Exploratory analysis of communities in co-authorship networks: A case study

Miloš Savić[1], Mirjana Ivanović[1], Miloš Radovanović[1], Zoran Ognjanović[2], Aleksandar Pejović[2], and Tatjana Jakšić Krüger[2]

[1] Department of Mathematics and Informatics, Faculty of Sciences
University of Novi Sad, Trg Dositeja Obradovića 4, 21000 Novi Sad, Serbia
{svc,mira,radacha}@dmi.uns.ac.rs
[2] Mathematical Institute of the Serbian Academy of Sciences and Arts
Kneza Mihaila 36, 11001 Beograd, Serbia
{zorano,pejovica,tatjana}@mi.sanu.ac.rs

Abstract. Digital libraries enable worldwide access to scientific results, but also provide a valuable source of information that can be used to investigate patterns and trends in scientific collaboration. The Electronic Library of the Mathematical Institute of the Serbian Academy of Sciences and Arts (eLib) digitizes the most prominent mathematical journals printed in Serbia. Using eLib bibliographical records we constructed a co-authorship network representing collaborations between authors who published their papers in eLib journals in the period from 1932 to 2011. In this paper we apply community detection techniques in order to examine the structure of the eLib co-authorship network. Such study reveals characteristic patterns of scientific collaboration in Serbian mathematical journals, and helps us to understand the (self-)organization of the eLib community of authors.

Keywords: digital library, co-authorship network, Serbian mathematical journals, scientific collaboration, social network analysis, community detection

1 Introduction

It has long been realized that the analysis of co-authorship graphs can help us to understand the structure and evolution of corresponding academic societies. Those networks can also be used to develop models for ranking most influential authors in a database [8], to automatically determine the most appropriate reviewers for a manuscript [21], or even to predict future research collaborations [12]. Nodes in a co-authorship network represent researchers – people who published at least one research paper. Two researchers are connected by an undirected link if they authored at least one paper together, with or without other coauthors. Additionally, link weights can be introduced in order to express the strength of collaboration: two researchers are connected by a link of weight w if they co-authored exactly w different research papers.

Community structure is a typical feature of social networks [16, 4]. A community (cluster or module) is a part of a network (group of nodes) where internal connections are denser than external ones. Uncovering communities helps us to understand the structure of the network, to identify cohesive subgroups, and to draw a readable map of the network.

This study explores structural properties of the co-authorship network that is formed from bibliographical records contained in the Electronic Library of the Mathematical Institute of the Serbian Academy of Sciences and Arts – eLib [13]. ELib started as a response to the increasing requirement for easier access to old issues of the journal *Publications de l'Institut Mathématique*. Currently, eLib digitizes 12 mathematical journals printed in Serbia. Therefore, the nature of the bibliographic data enables us to investigate the structure of scientific collaboration characteristic to authors who publish their results in Serbian mathematical journals.

The rest of the paper is structured as the folows. Related work is presented in Section 2. Section 3 describes the methodology that is used to examine structural properties of the network and identify cohesive subroups of co-authors. The obtained results are presented and discussed in Section 4. Finally, the last section concludes the paper.

2 Related work

A more recent resurgence of interest in networks of scientists and scientific papers was sparked by the observation of power-law degree distributions in various types of real-world networks [1] including networks of scientific collaboration [14, 15]. It is also observed that the largest connected component in collaboration networks tends to take up the majority of the network [14]. Collaboration networks also exhibit expected short paths between arbitrary researchers [15], i.e. they tend to be "small worlds."

The body of work most relevant to our study involves collaboration networks in the field of mathematics. Studies of collaboration networks focused around Paul Erdős include [9] and [2]. More general analysis of mathematics collaboration networks is performed by Grossmann [10, 11] who examined statistical properties of the network derived from Mathematical Reviews (MR). Brunson et al. [5] studied the evolution of the MR network, identifying two points of drastic reorganization of the network, as well as increased collaboration between mathematics researchers in more recent times.

