Science and ethics meet: a mathematical view on one kind of violation of publication ethics

Taisiya S Shinyaeva, Yuri Yu Tarasevich
Astrakhan State University, 20A Tatishchev Street, Astrakhan, 414056, Russia
E-mail: tarasevich@asu.edu.ru

Abstract. When a person who did not make a significant intellectual contribution to a published research is included into the co-author list, the person is called gift or guest author depending on the reason why the person has been added to the co-authors. Essential deviation of properties of a particular co-author network from typical values may evidenced that the network is artificial. Using network analysis, we have performed an attempt to characterize a typical co-author network. We performed analysis of the co-author networks using references in the thesis on Physics and Mathematics, Economics defended from 2012 to 2017 and planned to be defended in 2017 and 2018 in Russia. Properties of the co-author networks are expected to be a reference sample in future research.

1. Introduction
During last decade, a problem of malpractice and misconduct in scholarly publications attracts more and more attention from the academic community [1, 2]. The problem has many manifestations, including, but not limited to, duplicate publications and plagiarism [1], predatory journals [3], and closely related to them authors, diploma, and degree mills [4, 5]. Incorrect list of co-authors (i.e., gift, guest, ghost authors) is one of the possible kinds of malpractice [6]. Gift and guest authors are persons who did not make a significant intellectual contribution to a published research but included into the co-author list because of different reasons. During the last couple of years, “author contribution statement” became a mandatory section of articles in many journals. This fact evidenced that improper co-author lists are often suspected. Unfortunately, explicit “author contribution statement” is hardly able to stop Mr. Fraud. Academic community badly needs a tool to detect and prevent malpractice. Scientific methods are successfully used for plagiarism detection and analysis of publications in predatory journals [7], nevertheless, possibilities to use scientific methods for “gift author” detection are not clear yet [8]. Application of network analysis [9, 10, 11, 12, 13, 14, 15] to co-author networks looks very promising as a potential tool for improper author lists detection. Social network analysis based techniques are used for anomaly detection in different types of networks [16, 17, 18]. Method of finding transitive triads was used to identify the degree mill — the dissertational councils involved in the production of plagiarism, which allow successful protection of substandard works [19].

In this conference paper, we present some new attempts applying network analysis to indicate improper co-author lists. Our research is based on an assumption that there may exist some stable groups or persons which permanently include so-called “gift authors” into co-author lists. If this assumption is correct then such the artificially enlarged co-author networks presumptively
have some anomalies in their properties. https://www.dissernet.org/ evidenced that malpractice and misconduct in Russia in Physics and Mathematics is minimal in compare with other sciences, humanities, and arts. The greatest number of violations falls on the economics, jurisprudence and pedagogical sciences. This fact allows considering co-author networks in physical science as a potential reference sample.

2. Materials
In our study, we investigated the theses defended from 2012 to 2017 and planned to be defended in 2017 and 2018 in Russia relating to the two branches of science: physics and mathematics, economics. Their open access full texts can be found at the official portal of the Higher Attestation Commission of the Ministry of Education and Science of Russian Federation http://vak.ed.gov.ru/dis-list. Information about 11976 doctoral and candidate dissertations on examined fields of sciences is presented on this site.

Each record of the thesis at the portal contains the following information

- thesis’ ID,
- the type of dissertation (candidate/PhD or doctoral/Habilitation),
- full name of the applicant,
- title of the dissertation,
- scientific specialty,
- branch of science,
- ID of the dissertational council,
- name of the organization in which the defense took place,
- thesis (link to file),
- date of defense of the thesis.

The records are uniform that allows automatic processing with very small amount of improper identification. Each record has to be accompanied with full text of the thesis as a file in pdf format. We were able to process only 7093 files out of 11976 records, because in some cases the files were improper encoded, absent or secured (Table 1).

| Subject area                  | N     | Improper encoded files | Absent files | Secured files |
|-------------------------------|-------|------------------------|--------------|---------------|
| Physics and mathematics       | 6199  | 1500 (24.2%)           | 823 (13.3%)  | 105 (1.7%)    |
| Economics                     | 5777  | 232 (4%)               | 2200 (38.1%) | 23 (0.4%)     |

An analysis of obtained files was carried out to obtain the necessary additional information

- the organization in which the thesis was fulfilled,
- information about supervisors and opponents:
  - full name,
  - scientific degree and title,
  - affiliation,
- list of journal articles published on the topic of the dissertation.
Automatic extraction of metadata from these files are obstructed by lack of uniformity in presentation of information about supervisors and opponents, as well as in the preparation of lists of publications on the topic of the dissertation. Thus, we were extracted necessary information from slightly less than 90% of all files. Information about found defense participants (persons who were involved in the defense of the thesis: the applicant, the supervisor or the opponent), main publications and authors is presented in Table 2.

