certainly measured [21]. Citation analysis often starts with the assumption that references are indicators of influences on other scientists work. According to normative theory, norms of science suppose that authors cite works that they found useful for their own research, and it is assumed that they abide these norms, citing some authors and thus giving credits where they are due (as citing A by B means that A's works influenced B's thinking) [22]. Citation is as well considered as an indicator of reward in the science system in evaluation studies for science policy purposes (Martin and Irvine; Moed; Luukkonen; Merton; Latour and Woolgar in [21]), symbolic payments of intellectual debt, or *representation of trust* in virtual environments, that makes citation indexes to be "recommender systems" for other scientists [2]. Some practical reasons of citation-making process were also enumerated by Garfield (in [28]). In other scientific traditions, citation was also seen as a *function in scientific communication among texts* (Cronin, in [21]). Much attention was paid to the perfunctory and *rhetorical functions of citations* within the scientific community by B. Latour. Some investment into the citation analysis' theoretical legitimation (the theory of what is being analyzed) was made in the field of Science and Technology Studies (STS) itself, which included formalized measurement of citation analysis into empirical studies in 1980s. Nevertheless, theoretical and methodological reflection is still needed as there is a need of "translation" qualitative side of STS and merging it with formal approach [21].

During its history, citation analysis has proved to be a well-established tool for different aims of science analysis, including measurement of research impact and value [5, 14, 23]. However, besides all the discoveries, there was a substantial critique of this method [22, 28, 30, 34]. The scientists came up to the idea that it is not advisable to use citation analysis as a single and absolute criterion for judging the importance of a publication. Giving objective information regarding an individual, research group, journal or higher education institution, this method should be supplemented by other kinds of analysis, including qualitative approach (qualitative review, peer assessment, studying the authors behavior, characteristics of documents cited and not cited) [3, 28, 30, 34].

2.2 Citation Analysis as a Tool for Studying Scientific Fields

Methods of citation and co-citation analysis were used for studying of different aspects of scientific communication: coauthorship networks as complex systems [1, 8], dynamic aspects of collaboration networks [24], international collaboration as a self-organizing network based on the principle of preferential attachment [32], social ties, co-citations and inter-citations of Globenet, offline and online collaboration [35]. More examples of the studies are presented in [2].

One of the first examples of bibliometric tools usage for studying Social Network Analysis field was conducted by Hummon and Carley [15], who analyzed first volumes of Social Network Journal and other articles and have found that there is an invisible college in the growing SNA field. As the journal so specifically displays the people involved into network discipline it is quite often used as a looking glass on the social networks community [11]. Basing on the articles in the same journal, Lewis compared two types of social networks—formal collaborative relationships represented by coauthorship (who publishes papers with whom) and the informal collaborative relationships represented by acknowledgment (who thanks whom in published papers) in the scientific community of social network analysts [20]. Analyzing the Sociological abstracts data base, Otte and Rousseau [25] studied the underlying collaborative relationships between authors, built coauthorship network, pointed out central players of the field, and showed connections between SNA and other subfields (especially Information sciences).

In Russian scientific space studies, bibliometrics methods are not so much developed, even though in 1980s the technique of co-citing was developed by I. Marshakova, in parallel with G. Small [36]. However, there are also some examples of citation analysis usage for the studying of scientific fields. Cognitive structures of Russian Sociology and Ethnology by the method of co-citation analysis were studied by B. Winer, K. Divisenko and M. Safronova [27, 36], who tested different methods of the cognitively closed groups detection. Among other methods, citation analysis was used in the study of intelligent landscape and social structure of the local academic community (the case of St. Petersburg) [29]. The authors found three groups among Russian sociologists (West-side, East-side and Transition zone), who tend to see (in the meaning of citing) the representatives of their own groups, while the authors from other groups stay almost “invisible” for them.

It is important to note that most of the studies regard citation in the first sense—as the acknowledgement that the author gets from other scientists, but not a credit that he or she gives to them,—that is why what is being used is the method of citation analysis. In our project, we understand the citation in the second meaning—as an acknowledgement of another study to the current study—and propose to use the method that can be called reference analysis, where reference means a tie between the author (writer) of article and the author whom he or she cites in publication. Basing on authors of articles and authors from their bibliography lists allows us to build networks of relations between different groups of authors and study the structure of a community of researchers involved into network studies in Russian scientific space.

3 Data and Methodology

In this section we present some practical information concerning our data source and process of data collection and preprocessing (such as author disambiguation and paper classification) and discuss some problems associated with these procedures.

