Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Computer Science 00 (2016) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

13th International Conference on Current Research Information Systems, CRIS2016, 9-11 June 2016, Scotland, UK

Scientometric indicators and collaboration network as a potential tool for *gift author* finding

T S Shinyaeva*, Yu Yu Tarasevich

Astrakhan State University, 20a Tatishchev Street, Astrakhan, 414056, Russia

Abstract

In our research, we were looking for some criteria detecting incorrect lists of co-authors. We focused our efforts on a *gift author* problem. The term *gift author* describes a situation where a person is listed as a co-author but this person has an indirect relationship to the work. Scientometric analysis of the main publications of the Russian dissertations is performed. The open database of dissertations (website http://vak.ed.gov.ru/dis-list, websites of organizations, where dissertations were defended) and CRIS of Astrakhan State University (ASU) http://science.asu.edu.ru were used for the research.

Peer-review under responsibility of the Organizing Committee of CRIS2016.

Keywords:
gift author; publication ethics; co-author networks; scientometrics; CRIS

1. Introduction

The problem of compliance with the principles of publication ethics in the preparation of scientific publications is acute. Committee on Publication Ethics (COPE) http://publicationethics.org and International Committee of Medical Journal Editors (ICMJE) http://www.icmje.org/ deal with some practical aspects of publication ethics. CrossCheck (http://www.ithenticate.com/) and its Russian analog Antiplagiat (https://www.antiplagiat.ru/) provide a direct check of incorrect borrowing (plagiarism) in scientific publications. Nevertheless, the publication ethics is not limited to plagiarism. Violations of the principles of publication ethics also include the incorrect list of co-authors, data fabrication and falsification, multiply publications and 'salami slicing', i.e. the 'slicing' of research that would form one meaningful paper into several different papers (http://www.icmje.org/icmje-recommendations.pdf).

In our research, we were looking for some criteria detecting incorrect lists of co-authors. We focused our efforts on a *gift author* problem. The term *gift author* describes a situation where a person is listed as a co-author but this person has an indirect relationship to the work, i.e.

*Corresponding author. Tel.: (8512) 61-09-25 ; fax: (8512) 49-41-57.
E-mail address: taisiya.danilova@asu.edu.ru

1877-0509
Peer-review under responsibility of the Organizing Committee of CRIS2016.

- his/her contribution to the obtaining scientific results, to the interpretation of the results, and to preparation of the work is insignificant
- or this contribution is purely technical or organizational, but not scientific.

Worthy of note that Committee of Ethics of Scientific publications (Russia) http://publicet.org/code/ offers rather uncertain definition of co-authorship: 'Co-authors of the article should include all the persons who have made a significant contribution to the study. The persons who are not involved in the study cannot be indicated among the co-authors.' This formulation opens up wide possibilities for manipulating, since the 'significant contribution to the study' is not specified. In our research, we used the recommendations of ICMJE.

In some cases, a large list of the co-authors is not only justified but also inevitable, it itself may not indicate malpractice in the specifying of the co-authors of the publication. We perform scientometric analysis of the main publications of the dissertations. The open databases of dissertations (website http://vak.ed.gov.ru/dis-list, websites of organizations, where dissertations were defended) and CRIS of ASU http://science.asu.edu.ru are used for the research. Based on the analysis, we try to develop and offer the quantitative scientometric criteria for determining 'risk groups', i.e. dissertations in which all formal requirements are met, but the real scientific contribution of the applicant may be negligible. We designed and implemented a specialized software module for CRIS http://science.asu.edu.ru; this module provides automated collection of information, its analysis and visualization.

We took in our consideration only the journal articles concerning the dissertations. We analyzed the co-author networks and scientometric indicators of the articles and co-authors. Our investigation based on the hypotheses that *gift author* can be detected by analysis of objective data (co-authorship networks, scientometric indicators, etc.). This detection can be formalized and automated.

Additional criteria for the selection of dissertations in the risk group presumably are

- all articles are publication in the same journal, or in the several journals with low scientometric indices when there is a large number of journals concerning the subject of the research,
- an abnormally high number of the applicant's publications during a year,
- an abnormally high percentage of self-citations and citations by co-authors,
- an abnormally high fraction of short articles,
- the presence of the 'twin articles',
- research topic does not imply the presence of the co-authors (the study of one object by one method).

Practical application of our research results is the use of developed scientometric criteria for identifying dissertations, the main results of which were published in articles produced with violations of the principles of publication ethics.

Scientometric and statistical methods are used for the data analysis of

1. The database which includes
    - metadata of the dissertations,
    - metadata of the journal articles in which the main results of dissertations were published
    - scientometric indicators of the journals and publications.
2. Collaboration network.