Table 2: Result of files processing. $N$ — number of files, $P$ — number of entries about defenses participants, $Auth$ — total number of entries about publications authors, $Pub/D$ — average number of publications per thesis, $Auth/Pub$ — average number of authors per publication

| Subject area      | $N$  | $P$   | $Auth$ | $Pub/D$ | $Auth/Pub$ |
|-------------------|------|-------|--------|---------|------------|
| Physics and math. | 3771 | 13600 | 88338  | 8.3     | 3.47       |
| Economics         | 3322 | 11730 | 38362  | 9.4     | 1.45       |

It was necessary to determine the specific person for each such record because one person can participate in several defenses and to act both as an applicant and as an opponent or supervisor. Due to the fact that 98.4% of all participant records contain full names, and 89.7% of records contain information on the degree, rank or position of the defense participant, only 0.4% of the participants could not be identified (linked to the specific person). Searching for people in the list of authors of publications is a non-trivial task. The complexity of identification is primarily due to incomplete names (names and initials are indicated), with various versions of transliteration, misprints. Automatical linking assumes errors. Therefore, when there were ambiguous situations, no binding was performed. Thus, about 3% of the authors could not be identified.

The list of persons who took part in defenses with incorrect borrowings is available on the “Disseropedy” website (http://rosvuz.dissernet.org/person) and include 2647 records. 434 matches with people from our database were found. All coincidences refer to the economics.

3. Methods
We built network of the defenses participants and network of co-authors of the main publications on the topic of the dissertation. Depth-first search algorithm [20] has been applied to the networks. We built egonets (the 1-step neighborhood of a node) and studied the following characteristics of obtained ego networks.

- $N$ — number of nodes,
- $E$ — number of edges,
- $AvD$ — average degree,
- $d$ — diameter,
- $AvPath$ — average path length,
- $D$ — density,
- $C$ — average clustering coefficient [21],
- $T$ — total triangles,
- $NT$ — number of triangles related to node,
- $r$ — degree assortativity of graph [22]. Assortativity measures the similarity of connections in the graph with respect to the node degree.
4. Results

4.1. Network of defense participants

Network of defense participants contains 15434 nodes and 29853 edges, including 1299 weakly connected components. Two giant connected components contains 10231 nodes and 23067 edges (figure 1), their characteristics are presented in table 3.

Table 3: \( N \) — number of nodes, \( E \) — number of edges, \( AvD \) — average degree, \( d \) — diameter, \( AvPath \) — average path length, \( D \) — density, \( C \) — average clustering coefficient, \( T \) — total triangles, \( r \) — degree assortativity of graph.

| Sciences (Subject) area          | \( N \) | \( E \) | \( AvD \) | \( d \) | \( AvPath \) | \( D \) | \( C \) | \( T \) | \( r \) |
|---------------------------------|--------|--------|----------|--------|-------------|--------|-------|-------|-------|
| Physics and mathematics         | 3710   | 7734   | 4.169    | 38     | 12.033      | 0.001  | 0.804 | 5571  | -0.05 |
| Economics                       | 6521   | 15333  | 4.703    | 21     | 7.314       | 0.001  | 0.808 | 11091 | -0.03 |

Figure 1: Two giant connected components of defense participants network. On the left side, there is component related to physics and mathematics (green color), on the right side — to economics (blue color). The participants of the defenses included in the “Disseropedy” (http://rosvuz.dissernet.org/person) are marked in red.

Figure 2: Some ego networks in defense participants network. Green color indicates participants related to the physics and mathematics, blue — to the economics, red color indicates participants of the defenses included in the “Disseropedy”.

4.2. Co-author networks

Co-author network contains 31195 nodes and 109333 edges, including 2341 weakly connected components. 58.3% of nodes form giant connected component of co-author network (figure 3). Some ego networks are shown in the figure 4, their characteristics are presented in the table 5.
Table 4: Characteristics of ego networks. $N_T$ — number of triangles related to node, $N$ — number of nodes, $E$ — number of edges, $AvD$ — average degree, $D$ — density, $C$ — average clustering coefficient, $T$ — total triangles, $r$ — degree assortativity of graph.

| Node Label | $N_T$ | $N$ | $E$ | $AvD$ | $D$   | $C$   | $T$   | $r$    |
|------------|-------|-----|-----|-------|-------|-------|-------|--------|
| 30737      | 70    | 54  | 123 | 4.556 | 0.086 | 0.875 | 97    | -0.297 |
| 31293      | 35    | 26  | 60  | 4.615 | 0.185 | 0.877 | 50    | -0.324 |
| 30641      | 15    | 16  | 30  | 3.75  | 0.25  | 0.946 | 20    | -0.333 |
| 530        | 41    | 37  | 78  | 4.216 | 0.117 | 0.932 | 58    | -0.313 |
| 3164       | 31    | 20  | 50  | 5     | 0.263 | 0.861 | 49    | -0.296 |
| 3169       | 15    | 17  | 31  | 3.647 | 0.228 | 0.945 | 20    | -0.364 |

Figure 3: The giant connected component of co-author network contains 18173 nodes and 88247 edges.

