PLOS ONE

RESEARCH ARTICLE

Analysis of co-authorship networks among Brazilian graduate programs in computer science

Alex Nunes da Silva Junior[1], **Matheus Montanini Breve**[1], **Jesús Pascual Mena-Chalco**[2], **Fabrício Martins Lopes**[1]*

1 Computer Science Department, Universidade Tecnológica Federal do Paraná (UTFPR), Cornélio Procópio, PR, Brazil, **2** Center for Mathematics, Computing, and Cognition, Universidade Federal do ABC (UFABC), Santo André, SP, Brazil

* fabricio@utfpr.edu.br

OPEN ACCESS

Citation: Nunes da Silva A, Junior, Breve MM, Mena-Chalco JP, Lopes FM (2022) Analysis of co-authorship networks among Brazilian graduate programs in computer science. PLoS ONE 17(1): e0261200. https://doi.org/10.1371/journal.pone.0261200

Editor: Diego Raphael Amancio, University of Sao Paulo, BRAZIL

Received: April 5, 2021

Accepted: November 26, 2021

Published: January 18, 2022

Data Availability Statement: The data underlying the results presented in the study are available from https://sucupira.capes.gov.br/sucupira/ and http://lattes.cnpq.br/.

Funding: This study was financed by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) (Grant number 406099/2016-2) and the Fundação Araucária e do Governo do Estado do Paraná/SETI (Grant number 035/2019, 138/2021 and NAPI Bioinformática). All grants were awarded to FML.

Abstract

The growth and popularization of platforms on scientific production has been the subject of several studies, producing relevant analyses of co-authorship behavior among groups of researchers. Researchers and their scientific productions can be analysed as co-authorship social networks, so researchers are linked through common publications. In this context, co-authoring networks can be analysed to find patterns that can describe or characterize them. This work presents the analysis and characterization of co-authorship networks of academic Brazilian graduate programs in computer science. Data from Brazilian researchers were collected and modeled as co-authoring networks among the graduate programs that researchers take part in. Each network topology was analysed with complex network measurements and three proposed qualitative indices that evaluate the publication's quality. In addition, the co-authorship networks of the computer science graduate programs were characterized in relation to the assessment received by CAPES, which attributes a qualitative grade to the graduate programs in Brazil. The results show the most relevant topological measurements for the program's characterization and the evaluations received by the programs in different qualitative degrees, relating the main topological patterns of the co-authorship networks and the CAPES grades of the Brazilian graduate programs in computer science.

Introduction

Social networks have attracted a great deal of attention for decades. Some studies on this topic date as far back as the early 30s and were then mostly done by anthropologists and sociologists [1, 2]. With the increasing use of graph theory to represent social constructs [3], the concepts of small-world [4] and scale-free [5] networks, complex networks [6, 7], and their applications in different contexts, social networks have now drawn the attention of researchers from diverse disciplines, such as computer science, biology, mathematics, chemistry and physics.

One aspect of this research, namely the parallels between social networks and academic collaborations has not gone unnoticed. Research collaboration can be carried out at different

Competing interests: The authors have declared that no competing interests exist.

levels by researchers with some common goals to co-produce new scientific knowledge [8]. The limits of research collaboration may still be somewhat diffuse given the different forms of interaction between research actors. In this context, the tangible results of collaborations, such as scientific publications, have become important elements to study and analyse collaborations. Academic social networks are heterogeneous networks composed of entities that represent academic actors (e.g., researcher, institution, research group) or products of the result of the performance of these actors (e.g., conference paper, journal paper, book). The analysis of academic social networks allow to observe and study the way of communication and interaction between academic entities, as well as the dissemination of scientific knowledge [9].

A significant number of analyses on academic collaborations between researchers have been made since the establishment of the field of "Scientometrics" in the 70s, although works with similar ideas date as further back as the early 20th century [10]. Some examples of analyses of academic collaborations include nation-wide investigations, such as in Slovenia [11], Brazil [12–16], Germany [17] and Turkey [18]. Other works restrict the analysis to a certain discipline within a country, for example, conducting an analysis of only computer science publications in Brazil [19]. Limiting the scope to certain databases is also common, with, for example, the Zentralblatt MATH database in Germany being explored in [20] or two of Newman's works [21, 22], where publications between 1995 and 1999 in the areas of physics, biomedical research, and computer science in four specific databases were studied. International collaborations have also been researched, as in [23–25].

There is a rising competitiveness within academia [26], which leads to the development of indicators and world university rankings—such as the THE, QS and Shanghai rankings—or in the many ways to rank researchers based on their academic production with citation and productivity metrics, such as the SCI, *h*-index and PlumX. These metrics are now used worldwide for important decisions concerning funding, hiring and research directions in academia, a path that has led to criticisms [27, 28].

The majority of the contributions in this field, however, address the productivity of individual researchers, oftentimes analyzing how their academic collaboration networks evolve with time. For example, in [29] the question of whether more collaborative researchers tend to have more scientific impact was answered by analyzing their collaboration networks, with [30, 31] and [32] addressing similar questions. Few contributions focus on the performance of specific institutions or groups, such as universities or graduate programs. For example, high-quality Brazilian graduate programs were compared to international programs of excellence based on different universities and citations rankings in [33]. The relationship between a governmental quality assessment and internal academic collaborations among researchers in Brazilian computer science graduate programs was analysed with data from the DBLP database in [34]. Both [34, 33] use the results of the evaluation done by the governmental institution CAPES as a basis for comparison.

