Are papers all that counts? A bibliometric analysis of the Slovenian scientific community

Aymeric Dupuis
Jožef Stefan Institute
Ljubljana, Slovenia
aymeric.dupuis@etu.univ-nantes.fr

Sašo Džeroski
Jožef Stefan Institute
Ljubljana, Slovenia
saso.dzeroski@ijs.si

Boshko Koloski
Jožef Stefan Institute
Ljubljana, Slovenia
boshko.koloski@ijs.si

Matej Martinc
Jožef Stefan Institute
Ljubljana, Slovenia
matej.martinc@ijs.si

Abstract

We conduct a bibliometric analysis of the Slovenian science by scraping the data from Slovenian current research information system (SICRIS) and using it to build a knowledge graph, representing a network of all Slovenian scientific fields and a large majority of Slovenian researchers. By analyzing this network using different graph measures, we obtain valuable insights into the connections between different scientific fields and researchers in Slovenian science. Additionally, we show the importance of graph measures as measures of scientific excellence, since they measure very different aspects of scientific success than the traditional citation metrics.

Keywords

bibliometrics, Slovenian scientific community, knowledge graphs

1 Introduction

With the growth and diversification of the scientific enterprise, obtaining empirical evidence on the research process is crucial for enhancing its efficiency and reliability. Meta-research and bibliometrics are developing scientific disciplines, seeking to analyse, evaluate and refine research practices, and several studies have focused on the analysis of the global scientific endeavour, e.g., identifying most prominent scientists and fields [7]. These studies also focus on the problem of how to properly rank scientific excellence and scientific outputs in general, warning that one should not rely on just a few metrics to obtain a comprehensive picture of the actual impact a specific scientist has [8].

Until now, very few studies have tackled the analysis of scientific ventures at national level, and to our knowledge, there has been no study covering the Slovenian scientific landscape specifically. This kind of research is nevertheless important and could potentially influence policies that would improve scientific production and enable effective distribution of research funds and resources.

In this study, we try to address the identified research gaps by 1.) drawing the map of Slovenian scientific research that would enable proper decision making and policy formulation, and 2.) proposing new metrics of scientific excellence that would allow us to obtain a more complete view of the impact a scientist or a discipline as a whole has. More specifically, our contributions are the following:

- Using the collected data about the Slovenian scientists and their projects, covering different scientific fields and a large majority of researchers working in Slovenian science, we conduct a graph analysis of connections between different fields and researchers. By drawing a comprehensive map of connections between actors and fields, we identify the most important researchers and scientific fields that connect others and play a vital role in the Slovenian scientific ecosystem.
- We created a new ranked list of Slovenian scientists according to graph based metrics, which were not available in any of the previous analyses or databases. We argue that these metrics measure the importance of a role that a specific scientist has in a research community, i.e., their influence which allows them to act as a bridge or a hub connecting scientists from different fields.

2 Related work

Studies in bibliometrics (see [4] for a comprehensive survey of techniques used for measuring scientific excellence) have recently gained traction in parallel with the success of the scientific enterprise, which has grown in both size and diversity, and with the availability of data. According to Ioannidis et al. [7], research on research is becoming important due to the mounting evidence suggesting an alarming drop in reproducibility of research findings, the growing inefficiency of the scientific process, and the fact that the number of false positives in the literature is exceedingly high. To address these problems, they propose a meta-research divided into five main categories that should be studied: methods, reporting, reproducibility, evaluation, and incentives. Studying these five areas would correspondingly allow for five distinct insights into how to perform, communicate, verify, evaluate, and reward research.

Recently, several studies also tackled the problem of how to properly rank scientist and scientific outputs in general. For example, Ioannidis et al. [8] addressed the increasing prevalence of multiauthorship observed in several fields and how this phenomenon affects the effectiveness of the informativeness of citation metrics. They also explored how sensitive the indicators are to self-citation and alphabetic ordering of authors. They concluded that multiple indicators should be used for ranking, as a composite of different metrics gives a more comprehensive picture of the actual impact that a specific scientist has. They also acknowledged that no single or composite citation indicator can be expected to select all the best scientists.

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/https://doi.org/10.70314/is.2024.sikdd.11

Several studies employed graph-based metrics to enrich the assessment of bibliometric analysis [4, 1]. Network metrics such as degree of centrality, betweenness centrality, eigenvector centrality, closeness centrality, and PageRank were used to pinpoint the relative importance of research constituents (i.e., researchers and institution), which may not necessarily be reflected just through publications. In a large majority of cases, these metrics were calculated on co-authorship graphs.

