Communication of Scientists through Scientific Publications: Math-Net.Ru as a Case Study

Andrey Pechnikov[1,2][0000-0002-0683-0019], Dmitry Chebukov[3][0000-0001-9738-8707] and Anthony Nwohiri[4][0000-0001-7622-7533]

[1] Institute of Applied Mathematical Research of the Karelian Research Centre of the Russian Academy of Sciences, 11, Pushkinskaya str., Petrozavodsk, Russia
[2] Faculty of Applied Mathematics and Control Processes, Saint Petersburg State University, Saint Petersburg, 7/11 Universitetskaya Emb., Russia
[3] Steklov Mathematical Institute of RAS, 8 Gubkina Str., Moscow, Russia
[4] Department of Computer Sciences, University of Lagos, University Road, Akoka-Yaba, Lagos, Nigeria
pechnikov@krc.karelia.ru

Abstract. We present a study of two scientific collaboration graphs built using data drawn from Math-Net.Ru, an all-Russian mathematical portal. One of the graphs is a citation-based scientific collaboration graph. It is an oriented graph with no loops and multiple edges. Its vertices denote authors of papers, while the arcs connecting these vertices denote that the first author has, in at least one of his papers, cited the work of the second author. The second graph is a coauthorship-based graph. It is a non-oriented graph, where the vertices denote authors, while edges connecting two vertices indicates that the two authors have coauthored at least a paper. We conduct a traditional study of the main characteristics of both graphs, such as degrees of vertices, influence of vertices, diameter, mean distance, connected components and clustering. Both graphs are found to have a similar connectivity structure – both have a giant component and several small components. Using the two graphs, we split the set of Math-Net.Ru authors. In this set, it was revealed that more than 40% of authors who have co-authored a paper with someone have not ever cited their co-authors. This means there is no deliberate plan to cite each other's work in the journals registered in Math-Net.Ru.

Keywords: Scientific collaboration, Scientific journal, Citation, Co-authorship, Graph, Math-Net.Ru.

1 Introduction

Social networks have been the subject of empirical and theoretical study for the last 60 years – its structure is of great importance for information dissemination. Examples of such communities are scientific collaboration networks built on the basis of scientific papers. A study of scientific collaborations allows to assess trends in the development

of scientific areas, identify persons, research centers and schools, as well as detect relationships. There is no single universal scheme by which a scientific work can be evaluated. Nevertheless, the task of measuring individual quantitative characteristics of scientific information in specific scientific areas is being solved with varying degrees of success.

One of the first publications on this topic is a seminal paper by Derek John de Solla Price [1], devoted to citation networks and is associated with the emergence of electronic repositories of scientific papers. The main objects of research since then till now are still citation networks and coauthorship networks built using data drawn from various electronic sources, including Google Scholar. Graphs are constructed using this data and are investigated by mathematical methods with subsequent meaningful interpretation of the results.

Despite the great work done, the topic seems to be inexhaustible. There is no unified theory of scientific collaboration networks. Research on specific digital libraries brings new, sometimes unexpected results. The range of theoretical approaches and methods is expanding, new technical and software capabilities are emerging. A good review of English-language publications is given in [2].

Our research is based on information drawn from the database of all-Russian mathematical portal Math-Net.Ru (http://www.mathnet.ru). It is a well-known web resource with rich collection of full-text archives of leading Russian mathematical journals and information about their authors. As of March 30, 2020, there were 119,139 authors, 238,733 research papers and 135 journals (periodicals) registered on the portal.

A table containing metadata of articles is the key element of the relational database (used by the MSSQL DBMS). Initially, the database was created such that authors and affiliations were contained in separate tables, while the article table was connected to the tables of authors and affiliations through a one-to-many relationship. Each author and each affiliation are unique database elements combined into tables of persons and organizations.

This approach allows for each author to retrieve a list of his papers using his unique code (author_id). Unlike in the Web of Science and some other bibliographic resources, the author does not need to search for his works using his last name. The database of persons linked to articles easily allows you to select authors of the same papers, as well as individual authors with no co-authors.

Information system Math-Net.Ru also indexes bibliography and stores them in a database in a structured form [3]. Literature lists of all publications are combined into a single database table, where information about the author, title, year, volume, pages of cited paper are stored in separate columns. Each individual link corresponds to one record in the table. This technique facilitates the task of automatically placing hyperlinks to bibliometric databases. It solves the problem of finding backlinks, and also allows you to automatically export links in different formats, such as PDF, XML, and HTML. Among the hyperlinks from the bibliography items, there are also links to articles indexed in the Math-Net.Ru publications database. In this way, a link is made between the citing and cited articles.