3.1 Data Source

The data source that we use is the electronic library of scientific periodicals in Russian called eLibrary.ru—a leading electronic library of science periodicals in Russian, which contains more than 3 900 Russian-language and 4 000 foreign scientific journals, abstracts of nearly 20 000 journals and the descriptions of 1.5 million of Russian and foreign dissertation thesis, and has 1.1 million individual users and 2 200 organizations registered. The base is integrated with the Russian science citation index (RSCI)—a national information–analytical system which accumulates more than 6 million of publications of Russian authors in more than 4 500 Russian journals, as well as information on citing of these publications. Even though the system is based on indexed articles in Russian scientific journals, in recent years other types of scientific publications were included into the base—such as reports on conferences, monographs, tutorials, patents, and dissertation thesis. The chronological coverage of the system comes from 2005, but for many resources the depth of archives is deeper. In sum, eLibrary resource not only gives the support of scientists by the bibliographic information, but provides a tool for assessing the effectiveness of science and research organizations (more than 11.000) and scientists (more than 600.000), as well as scientific journals.

Unfortunately, when we go from the level of description of the resource to its practical usage, some problems associated with data collection appear. First of all, the resource does not offer any procedures for mass data downloading, as some scientific aggregators as Web of Science allow. The data collection process needs manual collection, which is impossible in the situation of a large amount of data, or the special crawling techniques.

3.2 Data Collection and Preprocessing

For each article, the eLibrary base contains information on publisher's imprint, paper's title, authors, their affiliations, keywords and disciplines, abstracts, and what is the most important in the terms of the current study—lists of bibliography (the references) and lists of other eLibrary papers—that cited the initial paper. That is why our data contains two parts. Data base (1) contains all the information on articles journal's name, discipline, year of publication, author's organization, keywords, annotation, scientometric indexes, etc. Data base (2) contains information on references—main data that contains authors and lists of bibliography.

The method of data collection that we used is based on expanding publication graph using two strategies:

1. Expansion strategy—a set of methods to increase the number of relevant papers (increases recall):
    a. Keyword search—we formed a list of 48 Russian and English keywords that we consider relevant to the domain of network research and collected all publications. Given the search query, eLibrary engine returns all articles that contain the search query in any field, including title, keyword, or annotation. Keyword list contains such keywords as *network analysis*, *relational sociology*, *actor-network theory*, *graph of a network*, etc. Having a list of relevant articles, obtained after filtering, we generated the distribution of articles for each keyword and selected the keywords that appeared more than the median value. Afterward, we repeated step a. with newly obtained keywords.
    b. Author search—if author had more than three publications in the domain of network research, we collected all his publications and filtered them with our keyword classifier;
    c. Citation search—we took all papers that cited the article from the network research domain.
2. Filter strategy—a set of methods to remove irrelevant papers, found during the expansion strategy (increases precision). It often happens that relevant keywords are used in a different meaning or separated by other words. Besides, many authors publish papers in multiple research areas or cite articles from other domain. In all these cases we often collect irrelevant papers that must be filtered. Traditionally, document classification task is made by machine learning classification, but by virtue of the fact that we have no annotated collection of documents, we used two-step strategy:
    a. Cluster documents using keywords and annotation text and manually mark relevant clusters.
    b. Convert chosen clusters to binary classes (relevant/irrelevant) and make a binary classification of the entire array. Filter strategy problem is described in more depth in Paper classification section.

Proposed method is very similar to the shark-search approach [13], but the relevance of each cluster is annotated manually. The following limitations should be taken into account:

1. Clustering forms lexically similar groups, but cluster center and periphery may contain very different articles;
2. Having a list of totally unconnected fields, we need at least one "seed" keyword in each field to make expansion strategy work;
3. Papers that do not contain common keywords can be lost during the filtering step.

There is no standard evaluation task for the proposed method as we cannot obtain recall for the 20 m collection, but we have made evaluation for our filtering strategy on the Cora Research Paper Classification task.[1] We compared each rubric with the most relevant cluster and obtained F1-score = 0.64 for automated and 0.92 for semi-manual clustering where F1-score is the mean score among all rubrics.

Brief description of typical parsing problems associated with collected fields is provided below:

- Author's name and surname of each author. We excluded papers with more than 10 authors because we consider their interconnectivity to be very weak and the problem of combinatorial explosion at the step of artifacts generation. We also cleaned up special eLibrary markup, such as editor or translator, and special author affiliations–organizations related to the certain author.
- Keywords—mentioned by author and splitted by comma. We excluded articles with one keyword consisting of more than five words.
- Language of the article. Article, which may differ from the abstract and keywords language. We used external language detection tool based on sequences of characters information.
- Abstract—short text, describing the article, which may be written in multiple languages (for example in Russian, French and English in paper 11897467 [2]). In this case we kept only one language (priority is Russian, English, other languages).
- Citations—list of papers, cited by this paper. Citation list is unstructured and very dirty, so we extracted only surnames information.

