Co-authorship network structure analysis

N G Scherbakova¹ and S V Bredikhin²

¹ Senior Researcher, Institute of Computational Mathematics and Mathematical Geophysics SB RAS, Ac. Lavrentiev ave., 6, Novosibirsk, 630090, Russia
² Leading Researcher, Institute of Computational Mathematics and Mathematical Geophysics SB RAS, Ac. Lavrentiev ave., 6, Novosibirsk, 630090, Russia

E-mail: scherbakova@sscc.ru

Abstract. The analysis of networks of collaboration between scientists reveals features of academic communities that help in understanding the specifics of collaborative scientific work and identifying the notable researchers. In these networks, the set of nodes consists of authors and there exists a link between two authors if they have coauthored one or more papers. This article presents an analysis of the co-authorship network based on bibliometric data retrieved from the distributed economic database. Here we use the simple network model without taking into account the strength of collaborative ties. The data were analyzed using statistical techniques in order to get such parameters as the number of papers per author, the number of authors per paper, the average number of coauthors per author and collaboration indices. We show that the largest component occupies near 90 % of the network and the node degree distribution follows a power-law. The study of typical distances between nodes and the degree of clustering makes it possible to classify the network as a ‘small world’ network.

1. Background
One class of complex networks is ‘collaboration networks’, constructed of groups of actors with common interests, such as athletes in the same sport; film actors participating in the general film. This also includes groups of scientists who jointly prepared one or more scientific publications (SP) – co-authors. Typically, these SPs are the result of well-documented scientific collaboration. If more than one author shares responsibility for a SP containing the results of specific research, then it is assumed that during the research process the coauthors must work together for some time. In a number of cases, co-authorship guarantees an effective solution to the problems posed in the SP, reducing the number of delusions and errors and, ultimately, ensuring the effectiveness of scientific activities.

For the reviews of factors affecting behavior, performance and motivation of scientific collaborations see works [1, 2] and the monograph [3]. The study of co-authorship networks sheds light on the degree of interpenetration of ideas, knowledge and technologies and characterizes the structure of the scientific community. In this case, different levels of cooperation can be considered: between countries, within a country, within an organization, between scientific areas, within a general subject, within a journal or conference proceedings, etc.

Citation networks and co-authorship networks represent different points of view on the SP-set. The source of information about SPs, as a rule, are bibliographic databases (DB). When studying a citation network, SPs are its nodes, and oriented links between nodes are established according a binary citation relation; in co-authorship networks, nodes are SP-authors, and links are established by their co-authorship of one or more SPs. However, there is a mutual influence of collaboration and citation...
practices (see [4–6]). For example, issues such as the impact of the number of researchers co-authoring a paper on its efficiency score or do authors more often cite their coauthors in comparison with other cited authors.

The study of collaboration networks is carried out in two main directions. Empirical measurements provide detailed network characteristics both at the network level and at the individual node level. An example of this approach is a series of papers [7–9] that present the results of a study of several large databases related to a number of scientific fields. It is shown that the corresponding co-authorship networks have common properties, such as a short node-to-node distance and a large clustering coefficient [10]; the degree distribution follows a power law. Another approach is to build network models and study their properties. Collaboration networks are modeled as undirected graphs, directed graphs, and hypergraphs and various coauthor weighting schemes are applied. The study of the dynamic properties of real networks and network models makes it possible to identify the structural principles that govern the evolution of networks; dynamic properties, in turn, can explain static ones [11, 12].

This paper presents a method for constructing a co-authorship network based on bibliographic information extracted from the RePEC database. A preliminary analysis of the initial data is presented. The parameters of the constructed co-authorship network are investigated and compared with the parameters of the “small world” network.

Co-authorship networks reflect professional interaction between scientists. Their research reveals much about the structure of scientific communities, allows predicting the results of their activities and to build maps of science.

2. Network constructing

Co-authorship networks are most commonly defined as undirected and weighted graphs, here we use this approach. Let $P$ be the set of papers indexed in the database, $P = \{p_1, p_2, \ldots, p_m\}$ and $V = \{v_1, v_2, \ldots, v_n\}$ the set of authors appearing in $P$. Let $E$ be the set of pairs $e = (v_i, p_j) \in E$ if $v_i$ is the author or co-author of $p_j$. The network of publications (network of authorship) $N^{pub} = (V, P, E)$ is a bipartite directed network, its nodes are divided into two disjoint independent sets $V$ and $P$. The adjacency matrix $A = (a_{ij})$ (authorship matrix) of the corresponding graph is an $n \times m$ matrix whose rows correspond to authors and columns to publications:

$$a_{ij} = \begin{cases} 1, & \text{if } v_i \text{ is the author of } p_j; \\ 0, & \text{otherwise.} \end{cases}$$