Three graphs were built according to data drawn from Math-Net.Ru. The first of them is a mutual citation graph of Math-Net.Ru journals. The other two are scientific collaboration graphs, reflecting citation and co-authorship information.

Since the question always arises as to how adequate (or acceptable) the database of a specific scientific area is to the ongoing research on scientific collaboration, we will use journal self-citation to show that the database is adequate to the research on scientific collaboration. Journal self-citation is a special form of publication scientific communication, but within a thematic community of journals. It should not have a major impact on journal rankings within that community. We have shown that self-citation has little impact on the rankings of journals in Math-Net.Ru. This suggests that the Math-Net.Ru database is adequate for the research being conducted.

Our graphs were constructed based on Newman's opinion regarding co-authorship: "*I study networks of scientists in which two scientists are considered connected if they have coauthored a paper. This seems a reasonable definition of scientific acquaintance: most people who have written a paper together will know one another quite well. It is a moderately stringent definition, since there are many scientists who know one another to some degree but have never collaborated on the writing of a paper*"[4, p. 404].

With regard to citation, we will assume that one scientist "knows" (not necessarily personally) another scientist if he has cited the work of that scientist in his own paper (the reverse is not true).

In this work, two types of scientific collaboration graphs are built based on citation and co-authorship data drawn from Math-Net.ru. A graph is understood in the traditional sense as a pair of sets: the first is a set of vertices, and the second is a set of pairs of vertices. Pair of vertices (a, b), depending on the context, is oriented when (a, b) ≠ (b, a); in this case, such a pair is called an arc. When the pair is non-oriented, (a, b) = (b, a); the pair is called an edge. A citation graph is a directed graph, because an author citing another does not necessarily mean a reverse citation. The co-authorship graph is an undirected graph because the co-authorship relationship is symmetric.

Here, our task is not to study the power of collaboration depending on the number of co-authored papers or citations of other papers by one author. Therefore, in the scientific collaboration graphs, built on the basis of co-authorship and citation, arcs (edges) have a multiplicity of 1.

We conduct a traditional study of the main characteristics of both graphs, such as degrees of vertices, influence of vertices, diameter, mean distance, connected components and clustering. Their similarities and differences are also discussed. Using the two graphs, we split the authors set into four disjoint subsets according to their connections through co-authorship and/or citation.

2   Self-citation as an indicator of the reputation of Math-Net.Ru journals

As mentioned above, there are 135 journals registered in Math-Net.Ru, and slightly more are indexed – 181 journals (not all indexed journals are registered). About 56% are related to fundamental mathematics, 20% to applied mathematics and computer

science, and 13% to physics. About 6% belong to a wide range of physical and mathematical sciences, while 5% are multidisciplinary journals.

Using the data drawn from Math-Net.Ru, we construct a journal citation graph, which we denote as *Journ-graph*. The set of vertices in this graph corresponds to a set of journals. Vertex i (beginning of the arc) is connected by an oriented arc to vertex j (end of the arc) if there is at least one citation of journal j in journal i. The weight of the (i, j) arc is equal to the number of citations of journal j in journal i. Thus, the graph is an oriented graph with weighted arcs.

There are 166 vertices in the *Journ-graph* graph. We excluded 15 isolated journals that have no references to and from other journals.

The sample from the Math-Net.Ru database for 2019 contains 264,000 records of journal mutual citations in Math-Net.Ru (of which 75,000 are self-citations). A large number of journal citations are multiples. So, in the *Journ-graph* graph, there are 4,856 weighted arcs, weighing from 1 to 10,375. *Journ-graph* is a connected graph, but not strongly connected. The maximal strongly connected component contains 100 vertices. The rest of the components (singular) contain 1 vertex each with only incoming or only outgoing arcs.

The significance of vertices for an oriented graph can be determined in various ways, and each of them requires meaningful interpretation. The PageRank (*PR*) allows to compare the relative "significance" of vertices in a similar way with the significance of web pages [5]. We can offer some meaningful interpretation of the significance of vertices based on *PR* scores in the *Journ-graph* graph as follows: if you imagine a certain "surfing scientist", who infinitely moves from one journal to another via bibliographic citations, he will most likely have high *PR* score in a journal.