The studies that would cover Slovenian scientific environment are very scarce. In fact, we are aware of just one, the study by [2], where they claim that research performance is highly dependent on the conditions of (national) research environments. They focus on analyzing research activity in six eastern European countries, namely Croatia, Estonia, Hungary, Latvia, Lithuania, and Slovenia, and try to determine and compare the effectiveness of research in a specific country by obtaining the number of articles belonging to the most cited 10% and the most cited 1% articles in the corresponding subject area and publication year for each country. Their empirical analysis addresses three levels: cross-country, cross-institution, and cross-researcher comparison. The study concludes that Hungary is the country with the highest output, followed by Croatia and then Slovenia, when it comes to the number of influential articles published.

3 Methodology

In this section, we describe our methodology, namely 1.) how we gather the data and 2.) how we analyze these data to obtain a map of the Slovenian scientific community.

3.1 Data Retrieval

Data were retrieved from the Slovenian Current Research Information System (SICRIS) website[1], which lists more than 35,000 researchers working in Slovenian research institutions. Data collection from the SICRIS website proved challenging, as information about a specific researcher can only be obtained by scraping his/her Web page on SICRIS. This required finding a solution to quickly retrieve data from more than 35,000 different pages, and to achieve this, we used the Python Asyncio[2] and BeautifulSoup[3] libraries, which allow the asynchronous connection to several dozen pages simultaneously and extraction of the required data.

Since the script sometimes took several seconds to connect to a specific page, which could quickly accumulate, resulting in considerable overall slowdowns, we optimized the procedure and identified potential slowdowns. Our proposed solution was to implement a strategy that involved canceling the connection and adding the URL to a list whenever a page failed to connect within a 0.5-second time frame. This timeframe was chosen after several trials and was found to be the best compromise. Once all pages had been visited, we repeatedly tried to reconnect to the URLs on this list until it was empty. This change significantly reduced the time required to retrieve all our data. Once all the data was retrieved, we used the Pandas library[2] for data manipulation, which allowed us to export the results into Excel spreadsheets, appropriate for further processing.

From SICRIS, we extracted research areas for each scientist and various bibliometric indicators of their impact, namely A", A', A1/2, citation metrics based on a quantitative assessment of publications in exceptional, high quality and important venues, respectively. We also extracted the A1 metric, which represents a weighted sum of these three metrics, a CI10 metric measuring the number of pure citations of scientific work in the last 10 years, the CImax metric measuring the number of citations in the most cited work, and the h10 metric representing the h-index in the last ten years. Furthermore, we extracted the SICRIS points, a conglomerate metric combing several distinct metrics mentioned above, and the A3 metric, which measures the amount of funds a specific researcher received for his research activity outside of the Slovenian National Research Agency (ARIS).

Finally, the SICRIS database also contains information on projects financed by the Slovenian national research agency in which a specific researcher participated. Scraping this information provided us with an important insight into collaborations between different scientists and fields, allowing us to build collaboration graphs, calculate several graph-based ranking criteria and draw the map of the Slovenian scientific community.

3.2 Methods

Once the data was obtained, we conduct two distinct analysis steps, namely 1.) graph construction and analysis, and 2.) correlation analysis

3.2.1 Graph construction and analysis. To construct the necessary graphs, we used the Python NetworkX library [6]. Using the data from SICRIS, which contain information about project collaboration, we created an undirected graph as follows: all researchers who participated in at least one project are represented by a node, and nodes of researchers who worked together on a project are connected by weighted edges, in which the weights represent the number of shared projects. By removing the isolated nodes, we ended up with a graph with a total of 20,012 nodes and 618,871 edges.

In the next step, we apply several graph statistics and measures in order to obtain several node rankings, each of them measuring a different aspect of the importance a specific node (i.e., a researcher) has in the graph. More specifically, we calculate PageRank (PR), Betweenness centrality (BC), and Eigenvector centrality (EC) measures.

In the context of our graph, the **PageRank** [3] algorithm is applied to evaluate the influence of researchers within the collaboration network. Thus, researchers who are strongly connected to other researchers, who also have many connections (i.e, the so-called hubs in the graph), will have a higher PR score, reflecting their importance and influence in the Slovenian research community. On the other hand, the **Betweenness Centrality** [5] measure evaluates the role of each researcher as an intermediary or a bridge between other researchers. This measure is based on the idea that researchers who are on many collaboration paths between other researchers are considered central and influential in the network. In our contexts, it helps to better understand the structure of the collaboration network among researchers. Researchers with high BC are those who play a crucial role in creating links between different subgroups of researchers and interdisciplinary connections. In practical terms, BC evaluates the number of times a researcher is traversed by the shortest paths connecting other researchers in the network. Thus, researchers who are frequently used as pathways for collaboration among their peers obtain higher BC scores.