Table 5: Characteristics of ego networks. $N_T$ — number of triangles related to node, $N$ — number of nodes, $E$ — number of edges, $AvD$ — average degree, $D$ — density, $C$ — average clustering coefficient, $T$ — total triangles, $r$ — degree assortativity of graph.

| Node Label | $N_T$ | $N$ | $E$ | $AvD$ | $D$   | $C$   | $T$   | $r$    |
|------------|-------|-----|-----|-------|-------|-------|-------|--------|
| 36176      | 114   | 20  | 133 | 13.3  | 0.7   | 0.921 | 521   | 0.005  |
| 31191      | 32    | 18  | 49  | 5.444 | 0.32  | 0.861 | 62    | -0.307 |
| 33410      | 15    | 7   | 21  | 6     | 1     | 1     | 35    | 0      |
| 1187       | 114   | 31  | 144 | 9.29  | 0.31  | 0.847 | 316   | -0.198 |
| 252        | 32    | 13  | 44  | 6.769 | 0.564 | 0.854 | 69    | -0.396 |
| 58         | 15    | 8   | 22  | 5.5   | 0.786 | 0.893 | 9     | -0.369 |

5. Conclusion
In our research, we have analysed some open access information presented at the official portal of the Higher Attestation Commission of the Ministry of Education and Science of Russian Federation http://vak.ed.gov.ru/dis-list. The information allows building co-author networks.
and networks as applicant–supervisor–opponent. Average clustering coefficient, graph density, and average degree have been calculated for particular clusters. We suppose that such the quantities may be used for identification of suspicious co-author networks which presumably are artificially built and may contain gift and/or guest authors.

Research is in progress.

Acknowledgments
The reported study was funded by RFBR according to the research project No. 6-33-00006.

Author contribution statement: YYT proposed the principal idea of the research, TSS performed data processing, network and statistical analysis, both co-authors took a part in discussion and interpretation of the results as well in the manuscript preparation.

References
[1] Fang F C, Steen R G and Casadevall A 2012 PNAS 109 17028–17033 ISSN 1091-6490
[2] Gasparyan A Y, Yessirkepov M, Diyanova S N and Kitas G D 2015 J. Korean Med. Sci. 30 1010 ISSN 1011-8934
[3] Beall J 2012 Nature 489 179–179 ISSN 0028-0836
[4] Sorokowski P, Kulczycki E, Sorokowska A and Pisanski K 2017 Nature 543 481–483 ISSN 0028-0836
[5] Angulo A 2016 Diploma mills: How for-profit colleges stiff students, taxpayers, and the American dream (Baltimore: Johns Hopkins University Press) ISBN 978-142142008-0
[6] Wislar J S, Flanagan A, Fontanarosa P B and DeAngelis C D 2011 BMJ 343 ISSN 0959-8138
[7] Xia J, Harmon J L, Connolly K G, Donnelly R M, Anderson M R and Howard H A 2015 J. Assoc. Inf. Sci. Tech. 66 1406–1417 ISSN 2330-1643
[8] Shinyaeva T S and Tarasevich Y Y 2017 Procedia Comput. Sci. 106 3–10 ISSN 1877-0509
[9] Yose J, Kenna R, MacCarron P, Platini T and Tonra J 2016 Adv. Complex Syst. 19 1650008 ISSN 0219-5259
[10] Kenna R and Mac Carron P 2016 Phys. World 29 22 ISSN 2058-7058
[11] Kenna R and Mac Carron P 2016 J. Phys.: Conf. Ser. 681 012002 ISSN 1742-6596
[12] Sarkanych P, Holovatch Y, Kenna R and Mac Carron P 2016 J. Phys. Studies 20 4801 ISSN 2310-0052
[13] Mac Carron P and Kenna R 2013 Significance 10 12–17 ISSN 1740-9713
[14] Mac Carron P and Kenna R 2013 Eur. Phys. J. B 86 407 ISSN 1434-6036
[15] Mac Carron P and Kenna R 2012 EPL (Europhysics Letters) 99 28002 ISSN 1286-4854
[16] Aloghi L, Tong H and Koutra D 2015 Data Mining and Knowledge Discovery 29 626–688 ISSN 1573-756X
[17] Vengertsev D and Thakkar H 2015 Anomaly detection in graph: Unsupervised learning, graph-based features and deep architecture Tech. rep. Department of Computer Science, Stanford University
[18] Kaur R and Singh S 2016 Egyptian Informatics Journal 17 199 – 216 ISSN 1110-8665
[19] Safonova M A and Sokolov M M 2016 Partija Vernadskogo. VVIPartija In russian URL http://vviparti.esrae.ru/2-23
[20] Tarjan R 1972 SIAM J. Comput. 1 146–160 ISSN 0097-5397
[21] Watts D J and Strogatz S H 1998 Nature 393 440–442 ISSN 0028-0836
[22] Newman M E J 2003 Phys. Rev. E 67(2) 026126 ISSN 2470-0053 (online), 2470-0045 (print)