2. Proposed approach for detecting risk groups

We used the following scientometric indicators: average number of authors per article, average length of articles (in pages), dissertation's rating, rating of dissertation's author. Dissertation's rating (R_d) and rating of dissertation's author (R_{da}) are calculated by the following formulas [1]

$$R_d = \sum_{i=1}^{n} IF_i \frac{NP_i}{TNP_i},$$

$$R_{da} = \sum_{i=1}^{n} IF_i \frac{NP_i}{TNP_i \cdot NCA_i},$$

where n is the number of publications concerning dissertation, IF_i is the journal impact factor in which i-th article was published, NP_i is the number of i-th article's pages, TNP_i is the typical article length (10), NCA_i is the number of co-authors of the i-th article. When a journal has not impact factor, then similar indicator, i.e. Cites/Doc. (2 years), calculated by Scimago http://www.scimagojr.com/journalrank.php was used. Since a large number of journals published in Russia have neither impact factor (WoS) nor Cites/Doc. (Scimago), then $IF = 0.01$ for journals included in WoS or Scopus DB's, $IF = 0.005$ for journals included in Russian Science Index, $IF = 0.001$ for other journals.

Study of real networks allowed to reveal their features. For example, most social networks are highly clustered and shows positive correlation between degrees of neighbour nodes (assortative mixing)[2]. Different network characteristics[3,4] are used for anomaly detection: for example, to discriminate real from imaginary social networks[5] or the real scientific manuscripts from the artificially generated[6].

In our research, we used the following parameters: assortativity, clustering coefficient and density. For the calculation of the assortativity coefficient we used the formula from[7].

We took in our consideration only the journal articles concerning the dissertations. Thus, we expected that co-authors networks built from the considered data, will be different from most social networks. Most likely that network consisting of supervisors and PhD students will be disassortative because of greatest probability that PhD student's article will be written in collaboration with the supervisor than with other PhD student.

3. Results

In Russia, there is no complete database of publications concerning dissertations. A thesis abstract contains a list of related publications. We took in our consideration only the journal articles, but requirements for mandatory publication of articles appeared only in 2002 for habilitation thesis and in 2006 for PhD thesis. Thus, dissertations that were defended previously do not always have associated journals articles. Moreover, absence of some abstracts in electronic format does not allow us to get full database.

In our research, we dealt with dissertations of ASU's employees. We processed about 143 dissertations and about 855 publications which have been mentioned as the results of the dissertations. It was possible to uniquely identify the co-author of publication in the CRIS of ASU only in the case, when the co-author is a registered user (employee of the University). So, the co-authors which are not employees of the University were not included in the study.

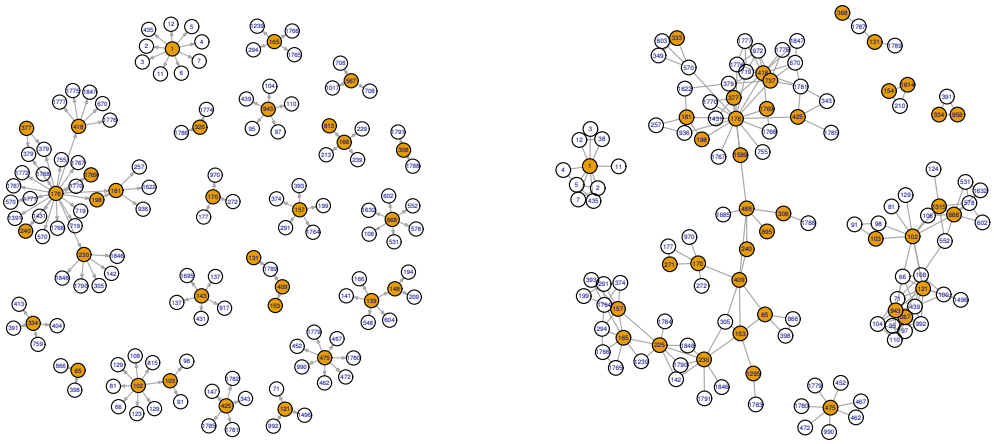
Some characteristics of the Humanities have been detected. High fraction of publications without co-authors (higher than the world level)[8]. Thus, in Humanities in Russia there is no *gift author* problem, and we excluded dissertations in the humanities from consideration. Moreover, some additional features were found, e.g.

- low fraction of publications in scientific journals, the results are mostly published in the Proceedings and books,
- small fraction of publications in international journals,
- high fraction of publications in local journals,
- presence of publications in non-core editions.