As there are different strategies of author counting in the literature (see [2, 22]) we decided to use all authors instead of first authors counting of citations and to include self-citing into the collection (as it is entirely appropriate for scientists to build on their own previous studies).

Such way of data collection allowed us to get not only the information on ties of "citing" ("referencing") type (1), but also the data on coauthorship in the article (2) and coauthorship in the citing article (3)—when there was more than one author in citing and cited articles. Also the data on ties between authors and "artifacts" was collected, where the latter were author's affiliation (organization) (4) and concepts that he or she uses in the works (5). Thus, the collected data potentially allows us to analyze five types of ties, and conduct more complex study in future, including semantic, citation, co-citation, analysis of coauthorship and affiliation networks, analysis of ties between authors and concepts, as well as work on such methodological issues as comparison of methods of articles sampling, as shown in Fig. 1.

[1] Download at http://sites.google.com/site/semanticbasedregularization/home/software/experiments_on_cora.

[2] Example is available by the link: http://elibrary.ru/item.asp?id=11897467.

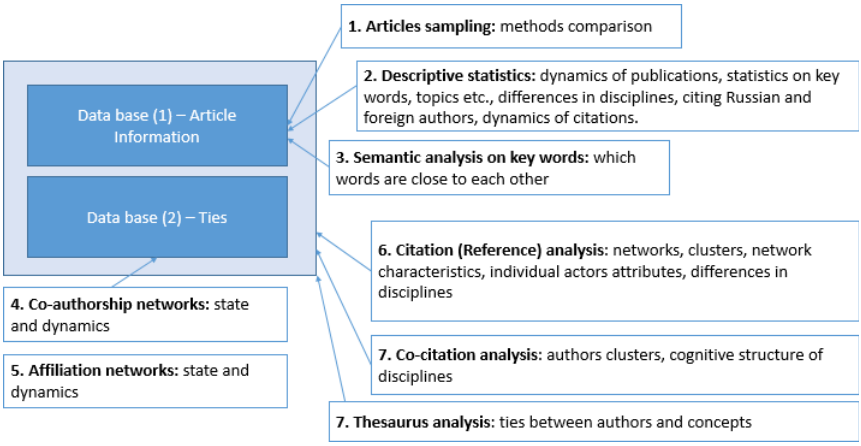

Fig. 1 Possible directions of analysis

3.3 Author Disambiguation

As many researchers work in the same field, papers made by different researchers with the same name should be searched out during the collection process. The process of resolving conflicts that arise when a potential name is ambiguous is called disambiguation. In some cases disambiguation is provided by eLibrary itself, and the authors are marked with special hyperlink and identifier, that is unique in the whole data base. In other cases, we have to solve two problems:

- How many unique authors with the same name are represented in the collection?
- How to classify each ambiguity to one of the existing classes?

Additional problem appears when we have the same author published in English and Russian languages. The core problem of matching Cyrillic and Latin names is that usually author tends to save phonetics of the word that does not match any transliteration rules. Our match was based on idea that most Russian authors obtain at least one English article with correct English name that can be used as correct transliteration. English authors were transcribed into Russian according to the GOST 7.79 2000 standart. We searched for all possible surnames with Levenshtein distance ≤ 3 [19] and manually validated them. Each author was trimmed to surname and initials, and we used hierarchical agglomerative clustering [16] to label each author-in-the-article pair to the certain cluster, based on the following features:

1. Keywords—mentioned in the article;
2. Coauthors—surnames of the coauthors for the certain author;
3. Affiliations—organizations, related to the ambiguous author;
4. Date of the publication—year of publication (is very weak feature as we have a contemporary field);
5. List of citations—surnames of the cited authors.

Affiliation feature needs to be additionally cleared as it varies from abbreviation to the full name of the laboratory and organization, which causes mismatches during clustering.

3.4 Paper Classification

As described above, some papers may be irrelevant to our research domain due to the specific of the proposed search method. We had no opportunity to get a representative collection of relevant papers, so we used unsupervised learning methods. We applied BIRCH [37] clustering algorithm, based on the following features:

- Keywords— mentioned in the article;
- Disambiguated authors of the paper.