[1] https://cris.cobiss.net/ecris/si/en
[2] https://docs.python.org/3/library/asyncio.html
[3] https://www.crummy.com/software/BeautifulSoup
[2] https://pandas.pydata.org/

[1] https://networkx.org/

Another graph centrality measure that we applied to the created graph is the **Eigenvector centrality** [9]. This measure evaluates the influence of a researcher taking into account both the quality and the quantity of connections. EC assigns more weight to connections that include influential researchers. Thus, a researcher connected to influential researchers will be assigned a high score, reflecting potentially greater influence within the network. This measure helps to detect researchers who, even with fewer direct connections, occupy strategic positions in the collaboration network. While this may seem similar to the PR algorithm, there are some differences. Unlike PR, which primarily focuses on the popularity of links, Eigenvector centrality also takes into account the quality of connections. This means that even if a researcher does not have a large number of direct connections, if they are connected to influential researchers, their Eigenvector centrality score can be high. In summary, while both measures aim to evaluate the influence of researchers in a network, they do so through slightly different approaches, thus offering complementary perspectives for analyzing the structure and importance of actors within the collaboration network.

Our second important area of focus in our research is the collaboration between different fields. To build a graph that would represent interdisciplinary collaboration between fields, we grouped all researchers from the same field into a single node, representing an entire field, i.e., we obtain a node for each scientific field found on SICRIS. Similar to the previous graph, edges and their weights represent collaborations on a project between researchers in the linked fields.

3.2.2 *Correlation analysis.*

In order to better understand the metrics from SICRIS and to evaluate the relevance of our scores, we deemed it pertinent to explore the correlation between all our data. This analysis has two main purposes. First, we aim to test the **hypothesis 1** that the new graph ranking we presented, measure different aspects of scientific excellence than the more established measures based on number of citations or publications available on the SICRIS web page. This hypothesis would be deemed correct if one-on-one correlations scores between the newly proposed graph measures and other measures would be low, and incorrect if correlations would be high.

Additionally, we wish to explore the correlation between the established measures available on the SICRIS web page. More specifically, we wish to test the **hypothesis 2** that these measures are strongly correlated, which would indicate that they essentially all measure a very similar aspect of scientific excellence, which is problematic. In order to obtain one-on-one correlations between all measures, we calculate the Spearman correlation coefficient among all of them and then display it through a heatmap.

4 Results

In Table 1, we present some of the results of the graph analysis conducted on the graph of nodes representing researchers, connected by edges representing project collaborations. More specifically, we present 10 best ranked researchers in the SICRIS dataset according to the average between ranks of the three newly proposed graph-based measures, their declared scientific fields, and their ranking (i.e., lower is better) according to the SICRIS points, BC, EC and PR measures.

Note that while the table does contain some highly ranked researchers according to the SICRIS points (e.g., Dr. Sašo Džeroski is ranked as 33rd out of roughly 20K researchers according to this criteria), several researchers in the table are ranked relatively low according to SICRIS points (e.g., the best ranked researcher according to our novel three measures, Dr. Branimir Leskošek, is ranked as 5731th according to the SICRIS points). This finding supports **hypothesis 1** that the proposed new measures measure different aspects of scientific excellence than the more established citation measures. Another important observation is that 7 out of 10 best ranked scientists appear to be active in two fields. This might suggest that they are (or have been) involved in several interdisciplinary projects, which could have a positive influence on the newly proposed graph-based metrics.

In Figure 1, we present the heatmap of the correlations between the different metrics extracted from SICRIS website and the newly proposed graph-based metrics. We observe a strong correlation between PR and BC, 0.7, which might suggest that researchers who collaborate with a wide range of colleagues from different fields are more likely to work with the most important ones.

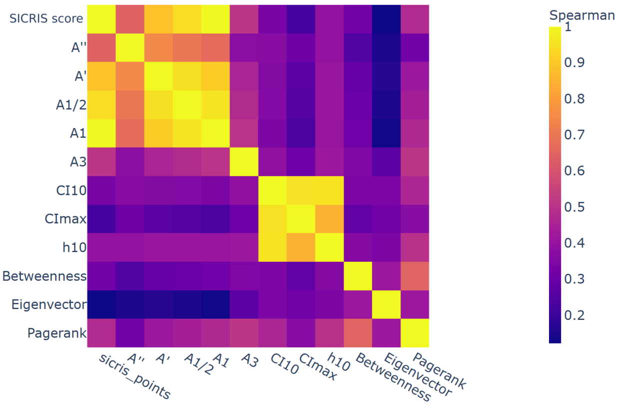

Figure 1: Heatmap of the Spearman correlation among metrics.