Relations between supervisors and their PhD students is shown in Fig. 1a. The only those supervisors who have two or more PhD students are shown. Different clusters in Fig. 1a have different topologies, i.e. star, chain, tree, and hybrid topology. A star topology represents a supervisor with one generation of PhD students. A chain topology occurs when a dissertation is an interdisciplinary and one PhD student has several (usually two) supervisors. A tree topology connects one star network to other star networks. So, we can see several generations of PhD students. A hybrid topology is a combination of two or more other network topologies.

We also built a network which nodes are authors of the dissertations and their supervisors and links points to co-authors of publications related to the dissertations (Fig. 1b). Single nodes and pairs of nodes were excluded from the network. Connected components on the Fig. 1a and Fig. 1b does not match. Several separate networks are merged in the following cases

- PhD student has common publications with not his/her supervisor,



(a) (b)

Fig. 1. (a) Relations between supervisors and PhD students. Arrow goes from a supervisor and points to a PhD student. Nodes representing supervisors are colored in orange. Node's label is an employee's identification number. (b) Network of co-authors publications related to the dissertations. Node's label is an employee's identification number. Nodes representing supervisors are colored in orange.

- two supervisors have some common publications,
- two PhD students have some common publications.

Also, using information about publications, we can find dissertations whose main results were published in the same articles (Fig. 2a). There are several types of sharing articles

- two dissertations belongs to the same author, i.e. PhD dissertation and the habilitation thesis (inaugural dissertation),
- one of the two dissertations belongs to the supervisor and other one belongs to his/her PhD Student,
- two dissertations are written under the supervision of the same person,
- two dissertations are written under the supervision of the different persons.

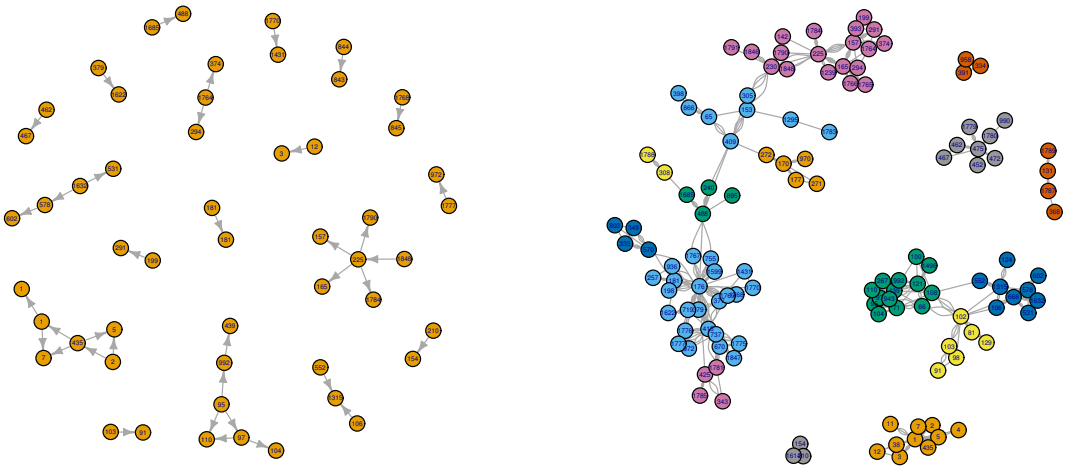
We considered separated networks from Fig. 1b. Big clusters were divided into groups according to the community structure algorithm of Girvan and Newman[9] (Fig. 2b). When we considered the clusters individually, we got the other partitions, which are not shown but which were included in our consideration.

For each group (except small groups), the following parameters were calculated, i.e. total number of articles, the number of unique articles published on the topics of dissertations, average number of articles co-authors, average length of articles (in pages), number of articles in journals published by ASU, average rating of dissertations, average rating of dissertations' authors, assortativity coefficient, clustering coefficient and graph density (Table 1). Table 1 divided into three sections. The first section corresponds to the clusters from Fig. 1b, the second section corresponds to the clusters from Fig. 2b and the last section corresponds to partitions which we got when considered the clusters individually.

When we consider only the journal articles concerning dissertations, network consisting of supervisors and PhD students are expected to be disassortative. Nodes of high degree (supervisors) will be connected with less linked nodes (PhD students). Negative assortativity coefficient was obtained for all groups except groups 3 ($r = 0.001$), 6 ($r = 0.083$), 10 ($r = 1$). The group 10 is unrepresentative because of small number of dissertations. Group 6 is a subset of group 3 (Fig. 3). The largest number of publications per dissertation (10.5) was found in group 6. Group 17 (which is a subset of group 6) has highest density (Fig. 4). Visible anomalies in scientometric indicators were not detected.