Clustering hyperparameters are as follows: number of clusters 64–255, number of top terms 50 000, term weighting method

$$W_i = \sigma^2 \left(\frac{TF_i}{IDF_i} \right)$$

where i is term id, TF_i—i-th is the term frequency, frequency of the given term in the document, IDF_i is the inverted document frequency, the measure of fraction of the documents that contain the i-th term in the whole collection, distance metric—Ward's method [33]. Overall collection process consisted of four consecutive phases: $Search\ phase\ 1 \rightarrow Filter\ phase\ 1 \rightarrow Search\ phase\ 2 \rightarrow Filter\ phase\ 2$. Resulting dataset is described in Table 1 below. The resulting number of articles is composed of 8 260.

Table 1 Resulting dataset statistics

	Search phase 1	Filter phase 1	Search phase 2	Filter phase 2
Number of articles	220 657	5 836	442 524	8 260
Added by title search	220 657	–	107 208	–
Added by author search	0	–	56 932	–
Added by citation search	0	–	181 867	–
Filtered by article type	–	121 880	–	234 114
Filtered by clustering	–	90 941	–	200 150

4 Results

In this section we will briefly provide the overview on the data on articles (Data base 1). We analyzed articles from the resulting data collection, which is more than 8 000 papers. Main information on this sample is shown in Table 2 (amount, min. and max. citing, the earliest and the latest year of publication).

The growing number of publications collected is shown in Fig. 2, with the first article published in 1988. During recent years we can see the growth of interest to the network topics—in last 5 years, from 2010 to 2015, the number of articles increased almost in four times. However, the low annual amounts for previous years might be associated with the quality of the data base itself.

Talking about types of the articles, most of the presented publications are articles in journals (scientific articles), which form 67% of the entire sample (Table 5). Second and third places are taken by PhD thesis (11%) and articles in the conference proceedings (9%). Other types of publications can be met less frequently.

Table 2 Main characteristics of the sample

Number of articles	8 260	100%
Mean citing	2.13	–
Min citing	0	–
Max citing	303	–
The earliest year of publication	1988	–
The latest year of publication	2016	–

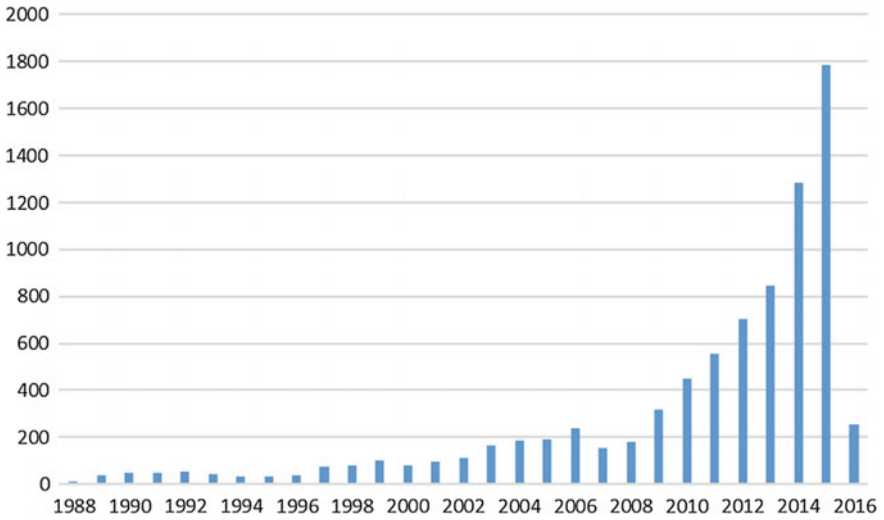

Fig. 2 Number of articles, by years

Table 3 Language of articles

	Frequencies	% by column
Russian	6 807	82
English	1 362	16.5
Other	58	0.7
Not defined	33	0.4
Total	8 260	100

Table 4 Number of citations

		Frequencies	%	valid %
Valid	0	5 132	62.1	62.5
	1–5	2 360	28.6	28.7
	6–20	579	7.0	7.1
	21–50	104	1.3	1.3
	51–100	26	0.3	0.3
	101–150	5	0.1	0.1
	151–500	4	0.05	0.05
	Total	8 210	99.4	100.0
Missing	System—missing	50	0.6	
Total		8 260	100.0	

The number of PhD thesis is quite high–948 dissertations, with the first published in 1997.

The language of the majority of articles is Russian (82%), while in other cases it is English (16.5%). Other 58 articles are written in the languages of different language groups (Table 3).

In terms of this study, the information on articles citing (from other authors) is quite interesting. We used the Russian science citation index (RSCI) values to compare the articles. It was found that even though the mean number of citing in the whole base is 2.13, the majority of articles (62%) do not have any citations in RSCI, i.e., actually located outside of the area of attention of other researchers. Another 29% of articles have between 1 and 5 citations. Thus, 91% of articles in general do not have more than 5 citations, and just 7% of articles have between 6 and 20 citations. Just 9 articles have more than 100 citations, where the maximum value—303—belongs to the textbook on social networks, models of information influence, management, and confrontation [12] (Table 4).

