Collaboration network analysis of scientific production at UB-SEE

Jovan Stojković, Marko Mišić, *IEEE Member*, Jelica Protić

***Abstract* — This paper presents an analysis of the coauthors collaboration network based on research papers published in journals, where at least one author is affiliated to the University of Belgrade, School of Electrical Engineering (UB-SEE). We have extracted data from the institutional database and processed 2000 papers published in journals during the period from 2000 to 2017. After the initial refinement of the dataset we created and visualized the network of coauthors and carried out further analysis based on metrics specifically used for social network analysis.**

***Keywords* — collaboration networks, visualization, scientific papers, social network analysis**

I. Introduction

SCIENTIFIC production is an important indicator of the development of scientific society. It can affect ranking of the institutions, ranking of the individuals within those institutions, but also evaluation and approval of projects and accreditation of higher education programs. Scientific production is often analyzed through bibliometric and scientometric indicators [1]. In such sense, co-authorship patterns are also studied [2].

At the University of Belgrade, School of Electrical Engineering, (UB-SEE), the school itself is organized in several departments. UB-SEE is mainly educational institution, but beside the obligation to teach students, the other important activity is research. Research results are usually published in scientific conferences and journals. Usually, researchers work in groups that publish papers together. They are called coauthors. The coauthors can work in the same institution or they can be affiliated to different institutions. Therefore, the papers published in scientific conferences and journals can have coauthors that are employed at UB-SEE together with the ones that are not employed at this school.

Our previous work has been done on collaboration between sole UB-SEE faculty members, and particularly on collaboration between authors from the same department [3]. The similar research is conducted in the case of Faculty of Sciences, University of Novi Sad, Serbia [4]. However, cooperation between the researchers from UB-SEE and the ones from the other institutions can show the direction of the future research and from such data we can find the institutions having the most similar research interest to that of UB-SEE.

In this paper we present the analysis of the collaboration network of coauthors of research papers published in journals where at least one author is affiliated to UB-SEE. We obtained the raw data from the school's database and processed 2000 papers published in journals from 2000 to 2017. The network of coauthors is created using MS Excel and Python scripts, while the visualization and further analysis is performed using Gephi software tool [5]. The network is analyzed based on the metrics used in social network analysis, since collaboration network is a typical example of such social networks. Also, we ranked the authors based on different network metrices.

The paper is divided into five sections. The second section explains the process of the data collection and cleansing. The third section presents network modelling and visualization. The results of our analysis are given in fifth section, while the last section gives short conclusion and suggestions for the future work.

II. Methods

In order to create the network of coauthors we first had to collect data and perform the initial refinement. We obtained raw data from the UB-SEE's database. In that database, all papers published in journals are inserted by one of the authors employed at UB-SEE. We extracted papers published from 2000 to 2017 based on the year of publishing. From all the fields in the original database, for our research we only needed the names of the authors, the names under which the paper was published and the identification of journal. A row with the necessary data for our research from the original database can be seen in Figure 1.

Authors' names	Publishing names	Journal ID
Vukašin Milovanović,Драган Милићев	V. Milovanović,D. Milićev	11527

Fig. 1. An example of the input database's row

While collecting data for further processing, there were couple of challenges which resulted from the freedom given to the authors when inserting new record in the institutional database. When inserting a new paper to the database, the authors fill in necessary paper data, such as: authors of the paper, title, journal or conference name, year of publishing, etc. To avoid duplicated entries for the same paper for authors that are employed together at UB-

This work has been partially funded by the Ministry of Education and Science of the Republic of Serbia (III44009 and TR32047). The authors gratefully acknowledge the financial support.

Jovan Stojkovic is with the School of Electrical Engineering, University of Belgrade, Bul. kralja Aleksandra 73, 11120 Belgrade, Serbia (e-mail: sj160151d@student.etf.rs).
Marko Misic is with the School of Electrical Engineering, University of Belgrade, Bul. kralja Aleksandra 73, 11120 Belgrade, Serbia (e-mail: marko.misic@etf.bg.ac.rs).
Jelica Protic is with the School of Electrical Engineering, University of Belgrade, Bul. kralja Aleksandra 73, 11120 Belgrade, Serbia (e-mail: jelica.protic@etf.bg.ac.rs).