For example, peer reviewed mathematical journal *Matematicheskii Sbornik* has 5,678 papers in Math-Net.Ru for 2019, 22,620 citations (including self-citations) and 0.06 PR. The *Mathematical Education* journal has 13 papers, no citations, and 0.0019 *PR*. That is, the probability of a "surfing scientist" getting on the *Matematicheskii Sbornik* journal is 30 times higher than on the journal with the lowest *PR* in the *Journ-graph* graph.

The influence of self-citation on Math-Net.Ru journal indicators was checked as follows. With an interval of 10 years, in 2019, 2009 and 1999, the $Journ\text{-}graph_{2019}$, $Journ\text{-}graph_{2009}$ and $Journ\text{-}graph_{1999}$ graphs were built. For each journal, *PR* values PR_{2019}, PR_{2009}, and PR_{1999} were calculated for the corresponding vertex in these graphs.

Afterwards, all loops (arcs corresponding to journal self-citation) were removed from graphs $Journ\text{-}graph_{2019}$, $Journ\text{-}graph_{2009}$ and $Journ\text{-}graph_{1999}$. The resulting graphs without self-citation are designated $Journ\text{-}graph^{nosc}_{2019}$, $Journ\text{-}graph^{nosc}_{2009}$ and $Journ\text{-}graph^{nosc}_{1999}$. Next, we calculated the values PR^{nosc}_{2019}, PR^{nosc}_{2009} and PR^{nosc}_{1999} for each journal.

For example, for *Matematicheskii Sbornik*, we calculated the following values: $PR_{1999}{=}0.025$, $PR^{nosc}_{1999}{=}0.023$, $PR_{2009}{=}0.053$, $PR^{nosc}_{2009}{=}0.054$, $R_{2019}{=}0.06$, $PR^{nosc}_{2019}{=}0.064$. This example shows that self-citation does not necessarily lead to higher *PR*.

So, we have 6 *PR* vectors calculated for the 1999, 2009 and 2019 graphs with and without journal self-citation.

In the general case, we introduce the following vector notation:

$$<PR>=\{PR(1), PR(2), ... , PR(166)\},$$

where $PR(i)$ is the PageRank of the journal corresponding to vertex i in the *Journ-graph* graph.

Pearson correlation coefficients, calculated in pairs between all six PR vectors, are given in Table 1.

Table 1. Pearson correlation coefficients for *PR* vectors.

	$<PR^{nosc}_{1999}>$	$<PR^{nosc}_{2009}>$	$<PR^{nosc}_{2019}>$
$<PR_{1999}>$	**0.9379**	0.9318	0.9375
$<PR_{2009}>$	0.8923	**0.9524**	0.9581
$<PR_{2019}>$	0.8827	0.9427	**0.9658**

All the vectors were found to have statistically strong positive relationship. The highest values in each row lie on the main diagonal of the table. This indicates a slight difference in the ranking of journals with and without self-citation in 1999, 2009 and 2019. The relationship grows with each decade, i.e. self-citation has less and less influence on journal rankings. Moreover, over 100 journals have approximately the same *PR* values with and without self-citation in 1999, 2009 and 2019.

This suggests that self-citation has little or no influence on journal rankings in the thematic group. This therefore confirms the high reputation of many Math-Net.Ru-registered journals.

3 Citation-based graph

We denote as *Cit-graph* a scientific collaboration graph built on the basis of citation.

The Math-Net.Ru database contains over 1 million paper citations. From these citations, all citations to external papers not registered in Math-Net were removed. Authors' self-citations were also removed. All multiple citations were replaced with one arc. Also, all authors who did not cite other authors or were not cited by other authors were removed from the authors set.

So, the resulting *Cit-graph* is an oriented graph with no loops and multiple arcs, without isolated vertices, and containing 52,728 vertices and 388,654 arcs. The graph vertices are denoted by author identifiers (author_id) in Math-Net.Ru, while the arcs are denoted by the (author_id-source, author_id-target) pairs.

The graph has the following main characteristics: very small density (0.00013), a diameter of 16, an average path length of 5.715, and 7.731 average degree of vertex.

One of the assessments of the significance of vertices is the degree of influence (more precisely, the eigenvector centrality) [6]. A high degree of influence means that a vertex is associated with many vertices that also have high degrees of influence.

Table 2 shows data on the first 7 vertices of the *Cit-graph* graph having the highest degrees of influence.