We also observe very strong correlations in the top left corner of the heatmap. While a strong correlation was expected, as A”, A’, A1/2 and A1 are all scores based on the number of publications (in venues of different qualities), the almost perfec correlation between the SICRIS points and A1 (which suggest they measure exactly the same aspect of the scientific impact) is surprising. This finding supports **hypothesis 2** that the current SICRIS measures all measure a very similar aspect of scientific excellence. On the other hand, there is no strong correlation between any of the newly proposed graph-based metrics and metrics extracted from the SICRIS website.

In Table 2, we present the results of our study of interdisciplinary collaboration between different scientific fields. The graph metrics were obtained from a graph of nodes representing fields and edges representing interdisciplinary project collaborations. Note that the field of Computer science and informatics ranks first according to all the criteria. On the other hand, most interdisciplinary collaborations are conducted by the researchers from the field of Chemistry, which ranked as third according to the average (AVG) between the ranks of three graph-based metrics, PG, BC and EV.

5 Conclusions

The graph based bibliometric analysis of the Slovenian scientific community shows that current citations based metrics do not cover some aspects of scientific excellence, such as researcher’s

Table 1: 10 best ranked researchers in the SICRIS dataset according to the average between ranks of the three newly proposed measures, BC, EC and PR. We do not show metric scores, but ranks according to scores (i.e., lower value is better).

Researcher	Field 1	Field 2	SICRIS points	BC	EC	PR	AVG
15355 PhD Branimir Leskosek	Public health (occupational safety)	Computer science and informatics	5731	8	4	31	14
06013 PhD Damjana Rozman	Biochemistry and molecular biology	Metabolic and hormonal disorders	704	21	2	33	18
11279 PhD Nives Ogrinc	Control and care of the environment	Animal production	182	7	50	3	20
27733 PhD Tina Kosjek	Control and care of the environment	Pharmacy	809	2	73	9	28
22459 PhD Tadeja Rezen	Neurobiology	Microbiology and immunology	1837	61	3	49	37
22621 PhD Polonca Ferk	Metabolic and hormonal disorders	Pharmacy	5059	13	8	103	41
12688 PhD Kristina Gruden	Biotechnology	/	219	44	139	6	63
08800 PhD Gregor Sersa	Oncology	/	71	3	185	1	63
12315 PhD Ester Heath	Control and care of the environment	Chemistry	208	62	115	23	66
11130 PhD Šašo Dzeroski	Computer science and informatics	/	33	1	195	20	72

Table 2: Scientific fields as defined in the SICRIS database, sorted according to the average (AVG) between the ranks (lower score is better) of three graph-based metrics, PG, BC and EV.