978-1-7281-4790-1/19/$31.00 ©2019 IEEE

SEE, those author's name can be chosen from a drop-down list. The paper is then showed in each author's profile in institutional database. However, for those author's that are not employed at UB-SEE, only forename and surname are required. Also, the authors can use alphabet of choice to store their references. For Serbian researchers English, Serbian Cyrillic and Serbian Latin alphabet can be used, while for the foreign researchers usually only English alphabet is used.

All aforementioned facts can produce high variance and non-uniform records which require some form of data cleansing. Similar problems were described in [6]. To perform cleansing of the data, we wrote several Python scripts. One script is used to translate Serbian letters (both Cyrillic and Latin) into English letters. However, there were still unsolved problems with the data refinement. The authors do not have to use any particular form for the names when filling in new record for the database. That means that the names can be in the any of the possible forms: full first name and full surname, initial of the first name and full surname, full first name and initial of the surname, or in the worst scenario only initials for both first and surname. To overcome such challenge, we used scripts to convert the author's name in one, canonical form. We decided to use the form with initial of the first name and full surname, except for the UB-SEE employees with the same first name initial and surname. The reason for using such short canonical form comes from the fact that names of the non-UB-SEE co-authors are stored in that form. However, we also needed to manually go through the database to see if there were still some forms for the names used by the authors that were not covered by our Python scripts.

We wanted to know not only the pure relations between the authors of the papers published in journals, but also to extract information specifically for the researchers affiliated to UB-SEE. For this purpose, we used another database provided by UB-SEE with all the people employed by the UB-SEE in the period of interest. Current school's faculty, but also people that worked at this school in the past are listed. Using these two databases: the database of the published papers and the database of the employees at UB-SEE, we can model network with two different types of nodes, researchers from UB-SEE and researches from other institutions. Furthermore, if we want to go further with the classification of the researchers, we can find more fine grain solution and differentiate people from different departments at UB-SEE.

III. Network Modeling

The data we collected and preprocessed is used to create the collaboration network of coauthors. The created network can be further analyzed by the social network theory and social network specific metrics. Nodes of the network stand for the actors of social activity while the edges present the relations between the nodes. In our network, nodes are authors of the papers published in journals. Two authors are connected if they co-authored at least one paper together. The weight of the relation is the number of papers where these two authors collaborated, therefore the network is weighted network. Such network is undirected because all the edges are bidirectional. Since there are researchers that do not have published papers in common and there is no path from one researcher to another, our network is disconnected network.

From the original databases we created MS Excel files with information necessary for the network modeling. For this purpose, we used Python scripts with NetworkX library. First, we extracted all the authors, e.g. the nodes of our network, from the database with published papers. Then, from the database with UB-SEE's employees we set flag to indicate the type of the node. There are two possible types: UB-SEE's researcher and non UB-SEE's researcher. After creating separate file for network nodes, we had to create file with edges, e.g. relations existing in our network. Our edge has five components: author A, indicator if author A is UB-SEE researcher, author B, indicator if author B is UB-SEE researcher and number of papers on which A and B collaborated. In Figure 2 an example of the record in the new database can be seen. In particular example, this record means that researcher B. Reljin is UB-SEE's employee, P. Kostic is not UB-SEE's employee and they have 8 papers published together.

TABLE 1: AN EXAMPLE OF THE RECORD IN THE NEW DATABASE CREATED FOR NETWORK MODELING

Author_A	Flag_A	Author_B	Flag_B	Weight
B. Reljin	1	P. Kostic	0	8

For the visualization of the network we used Gephi software tool [5]. Figure 2 shows the network visualized by that tool, while Table 1 shows the overview of the basic properties of the network. We can see that majority of the nodes are non UB-SEE's researchers, e.g. researchers from other institutions. According to Table 1 there are 234 UB-SEE's nodes and 1660 non UB-SEE's nodes. That is also noticeable in Figure 2, where the blue nodes are non UB-SEE's nodes and the red nodes are UB-SEE's nodes. Size of a node corresponds to its degree. We can see that in the center of the network there are couple of big red nodes, which indicates that they are UB-SEE's researchers connecting a lot of other researchers and they present the epicenter of our network.