Communities in co-authorship graphs may indicate groups of people with common research interest. For example, Girvan and Newman [7] used community detection techniques to identify groups corresponding to different research divisions at the Santa Fe institute. In our previous work [23] we studied statistical properties and evolution of the eLib co-authorship graph. The same article presented the methodology that is used to extract the network. This paper continues the work presented in [23]. Namely, in this paper we investigate the structure of the network using community detection methods.

3 Exploratory analysis

The analysis of structure of scientific collaboration in eLib journals is based on standard methods and metrics used in analysis of social networks. Firstly, we performed connected component analysis in order to isolate disjoint components of the network and to determine whether the network contains so called *giant connected component*. A connected component of an undirected network is a set of mutually reachable nodes, i.e. there is a path connecting each two nodes in the component. Giant connected component is a component that encompasses the vast majority of nodes. Secondly, we distinguish between two types of components in a co-authorship network: non-trivial and trivial components. A component of a co-authorship network is considered trivial if it is a complete sub-graph of the network and the weight of each link is equal to one. In other words, trivial components represent research collaborations that have not evolved in the examined time period.

We use different metrics to quantify nodes (authors) in the eLib co-authorship network. *Degree centrality* (DC) of author A is the number of links incident to A, i.e. the number of other authors with whom A collaborated. *Betweenness centrality* (BC) of A is the number of shortest paths between any pairs of nodes that pass through A. Unlike DC which is a local centrality measure, BC quantifies the centrality of a node considering the whole network. Nodes with high BC tend to be the most important actors in the network since they connect different groups of nodes and may control the flow of information in the network. To measure author productivity we use the normal counting method, i.e. the productivity of A is equal to the number of publications A (co-)authored. *Timespan* of author A is the number of years that passed from the publication of A's first article to the publication of A's last article in eLib journals.

An important advance in community detection was made by Girvan and Newman [17] who introduced a measure called *modularity* to estimate the quality of a partition of a network into communities. For weighted networks modularity Q is defined as

$$Q = \sum_{c=1}^{n_c} \left[\frac{W_c}{W} - \left(\frac{S_c}{2W} \right)^2 \right],$$

where n_c is the number of communities in the partition, W_c is the sum of weights of intra-community links of community c, S_c is the total weight of links incident to nodes in c, and W is the total weight of links in the network. In other words, modularity accumulates the difference between the total weight of links within a cluster and the expected total weight in an equivalent network with links placed at random. In this paper we use the Louvain method for community detection [3] to identify cohesive subgroups in the eLib co-authorship graph. Initially, we investigated the performance of five different community detection techniques on the largest connected component and showed that the Louvain method is the most suitable for our case study. The method uses a greedy multi-resolution approach to maximize Q starting from the partition where all nodes are put in different communities. When Q is optimized locally the algorithm builds the

coarse-grained description of the network (network of communities), and then repeats the same procedure until a maximum of modularity is attained. Although widely used, the modularity measure has a weakness known as the resolution limit problem – community detection techniques based on modularity maximization may fail to to identify modules smaller than a scale which depends on the total size of the network. Therefore, the application of modularity maximization methods requires investigation of the quality of obtained community partitions. In order to assess the reliability of the community detection method we use the definition of community proposed by Radicchi et al. [19] adopted for weighted networks. Namely, a community is called *Radicchi strong* if for each node in the community the sum of weights of links within the community (strength of intra-community links) is higher than the sum of weights of links connecting the node with the rest of graph (strength of inter-community links).

4 Results and discussion

In total 6480 research papers were published in eLib journals from 1932 to 2011. The majority of articles are single-authored papers: 4836 papers (74.63% of the total number of papers) are written by exactly one author. This situation is not surprising for mathematical journals, since researchers in mathematics and humanities usually engage in solitary work, while laboratory scientists tend to write articles with many co-authors.

The total number of authors that published papers in eLib journals during the examined period is 3597. Therefore, the co-authorship network formed from eLib bibliographic records contains 3597 nodes (authors). Those authors are connected by a significantly smaller number of links (2766) which means that there is a large number of authors (33% of the total number of authors) who have not collaborated with other eLib authors by publishing articles in eLib journals.