Table 2. First 7 vertices of the *Cit-graph* graph having the highest degrees of influence.

author_id	Authors	Eigenvector centrality	Incoming arcs	Outgoing arcs
8938	Izrail Moiseevich **Gelfand**	1.0	280	787
4537	Sergei Petrovich **Novikov**	0.99	702	699
4874	Vladimir Igorevich **Arnold**	0.91	307	830
9160	Lyudvig Dmitrievich **Faddeev**	0.82	401	740
8582	Olga Arsen'evna **Oleinik**	0.80	310	280
14007	Andrei Nikolaevich **Kolmogorov**	0.71	230	498
8485	Marko Iosifovich **Vishik**	0.70	207	467

All scientists on the list are mathematicians and their importance in the system of Russian mathematics is beyond doubt.

Connectivity is an important characteristic of a network. In our case, if we take orientation of arcs to be irrelevant, we get a maximal connected component (CC) containing 50,931 vertices, and the second largest component containing 24 vertices. In total, the graph contains 504 connected components. That is, there is a giant component and several small components that are not connected to it and to each other.

In the case where orientation of arcs is important, we get strong connectivity. The maximal strongly connected component (SCC) contains 21,108 vertices with a diameter of 16, and average length of (directed) path 5.29.

The second largest SCC contains only 19 vertices. Another 841 components contain from 2 to 13 vertices, and over 29,000 components consist of one "dangling" vertex.

Several verified SCCs of four to six vertices show that these are typically authors working at the same institution who have coauthored two or more papers, who, in their subsequent papers, have cited at least one previous paper.

A curious feature of real networks is that they have clustering properties based on which the graph topology is organized into communities (also called modules or clusters) [7]. One theoretically well-founded approach is called density-based clustering. Here, the modularity measure shows how qualitative the given split is in the sense that there are many arcs within communities, and few arcs outside communities. Modularity is used to determine the quality of graph splitting into communities. Here, we use the definition of modularity measure Q from [8]. The Q value lies in the $[-1,1]$ range. A Q value greater than 0.7 is considered a good split.

For the *Cit-graph* graph, the algorithm proposed in [9] gives $Q=0.722$. The maximum community contains 12,796 vertices, the total number of communities is 605, the smallest communities contain 2 vertices each.

If only a subgraph consisting of the maximal CC is left in the *Cit-graph* graph, the modularity measure remains almost the same, $Q_{CC}=0.724$. The number of communities decreases to 90, the maximum community contains 12,045 vertices. The smallest communities, numbering 10 in total, contain 3 vertices each.

If only a subgraph consisting of the maximal SCC is left in the *Cit-graph* graph, you get Q_{SCC}=0.687 and 24 communities of 3 to 5,472 vertices. There is only one community of 3 vertices; all its members work in one organization.

The next largest community contains 11 vertices. It contains 8 authors from Moldova, 2 from Spain and 1 from Canada. All citations here are made in papers published in the journal *Buletinul Academiei de Stiinte a Republicii Moldova. Matematic*a for papers from the same journal. Large communities defy meaningful interpretation just as easily.

4 Coauthorship-based graph

We denote as *Co_auth-graph* a scientific collaboration graph built on the basis of co-authorship. It contains over 105,000 authors and more than 340,000 coauthorship cases, i.e. pairs of authors in the event that they coauthored at least one paper. Thus, *Co_auth-graph* is an undirected graph with no loops and multiple arcs, without isolated vertices, and containing 105,327 vertices and 340,643 edges. As before, the graph vertices are denoted by author identifiers (author_id) in Math-Net.Ru, while the edges are denoted by the (author_id-source, author_id-target) pairs.

The graph has the following main characteristics: very small density (0.00006), a diameter of 24, an average path length of 19.062, and an average vertex degree of 6.468.

The degrees of vertices indicate the number of coauthors of a given author. Academician A.M. Prokhorov has the highest degree of vertex (798) in this graph.

Table 3 shows the degrees of influence (eigenvector of centrality) for the first 7 vertices having the highest degrees of influence.

Table 3. First 7 vertices of the *Co_auth-graph* graph having the highest degrees of influence.

author_id	Authors	Eigenvector centrality	Number of co-authors
4537	Sergei Petrovich **Novikov**	1.0	532
74733	Aleksandr Mikhailovich **Prokhorov**	0.959	798
44810	Aleksandr Nikolaevich **Skrinsky**	0.923	408
45287	Oleg Nikolaevich **Krokhin**	0.894	414
4406	Yurii Sergeevich **Osipov**	0.866	320
26158	Vladimir Evgenevich **Fortov**	0.839	611
21689	Evgenii Pavlovich **Velikhov**	0.837	399

Note that out of the first 7 vertices in the *Cit-graph* graph, only Novikov was among the first 7 vertices of the *Co_auth-graph* graph. Also, in *Cit-graph*, the first 7 vertices are entirely represented by mathematicians, while in *Co_auth-graph*, you have 3 mathematicians and 4 physicists.