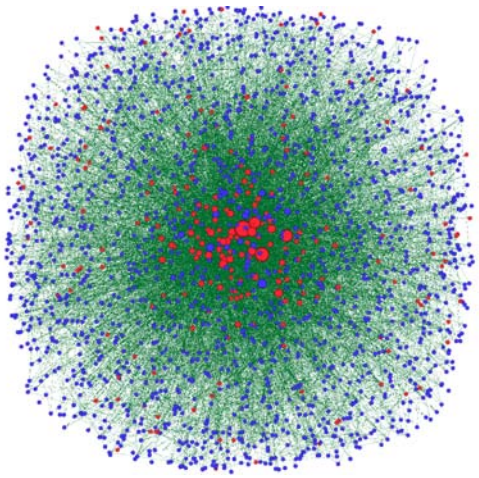

Fig. 2. Collaboration network visualized by Gephi software tool. Blue nodes present non-UB-SEE's researchers, while the red nodes stand for UB-SEE's researchers

TABLE 2: BASIC PROPERTIES OF THE COLLABORATION NETWORK

Property	Value
Number of nodes	1894
Number of UB-SEE's nodes	234
Number of edges	7407

In Figure 3 we present network consisting only of UB-SEE's researchers. We can see that even in such significantly smaller network there are obvious groups of researchers. Those groups actually present people from the same department that usually work together and publish as coauthors. There are no strong connections between different departments.

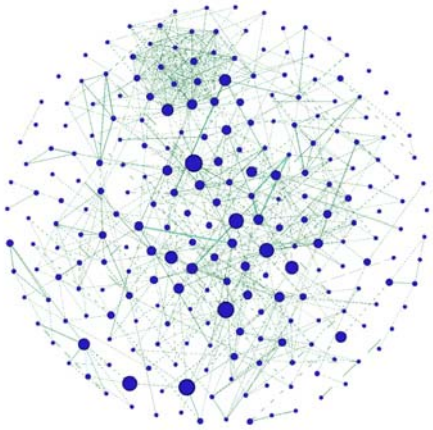

Fig. 3. Collaboration network consisting only of UB-SEE's researchers visualized by Gephi software tool

IV. NETWORK ANALYSIS

After initial modeling, we wanted to extract valuable information from our network. We used Python NetworkX package designed for creation, manipulation and study of complex networks and statistic module in Gephi to obtain deeper insight in the network's characteristics. Table 2 shows the properties we measured to understand the network and its values.

If the number of edges is E and number of nodes is N, then we can define graph density as 2*E/((N-1) * N). Our network has density of 0.004. In comparison to maximum graph density of 1, we can see that analyzed network is sparse, which means that there are a lot of edges which are not connected. That can be justified by the fact that there are many different fields which UB-SEE's researchers are involved. For that reason, they are not usually linked together. Also, researchers tend to work in small groups or well-established partnerships.

TABLE 3: AN OVERVIEW OF THE METRICS USED FOR NETWORK ANALYSIS

Metric	Value
Graph density	0.004
Average degree	7.585
Average weighted degree	14.082
Network diameter	11
Average path length	4.315
Connected components	10
Average clustering coefficient	0.82

In graph theory, the degree of a vertex of a graph is the number of edges that are incident to the vertex. Sum of the degrees for each node is equal to twice the number of edges: $\sum deg(n) = 2* E$. In our network, average degree of the node is 7.585. That means that each author has in average 8 co-authors. Since our network is presented as weighted graph, we can also define weighted degree as the sum of all the weights of edges that are incident to the vertex. Average weighted degree in our network is 14.082. The degree of a vertex can be interpreted as the number of nodes that a vertex relates to. We sorted all the authors based on their degree and the results can be seen in the Table 3. All 5 highest ranked authors based on the degree are UB-SEE's researchers except of P. Nikolic, who was a member of Serbian Academy of Sciences and Arts (SASA). P. Nikolic and K. Stankovic come from the field of engineering physics, while the other D. Popovic, A. Djordjevic, and Mirjana Popovic come from biomedical engineering, microwaves and antennas, and biomedical engineering, respectively. D. Popovic and A. Djordjevic are also SASA members.