4.1 Connected components

Connected component analysis revealed that the eLib co-authorship network is extremely fragmented: it contains 625 connected components (excluding isolated nodes) neither of which is a giant connected component. Additionally, the network contains nearly the same number of trivial and non-trivial components: 319 components are trivial (51.04%), while 306 of them are non-trivial. The average size of non-trivial components is 6.42, while the standard deviation is 17.83. This means that the eLib co-authorship graph contains components whose size is drastically larger than the average. In total, 19 components have size that is greater or equal to ten, while six of them have size greater than 20 authors. The largest connected component encompasses 249 authors, which is 6% of the total number of authors. The number of papers published by authors from the largest component is 997, which is 15.38% of the total number of papers, and the maximal number of papers per component.

4.2 Community structure of largest connected components

In order to select the best community detection method for our case study we initially investigated performance of five different community detection methods on the largest connected component. Results are presented in Table 1. It can be observed that the Louvain method shows the best performance for our network: this method reveals a community partition having the highest modularity and the largest percentage of Radicchi strong communities.

Table 1. Comparative analysis of performance of different community detection methods applied to the largest connected component: C – the number of detected communities, Q – modularity score, Strong – the percentage of Radicchi strong communities.

Method	C	Q	Strong [%]	Reference
Girvan-Newman edge betweenness	11	0.813	72.7	[7]
Walktrap	23	0.824	82.6	[18]
Infomap	30	0.802	66.7	[22]
Label propagation	29	0.803	79.3	[20]
Louvain	16	0.834	93.7	[3]

Since the Louvain method shows the best performance on the largest connected component we selected this method to investigate the community structure of ten largest connected components in the network. Results are summarized in Table 2. It can be observed that for each component the value of the modularity measure Q is higher than 0.3. Usually a value of Q larger than 0.3 is considered as a clear indication that the network possesses community organization according to the modularity based definition of community [6]. Moreover, the modularity score of the five largest eLib components is even higher than 0.5, and the largest component has the largest value of modularity.

Table 2. Results of community detection for ten largest connected components in the eLib co-authorship graph: N – the number of nodes in the component, Q – modularity score, C – the number of detected communities.

N	Q	C	N	Q	C
249	0.834	16	21	0.503	4
74	0.716	8	19	0.486	3
37	0.507	4	19	0.500	4
27	0.531	5	18	0.435	5
25	0.583	4	17	0.334	3

To investigate the quality of obtained community partitions we examine in detail the communities detected in the three largest connected components. Figure 1 shows the visualization of the largest connected component after community detection, while Table 3 provides a description of the obtained communities.

The largest cohesive subgroup is organized around Ivan Gutman who is the best connected eLib author and the most productive author. The central figure in the second largest community is Žarko Mijajlović who is the most central author according to the betweenness centrality metric. The third largest community which is organized around Jovan Karamata (1902–1967) encompasses the oldest generation of authors present in eLib journals, also including Paul Erdős. From this community the whole component started to emerge: the first collaboration among eLib authors is the collaboration between Jovan Karamata and Hermann Wendelin which was established in 1934. It can be observed that for each detected community the number of intra-community links (denoted by "IntraL" in Table 3) is significantly higher than the number of inter-community links (denoted by "InterL"). The same holds also for the sum of weights of intra-community ("IntraW") and inter-community ("InterW") links which means that the overall strength of collaboration among members of each community is higher than the strength of collaboration among authors belonging to different communities. Moreover, each of the detected communities, except community C6, is Radicchi strong which means that each author from a community collaborates more often with authors from his/her community than with authors from other communities. In case of community C6 there are only two authors who are not Radicchi strong: (1) Slobodan Simić has 9 joint publications with members of his community and 10 joint publications with members of communities C1 and C5, and (2) Vlajko Kocić has 1 joint publication with Slobodan Simić and 3 joint publication with Jovan Kečkić who belongs to community C5. For the majority of detected communities (all of them except for C3, C5 and C6) the author having the highest degree centrality in the community (shown in Table 3) is at the same time the author who is most central according to the betweenness centrality metric.