Quite interesting conclusions can be done from the cross-tables of type of publication and the mean number of citations. The type of publication which is most often met in the base—scientific articles—in average has just 1 citation. Small values are also characteristic of theses and articles in conference proceedings (0 and 1 citation, respectively). The highest amounts of citations are typical for monographs (2% of

Table 5 Citation in RSCI by types of citations

	Freq.	% by column	Mean	Min	Max
Article in journal—scientific article	5 521	67	1	0	86
Thesis	948	11	3	0	71
Article in conference proceedings	728	9	01	0	29
Article in journal	282	3	2	0	59
Monograph	195	2	18	0	292
Articles in the digest of articles	115	1.39	1	0	44
Article in journal—review article	96	1.16	2	0	27
Article in journal—conference materials	68	0.82	0	0	4
Tutorial	66	0.8	16	0	303
Abstract at the conference	49	0.59	0	0	3
Thesis abstract	42	0.51	4	0	28
Article in journal—other	39	0.47	0	0	3
Article in the journal—abstract	30	0.36	0	0	1
Article in the journal—a short message	12	0.15	1	0	7
Methodological guidelines	10	0.12	3	0	20
Digest of articles	10	0.12	1	0	5
Chapter in a book	9	0.11	0	0	3
Article in the journal—review	7	0.08	0	0	1
Dictionary or reference book	6	0.07	3	0	11
Article in the journal—editorial note	4	0.05	1	0	2
Article in the journal—scientific report	3	0.04	2	0	5
Report on research work	2	0.02	0	0	0
Article in the journal—correspondence	2	0.02	0	0	0
Article in the journal—personality	1	0.01	0	0	0
Deposited manuscript	1	0.01	0	0	0
Article in an open archive	1	0.01	0	0	0
Brochure	1	0.01	7	7	7
Other	12	0.15	0	0	0
Total	8 260	100	2	0	303

sample, mean value 18) and tutorial (1% of sample, mean value 16) (Table 5). Among monographs and tutorial just 29% and 18% of articles, respectively, have the number of citations between 1 and 5; while 22% and 26% of them, respectively, between 5 and 20, and 17% and 11%, respectively, between 21 and 50. This outcome correlates well with the results of foreign studies, in which it has been proved that there are the differences in citation practices between books and journals, with the greater emphasis given to the books [30].

Table 6 Network of keywords measure parameters

Vertices	143
Unique edges	1302
Edges with duplicates	2138
Total edges	3440
Avg degree	33
Maximum geodesic distance (Diameter)	4
Average geodesic distance	1.87
Graph density	0.23
Connected components	1

We provide these results with the thematic map of the data, which was constructed by network analysis of the main keywords used in the collected articles (Fig. 3). Extracted keywords which were met in the dataset more than 30 times (min. value = 30, max. value = 1888) were used for building this network. In the network, which was constructed by NodeXL software (by Harel-Koren Fast Multiscale layout), the scale of the nodes corresponds to the frequency the keyword was met in the dataset, the color of nodes corresponds to their degree value, edges are weighted according to the number of times two keywords were met in the same article, and the edges are filtered (min. value = 4 and max. value = 801). The information on network is provided in Table 6.

Figure 3 shows map of different groups of keywords, which are more likely to meet in the same article and thus are close to each other by topics. In the network center, there are words more frequently used in the dataset—social network and Internet, both in Russian and English. These words are connected to a large group of keywords above them associated with Internet communication technologies such as blogosphere, social networking, and some sites like VKontakte, Facebook, Twitter, Web 2.0, social media, and media itself. Quite large group is associated with Internet marketing promotion, content, site, SMM. Close to them we can also see the groups of words associated with Internet communication, self-presentation, and Internet addition, as well as information security and personal data.

Main keywords “Social Network” and “Internet” are also associated with information technologies, communication, and information society. Another large group (at the bottom) consists of the words associated with society—social capital, human capital, trust, civil society, and social inequality. On the right side, we can see a large group of words associated with education—educational organizations, educational programs, professional standard, modernization of education, and academic mobility. There is another large group on the left side, which consists of the vocabulary from humanities—Sociology, Philosophy, Social philosophy, Political science, Economics, International relationships, Conflictology—such as Social and economic institutions, Social processes, Social structure, Social community, State, Power, etc. Thus, we can see that the thematic map of our dataset covers different aspects of