Rank	Field	Collaborations	PR	EC	BC	AVG	Rank	Field	Collaborations	PR	EC	BC	AVG
1	Computer science and informatics	81248	1	1	1	1.0	36	Textile and leather	21080	27	41	39	35.67
2	Materials science and technology	88934	4	3	4	3.67	37	Animal production	34982	29	29	50	36.0
3	Chemistry	101139	2	2	12	5.33	38	Political science	13598	46	37	27	36.67
4	Control and care of the environment	52648	5	8	9	7.33	39	Anthropology	9860	53	36	24	37.67
5	Physics	50010	3	9	14	8.67	40	Ethnology	6698	65	39	11	38.33
6	Plant production	74535	6	6	16	9.33	41	Cardiovascular system	20793	28	43	45	38.67
7	Systems and cybernetics	45584	7	10	23	13.33	42	Telecommunications	14068	41	45	31	39.0
8	Biology	58879	12	7	21	13.33	43	Veterinarian medicine	30954	32	34	60	42.0
9	Civil engineering	36466	22	13	6	13.67	44	Metabolic and hormonal disorders	18518	30	46	55	43.67
10	Biochemistry and molecular biology	79725	11	5	25	13.67	45	Metrology	12978	34	52	47	44.33
11	Neurobiology	45680	14	12	19	15.0	46	Law	7480	54	49	32	45.0
12	Biotechnology	87261	8	4	33	15.0	47	Psychology	8583	51	55	29	45.0
13	Interdisciplinary research	22946	9	33	5	15.67	48	Human reproduction	21535	35	42	58	45.0
14	Public health (occupational safety)	30400	10	25	13	16.0	49	Process engineering	15340	36	47	53	45.33
15	Educational studies	23518	33	15	3	17.0	50	Hydrology	12396	40	53	44	45.67
16	Mathematics	30680	17	20	20	19.0	51	Architecture and Design	4242	58	57	22	45.67
17	Manufacturing technologies and systems	38874	18	14	26	19.33	52	Philosophy	7380	57	44	43	48.0
18	Forestry, wood and paper technology	30620	19	28	15	20.67	53	Sport	10013	43	54	49	48.67
19	Geography	18555	39	23	2	21.33	54	Geodesy	7760	45	56	51	50.67
20	Economics	26891	31	16	18	21.67	55	Electric devices	13633	42	51	59	50.67
21	Microbiology and immunology	54175	16	11	42	23.0	56	Literary sciences	6399	61	50	48	53.0
22	Sociology	19922	44	17	10	23.67	57	Traffic systems	4448	48	60	52	53.33
23	Pharmacy	41125	15	18	41	24.67	58	Culturology	7240	60	48	54	54.0
24	Linguistics	18176	49	19	7	25.0	59	Technology driven physics	6876	47	59	64	56.67
25	Chemical engineering	33753	13	27	38	26.0	60	Communications technology	4388	52	63	56	57.0
26	Energy engineering	32762	23	21	40	28.0	61	Psychiatry	2481	55	65	61	60.33
27	Computer intensive methods and applications	26942	20	32	34	28.67	62	Criminology and social work	2324	66	62	62	63.33
28	Mechanics	26444	24	31	36	30.33	63	Mining and geotechnology	2342	59	68	63	63.33
29	Oncology	37101	21	24	46	30.33	64	Theology	2941	67	58	66	63.67
30	Geology	26961	37	26	28	30.33	65	Ethnic studies	2398	63	61	67	63.67
31	Electronic components and technologies	28858	26	30	37	31.0	66	Art history	1408	70	64	57	63.67
32	Historiography	12390	56	22	17	31.67	67	Archaeology	1177	68	66	65	66.33
33	Urbanism	8669	50	40	8	32.67	68	Information science and librarianship	792	62	70	70	67.33
34	Mechanical design	22352	25	38	35	32.67	69	Stomatology	391	64	71	68	67.67
35	Administrative and organisational sciences	18563	38	35	30	34.33	70	Landscape design	1046	69	67	71	69.0
							71	Musicology	748	71	69	69	69.67

role of connecting a wider research community. Our correlation analysis indicates that existing measures of scientific excellence extracted from the SICRIS web page are strongly correlated. In the future, we plan to expand this analysis to also measure the impact of Slovenian scientists on the global scientific enterprise and conduct additional research to try to find certain patterns across disciplines, or institutions.

6 Acknowledgments

The authors acknowledge the financial support from the Slovenian Research Agency for research core funding for the programmes Knowledge Technologies (No. P2-0103).

References

[1] Njål Andersen. 2021. Mapping the expatriate literature: a bibliometric review of the field from 1998 to 2017 and identification of current research fronts. *The International Journal of Human Resource Management*, 32, 22, 4687–4724.

[2] Lutz Bornmann. [n. d.] Research excellence in eastern europe: a bibliometric study focusing on croatia, estonia, hungary, latvia, lithuania, and slovenia.

[3] Sergey Brin and Lawrence Page. 1998. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30, 1-7, 107–117.

[4] Naveen Donthu, Satish Kumar, Debmalya Mukherjee, Nitesh Pandey, and Weng Marc Lim. 2021. How to conduct a bibliometric analysis: an overview and guidelines. *Journal of business research*, 133, 285–296.

[5] Linton C. Freeman. 1977. A set of measures of centrality based on betweenness. *Sociometry*, 40, 1, 35–41. Retrieved June 27, 2024 from http://www.jstor.org/stable/3033543.

[6] Aric Hagberg, Pieter J Swart, and Daniel A Schult. 2008. Exploring network structure, dynamics, and function using networkx. In *Proceedings of the 7th Python in Science Conference (SciPy2008)*. Los Alamos National Laboratory (LANL), Los Alamos, NM (United States), 11–15.

[7] John PA Ioannidis, Daniele Fanelli, Debbie Drake Dunne, and Steven N Goodman. 2015. Meta-research: evaluation and improvement of research methods and practices. *PLoS biology*, 13, 10, e1002264.

[8] John PA Ioannidis, Richard Klavans, and Kevin W Boyack. 2016. Multiple citation indicators and their composite across scientific disciplines. *PLoS biology*, 14, 7, e1002501.

[9] Paul Turán, editor. 1969. *Publications of edmund landau. Number Theory and Analysis: A Collection of Papers in Honor of Edmund Landau (1877–1938)*. Springer US, Boston, MA, 335–355. ISBN: 978-1-4615-4819-5. DOI: 10.1007/978-1-4615-4819-5_23.