The maximal CC in the *Co_auth-graph* graph contains 79,517 vertices, and the second largest component contains 78 vertices. In total, the graph contains 7,939 connected

components. As with *Cit-graph*, there is a giant component and even more small components that are not connected to it and to each other. Moreover, there are 4,115 components with two vertices, 1833 with three vertices, 849 with four, 408 with five, and so on in descending order. In total, "small" CCs contain almost a quarter of all the vertices of the graph.

We take a closer look at the 78-vertex component. Most of it is made up of vertices identifying colleagues from Samara State University and several other scientists from other institutions. They conduct research in the field of organic chemistry, which are published in the multidisciplinary *Samara University Bulletin. Natural Science Series*, included in Math-Net.Ru.

Several randomly selected CCs of 4 to 6 vertices show that they are most often authors working at the same institution who have coauthored at least a paper. Although this example cannot be absolutized, there are instances of author groups from different universities in the same city, and sometimes (very rarely) from different cities and countries.

For the *Co_auth-graph* graph, the algorithm proposed in [9] gives $Q=0.857$. Obviously, such a high Q value is due to the presence of almost 8,000 connected components: the graph simply "splits" into disconnected parts.

The maximum community contains 11,502 vertices; there is a total of 8,229 communities; the smallest communities numbering 4,115 contain 2 vertices each, another 1,833 communities contain 3 vertices, etc. Over 20 communities have from 1,000 to almost 4,000 vertices. Communities formed in the maximal CC are of interest when it comes to meaningful interpretation.

For the maximal connected component of the *Co_auth-graph* graph, $Q_{CC}=0.842$, i.e. the tendency towards division into communities remains high. The maximum community is large enough (10,346 vertices), but the total number of communities decreased to 294. At the same time, all communities with 2 or 3 vertices disappeared, there were 27 communities of 4 vertices each, 24 of 5 vertices, and 19 containing 6 vertices each.

5   Splitting the Math-Net.Ru authors set

The size of the set of all authors of papers registered on Math-Net.Ru is slightly more than 119,000. We have the *Cit-graph* and *Co_auth-graph* graphs, characterizing collaboration between Math-Net.Ru authors through citation and co-authorship.

Therefore, the entire authors set can be split into four disjoint subsets:

(1) authors who have not co-authored any paper with someone, have not cited anyone and/or have not been cited by anyone,

(2) authors who have not co-authored any paper with someone, have cited someone and/or have been cited by someone,

(3) authors who have co-authored a paper with someone, have cited someone and/or have been cited by someone,

(4) authors who have co-authored a paper with someone, have not cited anyone and/or have not been cited by anyone.

To do this, we build a graph called *Combi* according to the following rule. We take the set of vertices of the *Cit-graph* graph and impose on it a set of arcs from the *Co_auth-graph* graph without adding the vertices that are not in the *Cit-graph* graph. The resulting graph is an undirected graph containing 52,728 vertices and 180,208 arcs. We now remove all isolated vertices. The resulting structure is a graph with the same number of arcs and 46,647 vertices. Vertices of the *Combi* graph correspond to authors who are co-authors of papers in Math-Net.Ru and at the same time cite other authors and/or are cited by other authors from Math-Net.Ru.

In this case, the structure of the authors set on co-authorship and citation can be represented in the same way as in Fig. 1. Set 1 contains 11.5% of 119,000. Set 2 represents authors who have cited other authors and/or have been cited by other authors but have not co-authored any paper with someone. Set 3 consists of authors who have co-authored papers in Math-Net.Ru and at the same time, have cited other authors and/or have been cited by other authors (in fact, this is the vertex set for the *Combi* graph). Set 4 contains authors who have co-authored at least one paper, but have not cited any author or have not been cited by someone.