Figure 2 shows the structure of the second largest connected component after community detection. The characteristics of the partition are given in Table 4. It can be observed that for each detected community the number of intra-community links is significantly higher than the number of inter-community links. The same also holds for the sum of weights of this two types of links. Moreover, each detected community is Radicchi strong which clearly suggests that the applied community detection technique produced a good partition into communities. The authors having the highest degree centrality in communities denoted by C1, C4, C5, C6 and C8 are Serbian mathematicians affiliated with the University of Novi Sad. Community C5 is organized around Bogoljub Stanković, a Serbian Academician from Novi Sad, who is the author with the maximal value of timespan for the whole network in the examined time period: the first paper of Bogoljub Stanković published in eLib journals is from 1953, while the last one is from 2011. For 6 out of 8 communities (all except C2 and C7) the author having the highest degree in the component is also the author with the highest betweenness centrality. The authors having the maximal betweenness centrality in C2 and C7 are Miroslava Petrović-Torgašev and Ratko Tošić, respectively.

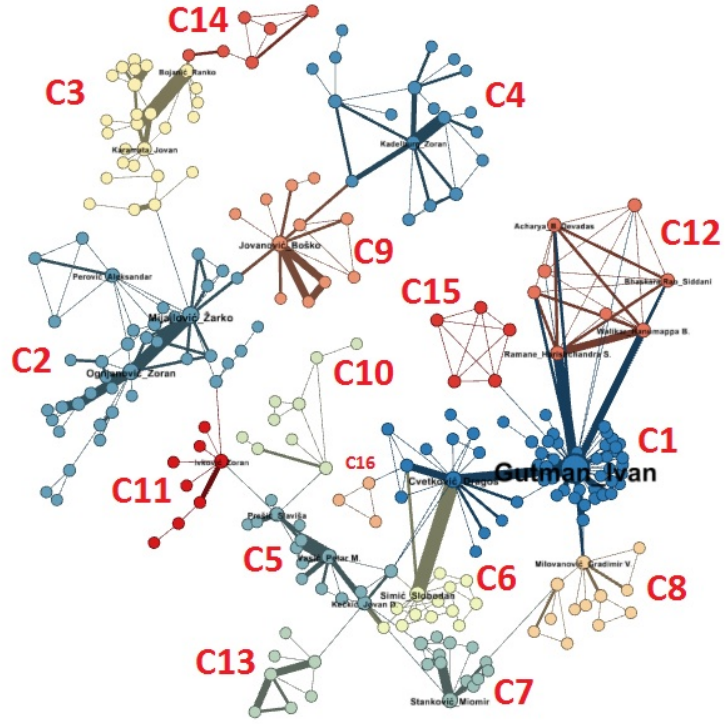

Fig. 1. Visualization of the largest connected component in the eLib co-authorship graph. Nodes from the same community are in the same color. Additionally, each community is marked with an appropriate identifier (C1, C2, etc.) used in Table 3.

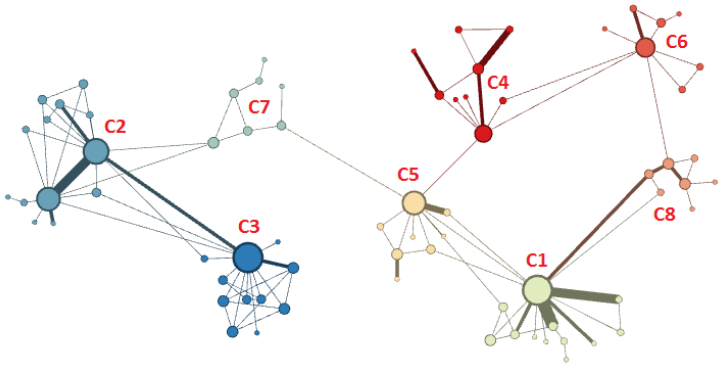

Fig. 2. Visualization of the second largest connected component in the eLib co-authorship graph after community detection.