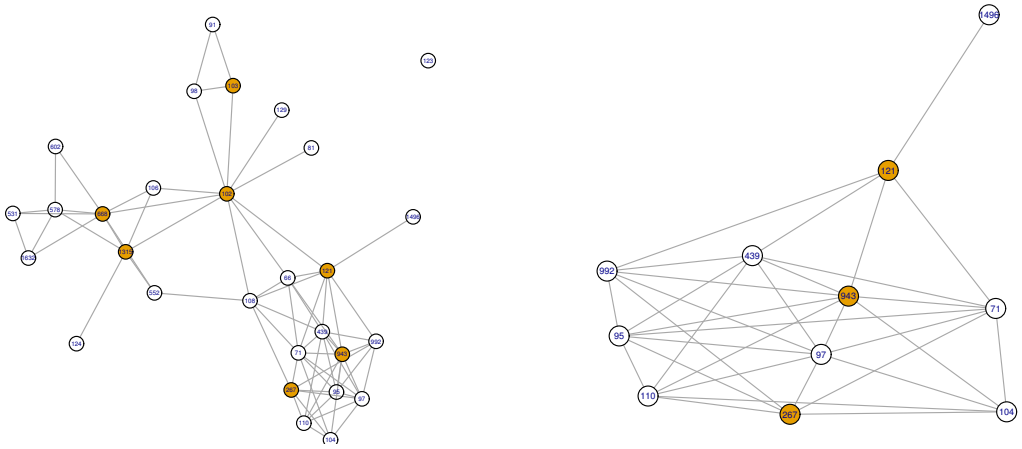
Table 1. Collaboration Network. N_{all} is the total number of articles; N is the the number of unique articles published on the topics of dissertations; $\langle NCA \rangle$ is the average number of articles co-authors; $\langle NP \rangle$ is the average length of articles (in pages), J_{ASU} is the number of articles in journals published by ASU, ND is the number of dissertations (the number of dissertations without articles is indicated in parentheses), R_d is the average rating of dissertations, R_{da} is the average rating of dissertations' authors, r is the assortative, C is the clustering coefficient, D is the graph density; r^*, C^*, D^* are the indicators for the network, taking into account not only the supervisors and candidates but all co-authors.

	N_{all}	N	$\langle NCA \rangle$	$\langle NP \rangle$	J_{ASU}	ND	R_d	R_{da}	r	C	D	r^*	C^*	D^*
1	29	24	2.417	7.375	10.5% (2)	9	0.745	0.363	−0.498	0.437	0.356	−0.436	0.42	0.345
2	35	34	2.471	4.618	28.6% (8)	7	0.01	0.005	−0.859	0.256	0.286	−0.643	0.419	0.289
3	186	166	2.645	6.03	45.7% (70)	31	0.145	0.054	0.001	0.219	0.175	−0.171	0.227	0.15
4	352	331	2.577	4.728	41.7%(116)	90 (20)	0.027	0.012	−0.121	0.325	0.045	−0.091	0.348	0.039
6	95	86	3.012	6.628	66.7%(56)	9	0.255	0.086	0.083	0.006	0.6	−0.14	0.027	0.425
7	49	48	1.708	6.104	21.9%(9)	10 (2)	0.009	0.006	−0.574	0.022	0.278	−0.383	0.113	0.242
8	42	32	3.063	4.313	35.7%(10)	8	0.261	0.102	−0.528	0.017	0.417	−0.322	0.111	0.283
9	72	67	3.075	4.373	22.2%(14)	15	0.116	0.048	−0.472	0.016	0.193	−0.307	0.035	0.151
10	16	16	3.375	5.438	50%(8)	8 (1)	0.006	0.002	1	0.167	0.238	−0.352	0.32	0.197
11	11	11	2.364	5.364	63.7%(7)	7 (3)	0.004	0.002	−0.806	0.119	0.333	−0.625	0.212	0.333
12	26	26	2.346	5.808	76.2%(16)	8 (1)	0.009	0.004	−1	0	0.3	−0.717	0.259	0.25
15	216	201	2.413	4.662	43.3%(69)	48 (14)	0.011	0.006	−0.297	0.02	0.121	−0.197	0.032	0.092
16	30	30	3	6.7	67.8%(19)	3	0.136	0.048	1	0	0.5	−0.272	0.105	0.341
17	65	58	3.017	6.638	63.7%(37)	6	0.314	0.105	−0.333	0.007	0.81	−0.486	0.031	0.582
18	49	44	3.205	4.205	4.9%(2)	10	0.169	0.07	−0.175	0.014	0.308	−0.16	0.014	0.276
19	23	23	2.826	4.696	54.6%(12)	5	0.01	0.005	−1	0	0.333	−0.336	0.087	0.303
20	170	158	2.259	4.456	37%(47)	34 (13)	0.012	0.007	−0.421	0.024	0.157	−0.293	0.038	0.106
21	17	17	2	7.059	85.7%(12)	5 (1)	0.011	0.006	−0.714	0.024	0.4	−0.714	0.024	0.4
22	31	28	3.5	4.286	52.6%(10)	10	0.014	0.006	−0.41	0.033	0.472	−0.374	0.062	0.352