Let $l(p_j)$ be the number of authors of $p_j$:

$$l(p_j) = \sum_{i=1}^{n} a_{ij}. \quad (2)$$

We assume that each SP analyzed below has at least two authors (co-authors), i.e. $l(p_j) > 1$ for each publication $p_j$. For $v_i, v_j \in V$, let’s define the binary co-authorship relation $R^a$:

$$v_i R^a v_j = (\exists p \in P), \quad v_i \text{ and } v_j \text{ are the authors of } p. \quad (3)$$

A co-authorship network is a weighted undirected network $N^a = (V, E, w)$, $V$ is the set of authors, $E \subseteq V \times V$ is the set of weighted edges, $e = (v_i, v_j) \in E$ if $v_i R^a v_j$ holds. Several weighting schemes are commonly used to assign weights to co-authorship links. In the simplest case, we are considering in this article, $w(e = (v_i, v_j)) = 1$, regardless of how many papers $v_i$ and $v_j$ have coauthored and how each author contributed to the publication, i.e., the $N^a$ network can be considered unweighted. Then the graph corresponding to the unweighted co-authorship network can be represented by $(0, 1)$ adjacency matrix $U = (u_{ij})$:

$$u_{ij} = \begin{cases} 1, & \text{if } v_i R^a v_j; \\ 0, & \text{otherwise.} \end{cases}$$

(4)
The matrix $U$ can be obtained from the matrix $A' = A \cdot A^{+}$ by replacing all nonzero elements with 1. Elements of main diagonal are set to 0.

3. Empirical research

3.1. Data extracting

The problem of recognizing individual authors based on the text of SP in a database of considerable size has been recognized as rather laborious and depends on how well the database is documented. The task includes not only parsing the headlines of articles and extracting such components as title, author, affiliation but also identifying authors. Two authors may have the same name, or the same author may be identified differently in different publications (full name or initials). This leads to ambiguity and erroneous conclusions. One way to avoid ambiguity is to exclude individuals with the same last name and first initial, see [13]. For pre-processing on the raw data in order to associate names of authors with unique people see [6]. In our case, metadata contained in the RePEc database [14] was used when selecting the authors and publications. Authors create individual ‘profiles’ using Author Service provided by the DB RePEc. The author's profile contains publication IDs, authored or coauthored by this author. For the date of data extracting (2020.01.31) the number of registered authors is 70549 and the number of declared publications is 470333. Among the publications there are those that are declared two or more times, we consider that these publications have more than one author. We select only publications declared by several authors, and the number of these authors coincides with the information about the number of authors contained in the publication's metadata. As a result, 91113 publications and 32434 authors that have declared these publications are considered. In total, the selected authors declared 364 989 publications (including those with one author). This filtered data is used as the basis for constructing the co-authorship network. The publications refer to the period from 1954 to 2019. It should be noted that the creation of the DB RePEc dates back to 1997.

3.2. Data analysis

Let $P$ be the set of papers and $A$ the set of authors or coauthors of at least one of these papers. Let $P_a$ be the set of publications having $a$ coauthors, $P_1$ the set of coauthored publication and $a$ the maximum number of authors in a single paper. So $P = P_1 \cup P_2; |P| = \sum |P_j|$. Table 1 presents some basic co-authorship statistics for the network studied here.

| Parameter                          | Value |
|-----------------------------------|-------|
| Number of authors $|A|$              | 32434 |
| Number of papers having coauthors $|P_a|$ | 91113 |
| Number of papers having one author $|P_1|$ | 273866 |
| Maximal number of authors in a single paper $a$ | 17    |
| Co-authorship frequency set       | \{1, 2, 3, …, 12, 15, 16, 17\} |
| Mean authors per paper (for papers of $P_a$) | 2.3   |
| Mean papers per author (for papers of $P_1$) | 2.8   |

Around 77% of the coauthored papers have two authors, and around 20% have three authors. It should be noted that the average number of authors per paper largely depends on the scientific discipline. An example for the most significant groups looks like this (number of coauthors, number of papers): (2, 70399); (3, 18073); (4, 2245); (5, 251). The histogram of the number of papers written by authors follows a power law $p_x \sim C \cdot x^{-\gamma}$ with parameters $x_{min} = 6$, $\gamma = 1.64$. That is, the most authors have a small number of coauthored papers. Note that the total number of papers in the database grows
with time, including the number of papers published in co-authorship (period, number of papers): (1954–1989, 3964); (1990–1999, 10781); (2000–2009, 33288); (2010–2019, 42941).

3.3. Co-authorship network parameters
We construct the unweighted network of co-authorship $N^{ca} = (V^{ca}, E^{ca})$ according (1) – (4) and analyze its structure. Table 2 gives a summary of the basic results.