Table 3. Description of detected communities for the largest connected eLib component.

Community	Size	Max. degree author	IntraL	InterL	IntraW	InterW	Strong
C1	54	Ivan Gutman (50)	82	15	108	33	yes
C2	40	Žarko Mijajlović (16)	66	4	106	6	yes
C3	26	Jovan Karamata (8)	35	2	64	3	yes
C4	19	Zoran Kadelburg (7)	25	1	42	2	yes
C5	15	Petar M. Vasić (10)	23	8	42	10	yes
C6	13	Slobodan Simić (12)	20	4	20	13	no
C7	13	Miomir Stanković (11)	23	3	34	3	yes
C8	12	Gradimir Milovanović (8)	15	3	18	4	yes
C9	11	Boško Jovanović (12)	14	3	26	6	yes
C10	9	Jovan Petrić (5)	10	1	11	1	yes
C11	8	Zoran Ivković (8)	7	2	9	2	yes
C12	8	Ramane Harishchandra (8)	21	8	31	18	yes
C13	7	Svetozar Milić (5)	8	1	16	1	yes
C14	6	Snežana Pejović (4)	8	1	10	2	yes
C15	5	Song Zhang (5)	10	1	10	1	yes
C16	3	Bolian Liu (3)	3	1	3	1	yes

Table 4. Description of detected communities for the second largest connected eLib component.

Community	Size	Max. degree author	IntraL	InterL	IntraW	InterW	Strong
C1	14	Stevan Pilipović (13)	18	5	28	6	yes
C2	13	Leopold Verstraelen (11)	19	6	25	7	yes
C3	11	Ryszard Deszcz (13)	19	4	20	5	yes
C4	9	Dragoslav Herceg (7)	10	3	14	3	yes
C5	8	Bogoljub Stanković (10)	9	6	12	6	yes
C6	7	Djurdjica Takači (8)	8	3	9	3	yes
C7	7	Mirjana Djorić (4)	7	3	7	3	yes
C8	5	Arpad Takači (5)	5	2	6	3	yes

The third largest connected component in the eLib co-authorship network encompasses eLib authors who published their papers in two eLib journals: "Computer Science and Information Systems" and "Review of the National Center for Digization". The scope of mentioned journals is not purely mathematical, but oriented to applications of mathematics and computer science, where the number of authors per paper is generally higher compared to pure mathematical research. Consequently, this component is denser than the previously two described connected components. The details of obtained communities for the third largest component are provided in Table 5. It can be observed that all detected communities are Radicchi strong. Additionally, for each component the author having the highest degree centrality has the highest betweenness centrality.

Table 5. Description of detected communities for the third largest connected eLib component.

Community	Size	Max. degree author	IntraL	InterL	IntraW	InterW	Strong
C1	12	Pedro Henriques (13)	25	16	49	19	yes
C2	11	Ivan Luković (10)	18	4	21	4	yes
C3	9	Marjan Mernik (17)	23	17	33	20	yes
C4	5	Bryant R. Barrett (5)	10	5	10	5	yes

5 Concluding remarks

The project of the electronic library of the Mathematical Institute of the Serbian Academy of Sciences and Arts (eLib) was founded in order to provide online presence and long-term preservation of mathematical journals printed in Serbia. In this study we used eLib bibliographical records to construct the co-authorship network of eLib authors and to identify cohesive subgroups in the network.

Analysis of connected components of the network revealed that the network contains a large number of components. The majority of them are isolated authors or small trivial components, but there is also a small number of relatively large, non-trivial components of connected authors. The main contribution of this article is that we showed that the largest connected components of the eLib co-authorship graph possess clear community structure. This means that authors belonging to the largest components are organized into non-overlapping cohesive subgroups. Additionally, we showed that the majority of identified groups tend to be strong in the sense that each author from a group collaborates more often with authors from his/her group than with authors from other groups.