TABLE 4: THE HIGHEST 5 RANKED AUTHORS IN THE NETWORK BASED ON THE DEGREE

Researcher	Degree
D. Popovic	122
Mirjana Popovic	111
A. Djordjevic	81
P. Nikolic	79
K. Stankovic	77

Degree distribution shows us that our network follows a power law, i.e. our network is scale-free network. Degree distribution is shown in Figure 3.

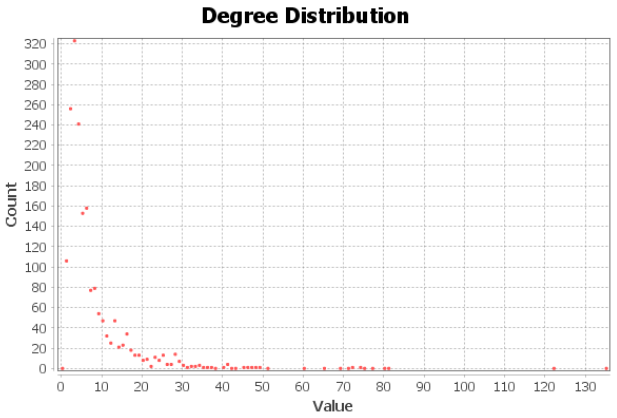

Fig. 3. Degree distribution of analyzed network

Network size is usually referred to the number of nodes, which is 1894 in our network, but to present linear size of a network, we can use network diameter. Network diameter is the longest of all calculated shortest paths in a network. To calculate that parameter, we first need to calculate all shortest paths in a network and then to choose the maximum value. In our case, network diameter is 11.

After calculating the shortest path lengths for all existing pair of nodes, we can find the average path length for our network. This shows how many steps in average are needed to get to one node from another if two nodes are connected. In our network this property has value of 4.315. Both average path length and diameter values showed us that collaboration network of UB-SEE

researchers exhibits small-world phenomenon, similar to Facebook and other social networks [7] .

Connected component of an undirected graph is a subgraph in which any two nodes are connected and there is no connection with the nodes outside of that subgraph. Number of connected components can show us how are the nodes grouped in the graph. In our network, there are 10 connected components. There is one big connected component consisting of 1839 nodes, and other 9 components have all together 55 nodes. Those components are mostly related to those authors that were employed at the UB-SEE for a short period and published papers only with non-UB-SEE authors.

Clustering coefficient of a node is the ratio of the number of existing links between node's neighbors and the maximum possible number of such links. That metric can show us "*all-my-friends-know-each-other*" property. If there are K neighbors and L links between them, then we define clustering coefficient of that node: 2*L/((K-1) *K). After calculating clustering coefficient for each node, we can then find the average clustering coefficient, which is 0.820 in our network. Since the maximum clustering coefficient is 1, we can state that there is medium to high probability that two arbitrary neighbors of a node are linked.

An indicator that shows how well network decomposes into modular communities is called modularity. This value shows us the structure of the network. It measures the strength of division of the network into departments. For our network modularity score is 0.842. We can interpret this value in the following way: our network has dense connections between the nodes within modules but sparse connections between nodes in different modules. That is another proof that within our collaboration network authors are grouped in small tightly connected groups and there are not many connections between different groups. Those groups roughly resemble the actual division of departments at UB-SEE.

We analyzed communities detected in our network. There are 30 classes having from 1 to 220 nodes. Distribution of size of communities can be seen in Figure 4. Researchers employed by UB-SEE are mostly grouped into communities based on their department. There are not many researchers from different departments in the same community, however one department can be separated into more than one community.