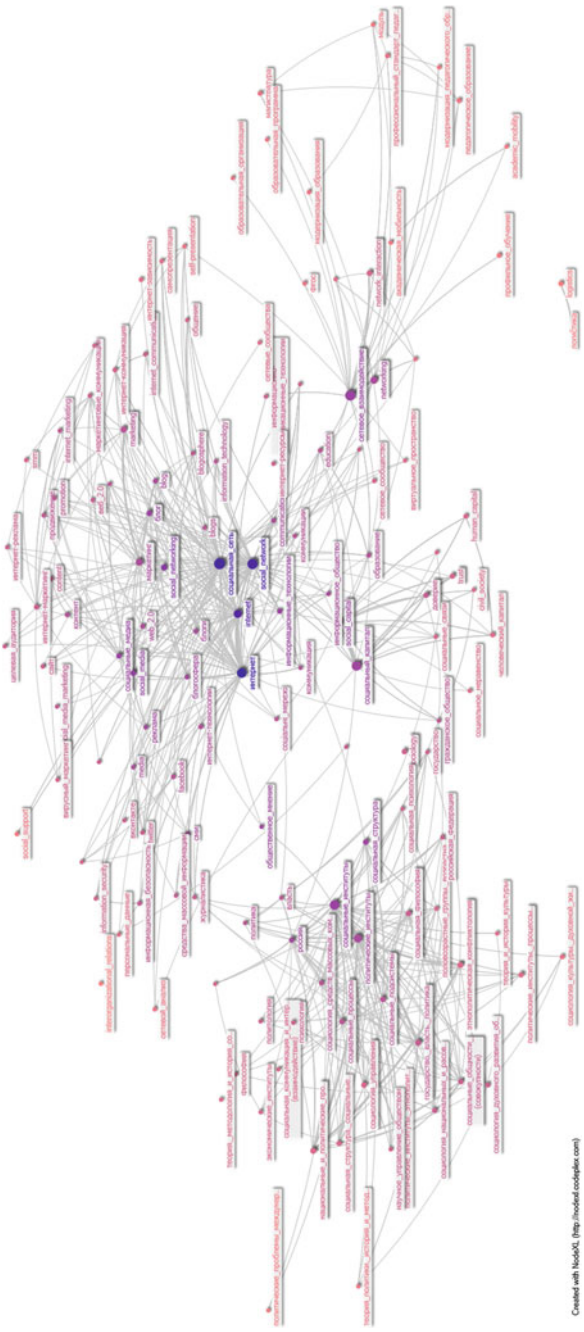

Fig. 3 Network of keywords

network topics. Another outcome that can be done after examining this network is the correctness of the data collection procedure, as we have not met any irrelevant topics among our main keywords.

5 Discussion

Having written all the details on the procedures, we should admit that there are a lot of things to be done in future for the implementation of the study and accomplishment of its aims. First of all, the dataset that we have now concerns only the information on articles on network topics. However, the main aim of our study is to build networks of scientists involved into this type of research according to their citation practices. It means that after the final cleaning of this dataset we have to work at collection of other dataset with citations.

At the same time, we still have some issues concerning the dataset with articles information (Dataset 1). Talking about author disambiguation, we can propose the usage of techniques based on classification of the network of coauthors, which can increase method performance. Having the information from eLibrary resource on the authors already familiar to it, we can check the efficiency of such classification technique. Such work would be practical not only for the current study, but also for solving the author disambiguation problem in general.

Basing on main discoveries of the previous studies in the field of citation analysis [30], we could expect that there are different closed groups of Russian scientists working in the field of network research, which appears in the following aspects:

- There are discrete groupings of researchers, with relatively little overlap between Russian authors;
- Russian authors more often cite foreign (North American and European) authors than each other;
- Russian authors tend to cite particular (different) groups of foreign authors, which are connected with topics and methods that they use;
- The significant number of Russian authors are isolated researches.

The last issue on isolated researches can be already verified with our preliminary results, according to which the significant number of articles in our collected sample does not have any citations from other authors, which means that they are invisible to the other scientists and do not provide any information that can be used by others in the field. Other formulated propositions should be checked in the future studies, during the analysis of ties between citing and cited authors. Analysis of the full set of data will also give us the opportunity to find the most active drivers of network studies and to see the structure of a network research community in Russia.