Acknowledgments. Miloš Savić, Mirjana Ivanović and Miloš Radovanović gratefully acknowledge the support of this work by the Serbian Ministry of Education, Science and Technological Development through project no. OI174023. Zoran Ognjanović, Aleksandar Pejović and Tatjana Jakšić Kruger gratefully acknowledge the support of this work by the Serbian Ministry of Education, Science and Technological Development through project no. III44006.

References

1. Barabasi, A.L., Albert, R.: Emergence of scaling in random networks. Science 286(5439), 509–512 (1999)
2. Batagelj, V., Mrvar, A.: Some analyses of Erdős collaboration graph. Social Networks 22(2), 173–186 (2000)
3. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment 2008(10), P10008 (2008)
4. Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., Hwang, D.: Complex networks: Structure and dynamics. Physics Reports 424, 175–308 (2006)

5. Brunson, J.C., Fassino, S., McInnes, A., Narayan, M., Richardson, B., Franck, C., Ion, P., Laubenbacher, R.: Evolutionary events in a mathematical sciences research collaboration network. ArXiv e-prints (2012)
6. Fortunato, S., Barthlemy, M.: Resolution limit in community detection. Proceedings of the National Academy of Sciences 104(1), 36–41 (2007)
7. Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. Proceedings of the National Academy of Sciences 99(12), 7821–7826 (2002)
8. Gollapalli, S.D., Mitra, P., Giles, C.L.: Ranking authors in digital libraries. In: Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries. pp. 251–254. JCDL '11, ACM, New York, NY, USA (2011)
9. Grossman, J.W., Ion, P.D.F.: On a portion of the well known collaboration graph. Congressus Numerantium 108, 129–131 (1995)
10. Grossman, J.: The evolution of the mathematical research collaboration graph. Congressus Numerantium 158, 201–212 (2002)
11. Grossman, J.: Patterns of collaboration in mathematical research. SIAM News 35(9), 8–9 (2002)
12. Liben-Nowell, D., Kleinberg, J.: The link prediction problem for social networks. In: Proceedings of the Twelfth International Conference on Information and Knowledge Management. pp. 556–559. CIKM '03, ACM, New York, NY, USA (2003)
13. Mijajlović, Z., Ognjanović, Z., Pejović, A.: Digitization of mathematical editions in Serbia. Mathematics in Computer Science 3(3), 251–263 (2010)
14. Newman, M.E.J.: Scientific collaboration networks I: Network construction and fundamental results. Physical Review E 64, 016131 (2001)
15. Newman, M.E.J.: Scientific collaboration networks II: Shortest paths, weighted networks, and centrality. Physical Review E 64, 016132 (2001)
16. Newman, M.E.J.: The structure and function of complex networks. SIAM Review 45, 167–256 (2003)
17. Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. Physical Review E 69(2), 026113 (2004)
18. Pons, P., Latapy, M.: Computing communities in large networks using random walks. Journal of Graph Algorithms and Applications 10(2), 191–218 (2006)
19. Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., Parisi, D.: Defining and identifying communities in networks. In: Proceedings of the National Academy of Sciences of the United States of America. vol. 101, pp. 2658–2663 (2004)
20. Raghavan, U.N., Albert, R., Kumara, S.: Near linear time algorithm to detect community structures in large-scale networks. Physical Review E 76, 036106 (2007)
21. Rodriguez, M.A., Bollen, J.: An algorithm to determine peer-reviewers. In: Proceedings of the 17th ACM Conference on Information and Knowledge Management. pp. 319–328. CIKM '08, ACM, New York, NY, USA (2008)
22. Rosvall, M., Bergstrom, C.T.: Maps of information flow reveal community structure in complex networks. In: Proceedings of the National Academy of Sciences of the United States of America. vol. 105, pp. 1118–1123 (2007)
23. Savić, M., Ivanović, M., Radovanović, M., Ognjanović, Z., Pejović, A., Jakšić Krüger, T.: The structure and evolution of scientific collaboration in Serbian mathematical journals. Scientometrics, to appear (2014)