**Table 2. Parameters of the co-authorship network $N^{ca}$.**

| Parameter                           | Value / Description                                      |
|-------------------------------------|----------------------------------------------------------|
| Order $n = |V^{ca}|$                      | 32434                                                    |
| Size $m = |E^{ca}|$                        | 73969                                                    |
| Degree distribution                 | Power law: $x_{\text{min}} = 4$, $\gamma = 1.3$. See Fig. 1 below. |
| Mean node degree ($k$)              | $\approx 4.6$                                           |
| Number of components                | 1242                                                     |
| Giant component dimensions:         | 29270, 71780, $\approx 4.9$                            |
| Second largest component order      | 25                                                       |
| Edge density [15]:                  | 0.000168                                                 |
| Radius                              | 13                                                       |
| Diameter                            | 25                                                       |
| Mean distance $L(N^{ca})$ [16]      | 6.577475                                                 |
| Local clustering coefficient $C(N^{ca})$ [10] | 0.264428                                              |

**Fig. 1.** Histogram of the number of coauthors of scientists. The abscissa shows the number of coauthors. The ordinate shows the proportion (in $\%$) of authors with a given number of coauthors; the scales are logarithmic.

Thus the single giant component of the network makes up around 90\% of the network. The average distance ($\approx 6.5$) is similar to that of a random graph with the same number of nodes and edges ($\approx 6.7$); the clustering coefficient ($\approx 0.3$) is much larger than that of a random graph ($\approx 0.0001$). The distribution of the numbers of collaborators of authors follows a power law $p_x \sim C \cdot x^{-\gamma}$ with parameters $x_{\text{min}} = 4$, $\gamma = 1.3$. 


Summary
The co-authorship network of scientists in economics $N^{ca}$ has been analyzed by using the bibliometric information retrieved from the distributed DB RePEc. In the collection of papers under consideration multi-authorship is less as compared to sole authorship (around 25 %), and, the prevailing trend is the presence of two authors (77 % of papers).

The statistical network analysis is based on a simple network model representing the co-authorship network as an undirected unweighted graph. We find that the giant component of $N^{ca}$ fills a large portion of the network (around 90 % of the total volume). Thus, most authors at the time of the study are indirectly connected via collaboration. The edge density in the giant component is quite low. The second largest component is far smaller than the giant component. The distribution of the number of coauthors of scientists follows a power law; moreover, this distribution follows a similar power-law form to the distribution of papers.

An important property of many real networks is their ‘small-worldness’. We find that the giant component of the network constitutes a ‘small world’ [17] in which the average distance between nodes varies logarithmically with the number of nodes. According [10] we define ‘the small-worldness’ on the base of the trade off between the mean shortest path length that is similar to that of a matched random graph and the local clustering coefficient, which is higher than that of a random graph.

The relationship between co-authorship and citation needs to be further studied, namely, whether papers published in co-authorship receive more citations than those published by a single author. The parameters of the weighted co-authorship network, for example, the stability of coauthored groups of authors, are also of our interest.

Acknowledgement
This work was carried out under state contract with ICMMG SB RAS (0251-2021-0005).

References
[1] Beaver D D 2001 Reflections on scientific collaboration (and its study): past, present and future Scientometrics 52 pp 365–37
[2] Laudel G 2002 What do we measure by co-authorships? Research Evaluation 11 pp 3–15
[3] Stephan P 2012 How Economics Shapes Science (Cambridge, Mass.: Harvard University Press)
[4] Ding Y 2011 Scientific collaboration and endorsement: Network analysis of coauthorship and citation networks J. Informetrics 5 pp 187–203
[5] Wallace M L, Larivière V and Gingras Y 2012 A smallworld of citations? The influence of collaboration networks on citation practices PLoS ONE 7(3) 0033339
[6] Martin T, Ball B, Karrer B and Newman M E J 2013 Coauthorship and citation patterns in the Physical Review Phys. Rev. E 88 012814
[7] Newman M E J 2001 The structure of scientific collaboration networks Proceedings of the National Academy of Sciences 98(2) pp 404–9
[8] Newman M E J 2001 Scientific collaboration networks. I. Network construction and fundamental results Phys. Rev. E 64 016131
[9] Newman M E J 2001 Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality Phys. Rev. E 64 016132
[10] Watts D J and Strogatz S H 1998 Collective dynamics of ‘small-world’ networks Nature 393 pp 440–2
[11] Barabási A-L et al 2002 Evolution of the social network of scientific collaborations Physica A: Statistical Mechanics and its Applications 311(3) pp 590–614
[12] Savić M et al 2014 The structure and evolution of scientific collaboration in Serbian mathematical journals Scientometrics 101(3) pp 1805–1830
[13] Youngblood M and Lahti D 2018 A bibliometric analysis of the interdisciplinary field of cultural evolution Palgrave Communications 4(1) pp 1–9
[14] RePEc project http://repec.org/
[15] Distel R 2005 Graph Theory (New York: Springer-Verlag Heidelberg) chapter 7 p 164
[16] Watts D J 1999 Networks, dynamics and the small-world phenomenon American Journal of Sociology 105(2) pp 493–527
[17] Milgram S 1967 The small world problem Psychology Today 2(1) pp 60–7