6 Conclusion

The main aim of this article was to present some methodological issues concerning our proceeding research on the network studies field in Russia. Providing the overview of the method of citation (reference) analysis as a tool for studying scientific fields, we showed its power and relevance to the studies of scientific communities structure. Then we described the process of data collection in deep details, in order anyone interested could repeat the data collection procedure. We emphasized some methodological and technical issues typical for the process of data collection, network expansion and filtering strategies, authors disambiguation and transliteration for the specific problem of domain-oriented information retrieval. Data collection and extraction code was published online,[3] so that any researcher could make experiments in his domain. We tried to enumerate all the problems that we faced to in our study and proposed the procedures of their overcoming. From one side these problems are standard for the data collection, but we also see some specific characteristics of the eLibrary resource. Providing the brief information on collected data on articles, we made a description of our dataset. We considered the thematic map of the data, which was constructed by network analysis of the main keywords used in the collected articles. Finally, basing on previous studies and collected data, we made some propositions that should be checked during next steps of analysis. These steps should be done in the following directions:

- on methodology: future work on the author disambiguation techniques based on classification of the network of coauthors, which can increase method performance;
- on data collection: collection of the full dataset on ties between authors (Dataset 2);
- on data analysis: the analysis of ties data base, which will allow us to build the network of research community of scientists involved into network studies in Russia and answer the main research questions of this study.

Acknowledgements The study has been funded by the Russian Academic Excellence Project '5–100' and RFBR grant 16-29-09583 "Methodology, techniques and tools of recognition and counteraction to organized information campaigns on the Internet". We thank the participants of the Sixth International Conference on Network Analysis NET 2016 and its organizer Laboratory of Algorithms and Technologies for Networks Analysis (LATNA) of National Research University Higher School of Economics (Nizhny Novgorod, Russia) for fruitful discussions and valuable comments on the platform of LATNA Laboratory.

[3] Available at https://github.com/lab533/elibrary.

References

1. Barabsi, A.L., Jeong, H., Neda, Z., Ravasz, E., Schubert, A., Vicsek, T.: Evolution of the social network of scientific collaborations. Phys. A **311**(34), 590–614 (2002)
2. Bar-Ilan, J.: Informetrics at the beginning of the 21st century-a review. J. Inf. **2**, 152 (2008)
3. Belotti, E.: Getting funded: multi-level network of physicists in Italy. Soc. Netw. **34**(2), 215–229 (2012)
4. Casey, D.L., McMillan, G.S.: Identifying the "Invisible Colleges" of the "Industrial and Labor Relations Review": a bibliometric approach. Ind. Lab. Relat. Rev. **62**(1), 126–132 (2008)
5. Cronin, B., Atkins, H.: The Web of knowledge-a Festschrift in honor of Eugene Garfield. ASIS Monograph Series. Information Today, Medford (2000)
6. De Bellis, N.: Bibliometrics and Citation Analysis: From the Science Citation Index to Cybermetrics. The Scarecrow Press, 417 p. (2009)
7. Devyatko, I.: Sotsiologicheskie teorii deyatel'nosti i prakticheskoy ratsional'nosti [Sociological theories of activity and practical rationality]. Chistie vodi, 336 pp (2003)
8. Farkas, I., Derenyi, I., Jeong, H., Nda, Z., Oltvai, Z.N., Ravasz, E.: Networks in lifescaling properties and eigenvalue spectra. Phys. A **314**(14), 25–34 (2002)
9. Garfield, E.: Citation Analysis as a Tool in Journal Evaluation. Science New Series **178**(4060), 471–479 (1972)
10. Gradoselskaya, G.: Setevye izmereniya v sotsiologii: Uchebnoe posobie [Network Measurement in Sociology: the textbook], Batygin G.S. (ed.). Publishing House New Textbook, 248 pp (2004)
11. Groenewegen, P., Hellsten, L., Leydesdorff, L.: Social networks as a looking glass on the social networks community. In: Sunbelt XXXV International Sunbelt Social Network, Abstracts. Hilton Metropole, Brighton, UK, 23–28 June 2015, p. 118 (2015)
12. Gubanov, D., Novikov, D., Chkhartishvili, A.: Sotsial'nye seti: modeli informatsionnogo vliyaniya, upravleniya i protivoborstva [Social networks: Information models of influence, control and confrontation]. Izdatel'skaya firma "Fiziko-matematicheskaya literatura" M., 228 pp (2010)
13. Hersovici, M., Jacovi, M., Maarek, Y. S., Pelleg, D., Shtalhaim, M., Ur, S. 1998. The shark-search algorithm. An application: tailored Web site mapping. Comput. Netw. ISDN Syst. **30**, 317–326 (1998)
14. Hoffman, K., Doucette, L.: A Review of Citation Analysis Methodologies for Collection Management. College and Research Libraries (2012)
15. Hummon, N., Carley, K.: Social networks as normal science. Soc. Netw. **15**, 71–106 (1993)
16. Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, 368 p. (2005)
17. Kharhordin, O.: Predislovie k Latur B. Nauka v deystvii: sleduya za uchenymi i inzhenerami vnutri obshchestva [Preface to B. Latour, Science in Action: How to Follow Scientists and Engineers through Society] (trans: St.Peter, F.K.). European University in St. Petersburg, pp. 7–19 (2013)
18. Kravchenko, S.: Sotsiologiya: paradigmy cherez prizmu sotsiologicheskogo voobrazheniya: uchebnik [Sociology: paradigm through the prism of sociological imagination: the textbook], 3rd ed. Examen, 750 pp (2007)
19. Levenshtein, V.I.: Binary codes capable of correcting deletions, insertions, and reversals. Sov. Phys. Dokl. **10**(8), 707–710 (1966)
20. Lewis, K.: Collaboration and acknowledgment in a scientific community. In: Sunbelt XXXV International Sunbelt Social Network, Abstracts. Hilton Metropole, Brighton, UK, 23–28 June 2015, p. 176 (2015)
21. Leydesdorff, L.: Theories of citations? Scientometrics **43**(1), 5–25 (1998)
22. MacRoberts, M.H., MacRoberts, B.R.: Problems of citation analysis: a critical review. J. Am. Soc. Inf. Sci. (1986–1998) **40**(5) (1989)
23. Moed, H.F.: Citation Analysis in Research Evaluation. Springer, Dortrecht (2005)