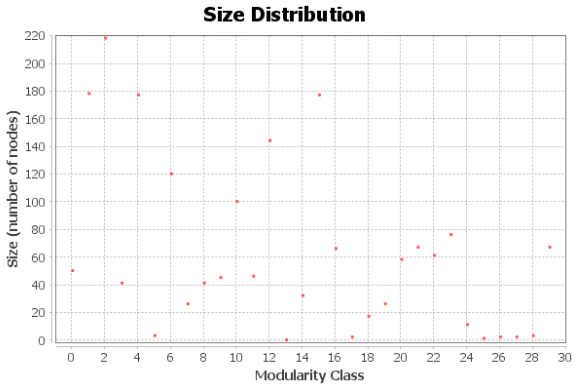

Fig. 4. Distribution of size of communities

Betweenness centrality is the measure of centrality of a node based on shortest paths. This value indicates how many shortest paths pass through a node. The higher betweenness centrality is the more important node is, since it connects more nodes to each other. Table 5 shows the 5 highest ranked authors based on normalized betweenness centrality metric. The metric is normalized by 2/((N-1) *(N-2)) where N is the number of nodes.

The results from Table 5 show that Mirjana Popovic, D. Popovic, and A. Djordjevic also serve as bridges in their scientific fields, not only that they have many co-authors. Two other important bridges are professor emeritus S. Stankovic from signals and systems department, and V. Milutinovic, IEEE fellow, from the computer engineering department.

TABLE 5. THE HIGHEST 5 RANKED AUTHORS IN THE NETWORK BASED ON BETWEENNESS CENTRALITY

Researcher	Betweenness centrality
D. Popovic	0.117621
A. Djordjevic	0.098955
Mirjana Popovic	0.098258
S. Stankovic	0.092957
V. Milutinovic	0.061902

V. CONCLUSION

Institutional databases offer great possibilities to analyze scientific production and co-authorship (collaboration) patterns of its employees. In this paper, we have analyzed collaboration network of UB-SEE employees and their collaborators from other institutions. We showed that analyzed network exhibits the properties of a social network and point out the most important researchers in terms of their collaboration patterns.

In the future, we plan to compare the data obtained from institutional database with the data from external sources, such as Web of Science and Scopus index databases. Also, it is interesting to analyze the institutions of the non-UB-SEE authors and their connections with researchers from different UB-SEE departments.

REFERENCES

[1] É. Archambault, D. Campbell, Y. Gingras, and V. Larivière, "Comparing bibliometric statistics obtained from the Web of Science and Scopus," *Journal of the American Society for Information Science and Technology,* vol. 60, no. 7, pp. 1320-1326, 2009.

[2] W. Glänzel and A. Schubert, "Analyzing scientific networks through co-authorship," in *Handbook of Quantitative Science and Technology Research*, H. F. M. e. a. (eds.), Ed.: Kluwer Academic Publishers, 2004, ch. 11, pp. 257-276.

[3] D. Milovančević, M. Mišić, and J. Protić, "Mrežna analiza naučne kolaboracije zaposlenih na Elektrotehničkom fakultetu u Beogradu na osnovu institucionalne evidencije objavljenih radova," XXIV Skup TRENDOVI RAZVOJA, Kopaonik, Serbia, 2018.

[4] M. Savić, M. Ivanović, and B. D. Surla, "Analysis of intra-institutional research collaboration: a case of a Serbian faculty of sciences," *Scientometrics,* vol. 110, no. 1, pp. 195-216, 2017.

[5] M. Bastian, S. Heymann, and M. Jacomy, "Gephi: an open source software for exploring and manipulating networks," *ICWSM,* vol. 8, pp. 361-362, 2009.

[6] I. Mitrović and J. Protić, "Problems with affiliations, names and personal identity in the process of evaluating higher education institutions," *EDULEARN14 Proceedings,* pp. 2524-2533, 2014.

[7] L. Backstrom, P. Boldi, M. Rosa, J. Ugander, and S. Vigna, "Four degrees of separation," in *Proceedings of the 4th Annual ACM Web Science Conference*, 2012: ACM, pp. 33-42.