(a) (b)

Fig. 2. (a) Network of dissertations according to CRIS ASU DB. Node's label is an identification number of a dissertation's author. Two dissertations are connected when they have the same articles. Arrow points to an earlier dissertation; (b) Network of co-authors publications related to the dissertations. Node's label is an employee's identification number. Colors of nodes indicates different groups.



(a) (b)

Fig. 3. Networks of co-authors (only supervisors and candidates) publications related to the dissertations. Node's label is an employee's identification number. Nodes representing supervisors are colored in orange. (a) Group 3. $r = 0.001, D = 0.175$; (b) Group 6. $r = 0.083, D = 0.6$.

4. Conclusion

In our research, information and communication technologies were used, which provide automation of processing and analyzing large amounts of information, including the search for and analysis of fuzzy duplicate publications.

Our study should answer the question whether there is objective, formalized, adequate and informative indicators to detect signs of violation of the publication ethics in scientific articles.

Quantitative criteria for the detection of risk groups will contribute to observance of the publication ethics principles and provide a higher quality of research training.

The obtained results will be useful to evaluate the activity of research groups and institutions that will improve the efficiency of management decision-making in matters of personnel placement, the use of material resources and funding.

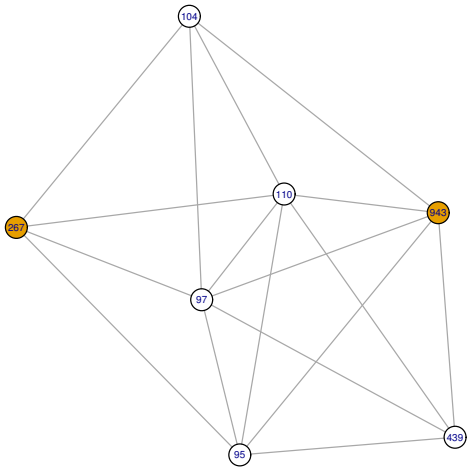


Fig. 4. Networks of co-authors (only supervisors and candidates) publications related to the dissertations. Node's label is an employee's identification number. Nodes representing supervisors are colored in orange. Group 17. $r = -0.333$, $D = 0.81$.

References

1. Y. Y. Tarasevich, Scientometric information and its interpritation, Informatizaciya obrazovaniya i nauki (2014) 141–148(in Russian).
2. M. E. J. Newman, J. Park, Why social networks are different from other types of networks, Phys. Rev. E 68 (2003) 036122. doi:10.1103/PhysRevE.68.036122.
3. R. Albert, A.-L. Barabási, Statistical mechanics of complex networks, Rev. Mod. Phys. 74 (2002) 47–97. doi:10.1103/RevModPhys.74.47.
4. M. E. J. Newman, The structure and function of complex networks, SIAM Review 45 (2) (2003) 167–256. doi:10.1137/S003614450342480.
5. P. M. Carron, R. Kenna, Universal properties of mythological networks, EPL 99 (2) (2012) 28002. URL http://stacks.iop.org/0295-5075/99/i=2/a=28002
6. D. R. Amancio, Comparing the topological properties of real and artificially generated scientific manuscripts, Scientometrics 105 (3) (2015) 1763–1779. doi:10.1007/s11192-015-1637-z.
7. M. E. J. Newman, Assortative mixing in networks, Phys. Rev. Lett. 89 (2002) 208701. doi:10.1103/PhysRevLett.89.208701.
8. C. King, Single-author papers: a waning share of output, but still providing the tools for progress, ScienceWatch (September 2013). URL http://sciencewatch.com/articles/single-author-papers-waning-share-output-still-providing-tools-progress
9. M. Girvan, M. E. J. Newman, Community structure in social and biological networks, PNAS 99 (12) (2002) 7821–7826. doi:10.1073/pnas.122653799.