The union of sets 2 and 3 consists of a set of the vertices of the *Cit-graph* graph, while the union of sets 3 and 4 represents the set of vertices of the *Co_auth-graph* graph.
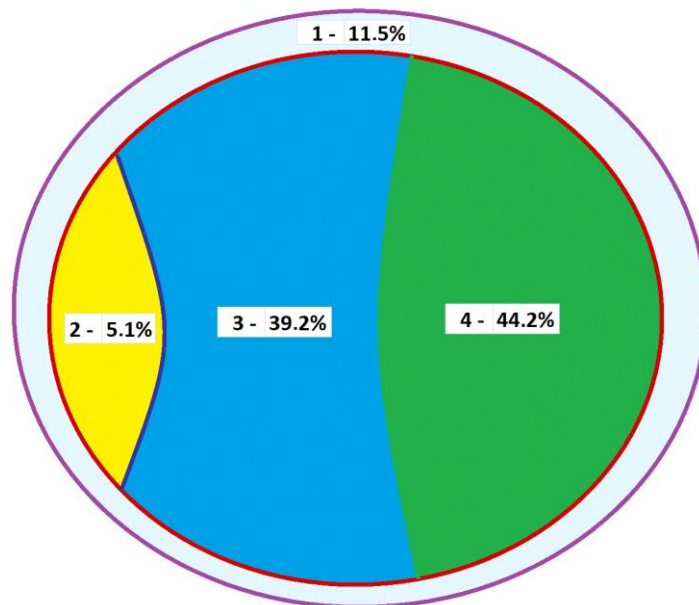

Fig. 1. Split set of Math-Net.Ru authors.

6   Discussion and conclusions

The *Cit-graph* and *Co_auth-graph* graphs were built using different approaches. The first one is oriented, while the second one is not. The cardinality of their sets of vertices differ significantly. However, they have similar characteristics – very small density, large diameter and high modularity. In both graphs, we observe a similar connectivity structure: the presence of a giant component containing tens of thousands of vertices, the next largest component is a thousand times smaller than the giant one, and a large number of small connected components. Obviously, this is an indicator that "acquaintance by citation" and "acquaintance by co-authorship" are more often found in small groups of authors.

This also explains why we have large modularity coefficients. Graphs with this connectivity structure obviously tend to cluster into a large number of loosely connected or disconnected communities.

There is no doubt that the discrepancy between personalities, which correspond to the most significant vertices in the *Cit-graph* and *Co_auth-graph*, is due to the different approaches used in building the graphs. Math-Net.Ru journals are characterized by a high degree of mutual citation among mathematicians and co-authorship among physicists.

Meaningful analysis of some of the smaller connected components and communities may lead to misconceptions that their structure is only due to personal relationships among the authors. This thesis is undoubtedly true. However, the constructed splits of the set of Math-Net.Ru authors shows that over 40% of authors who have co-authored a paper with someone do not have a citation-related relationship, and almost 17% of them have not coauthored any paper with someone, although most of them have cited other authors. This is a largely substantial evidence that there is no unmotivated deliberate mutual citation in Math-Net.Ru journals [10].

In conclusion, it should be noted that the Math-Net.Ru portal contains limited amount of information even on Russian journals in the field of mathematical sciences. References have been indexed since 2000, and selectively processed for earlier publications. This certainly affects the results obtained in this work, making them differ from those that could be arrived at in a potential "global" study. Nonetheless, that the Math-Net.Ru journals represent a good sample of many mathematical periodicals in Russia should not be underestimated.

References

1. de Solla Price, D.J.: Networks of scientific paper. Science 149(3683), 510–515 (1965).
2. Kas, M., Carley, K.M., Carley, L.R.: Trends in science networks: understanding structures and statistics of scientific networks. Social Network Analysis and Mining, 2, 169–187 (2012).
3. Chebukov, D., Izaak, A., Misyurina, O., Pupyrev, Yu., Zhizhchenko, A.: Math-Net.Ru as a digital archive of the Russian mathematical knowledge from the XIX century to today. Lecture Notes in Computer Science, 7961, 344–348 (2013).

4. Newman, M.E.J.: The structure of scientific collaboration networks. Proceedings of the National Academy of Sciences of the USA, 98 (2), 404–409 (2001).

5. Brin, S., Page, L.: The Anatomy of a Large-Scale Hypertextual Web Search Engine. Computer Networks and ISDN Systems, 30, 107–117 (1998).

6. Easley, D., Kleinberg, J.: Networks, Crowds, and Markets. Cambridge University Press (2010).

7. Malliaros, F.D., Vazirgiannis, M.: Clustering and community detection in directed networks: A survey. Physics Reports, 533(4), 95–142 (2013).

8. Newman, M.E., Girvan, M.: Finding and evaluating community structure in networks. Physical Review E, 69(2), P 026113 (2004).

9. Blondel, V.D., Guillaume, J-L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, P 10008 (2008).

10. Gulyova, M.: Radost vzaimnogo citirovaniya. Troickii variant – Nauka 287, 3 (2019).