24. Newman, M.E.J.: The structure of scientific collaboration networks. Proc. Nat. Acad. Sci. U.S.A. **98**(2), 404409 (2001)
25. Otte, E., Rousseau, R.: Social network analysis. J. Inf. Sci. **28**, 441 (2002)
26. Radaev, V.: Osnovnye napravleniya razvitiya sovremennoy ekonomicheskoy sotsiologii [The main directions of development of modern economic sociology]. In: The book Ekonomicheskaya sotsiologiya: Novye podkhody k institutsional'nomu i setevomu analizu [Economic Sociology: New approaches to institutional and network analysis] (trans: Dobryakova M.S.); M.: POSSPEN, pp. 3–18 (2002)
27. Safonova, M., Winer, B.: Setevoy analiz sotsitirovaniy etnologicheskikh publikatsiy v rossiyskikh periodicheskikh izdaniyakh: predvaritel'nye rezul'taty [Network analysis of co-citations of ethnological publications in Russian periodicals: preliminary results]. Sotsiologiya: Metodologiya, metody. matematicheskoe modelirovanie **36**, 140–176 (2013)
28. Smith, L.: Citation analysis. Library Trends (1981)
29. Sokolov, M., Safonova, M., Guba, K., Dimke, D. 2012. Intellektual'nyy landshaft i sotsial'naya struktura lokal'nogo akademicheskogo soobshchestva (sluchay peterburgskoy sotsiologii) [Intelligent landscape and social structure of the local academic community (the case of the St. Petersburg Sociology)]. M.: Izd. dom Vysshey shkoly ekonomiki, 48 pp
30. Tight, M. Citation and co-citation analysis. In: Brew, A., Lucas, L. (eds.) Academic Research And Researchers: Policy and Practice. Open University Press (2009)
31. Vakhshtain, V.: Vozvrashchenie material'nogo. Prostranstva, seti, potoki v aktorno-setevoy teorii [The return of the material. "Spaces", "networks", "flows" in the actor-network theory]. Sotsiologicheskoe obozrenie **4**(1), 94–115 (2005)
32. Wagner, C.S., Leydesdorff, L.: Network structure, self-organization, and the growth of international collaboration in science. Res. Policy **34**(10), 1608–1618 (2005)
33. Ward, J.H.: Hierarchical grouping to optimize an objective function. J. Am. Stat. Assoc. **58**(301), 236–244 (1963)
34. Warner, J.: A critical review of the application of citation studies to the research assessment exercise. J. Inf. Sci. **26**, 453 (2000)
35. White, H.D., Wellman, B., Nazer, N.: Does citation reflect social structure? longitudinal evidence from the Globenet interdisciplinary research group. J. Am. Soc. Inf. Sci. Technol. **55**(2), 111–126 (2004)
36. Winer, B., Divisenko, K.: Kognitivnaya struktura sovremennoy rossiyskoy sotsiologii po dannym zhurnal'nykh ssylok [Cognitive structure of modern Russian sociology according to the journal links]. Zhurnal sotsiologii i sotsial'noy antropologii **4**(60) (2012)
37. Zhang, T., Ramakrishnan, R., Livny, M.: BIRCH: an efficient data clustering method for very large databases. In: Proceedings of the 1996 ACM SIGMOD International Conference Management data SIGMOD 96, vol. 1, pp. 103–114 (1996)