The evaluation of Brazilian graduate programs is performed by the Coordination of Superior Level Staff Improvement (CAPES), a governmental institution of the Ministry of Education. The evaluation process takes into account several aspects, such as academic personnel, ongoing research projects, program curriculum, academic production, regional economic and social impacts. The results are released every 4 years and each graduate program is granted a score, called CAPES grade. These vary between 1 and 7 with the latter being the highest possible grade. Since all graduate programs must be evaluated by CAPES and their grades determine whether they can continue to operate and, to an extent, how much government funding they receive, it would be interesting to investigate how the structure of academic collaboration within graduate programs correlate to their CAPES grade.

In this context, this work presents an approach for the analysis of co-authorship networks of Brazilian graduate programs in Computer Science. Thus, the aim of this work is to characterize and to identify topological patterns that correlate with the grade received from CAPES, considering three evaluation periods. For this, complex network and vulnerability measurements were adopted. The co-authorship networks were built based on collected data from the Lattes [35] and Sucupira platforms [36], which, respectively, contain the publications by Brazilian researchers and in which graduate programs they are currently active.

The achieved results can be of great relevance for the coordination of graduate programs, given the factors that differentiate the best evaluated from the least evaluated programs. Thus, being able to improve the understanding of the evaluations received by graduate programs and suggesting the topological patterns in the co-authorship networks that can help the coordinators of graduate programs visualize possible improvements in their programs and work towards better CAPES grades, i.e., with direct propositional actions in their respective programs.

Materials and methods

This section presents the materials and methods adopted to extract and process data in order to generate the academic co-authorship networks. These networks are then analyzed using complex network measurements. Fig 1 presents an overview of the proposed approach.

Data sources

The Lattes [35] is a Brazilian online platform where researchers can create their academic resumes and list their publications, research projects, etc. This platform is used throughout the country as a decision factor for hiring university staff, distributing federal financial support for research, and university scoring, among others. Researchers are also evaluated based on their production listed in Lattes resumes, a determining factor for obtaining research grants, for example. As a result, government institutions encourage researchers to keep their resumes up to date and complete. The Lattes platform has over 6 million online resumes. It provides a significant amount of reliable data, which can be extracted and analyzed to determine the key features that distinguish low and high-ranking universities based on the academic performance of their researchers.

The Lattes Resume of 1,644 researchers, affiliated with Brazilian graduate programs in Computer Science, were extracted from the Lattes Platform. The data acquisition process is further detailed in Sec. Data gathering.

The quality of Brazilian graduate programs is assessed since 1998 by CAPES and the results of the evaluation are publicly accessible on the CAPES website. The CAPES assessment is carried out with internal analyses with information systems and professionals who, through studies, reach the CAPES grade [37].

The assessment is carried out in periods, and until 2012 the periods were 3 years; however, due to CAPES changes, from 2013 the assessment period is every 4 years. The CAPES evaluation is carried out by specialized committees in each area of knowledge. Currently, CAPES has 49 evaluation areas [38], which publish specific documents with the guiding criteria for evaluation following common guidelines, however also presents their particularities in the evaluation of each area. In addition, considering the CAPES grades, each graduate program is evaluated relatively within each area of knowledge, i.e. the programs of the same area are compared among themselves for the attribution of grades within each area. The grade received leads to several consequences for the graduate programs, such as programs with grades 1 and 2 are not recommended by CAPES, programs with grades 3 may offer only masters courses, programs

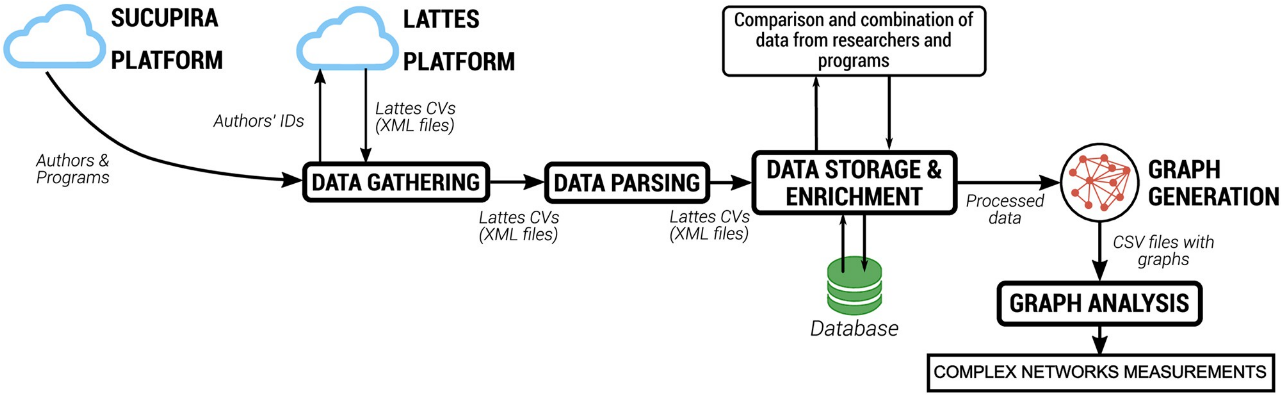

Fig 1. Process flow. Clouds show internet data access. Blocks in bold represent processes. Artifacts generated by each process are described next to the arrows that show the directions of the flow.

https://doi.org/10.1371/journal.pone.0261200.g001

with grades 4 or higher may offer masters and PhD courses, programs with grades 5 or higher may participate in some government research funding initiatives, while others do not, to name a few. Therefore, for the graduate programs in Brazil it is mandatory the evaluation and the attribution of the CAPES grade. More detailed technical data on how CAPES performs this assessment can be accessed through the area document and the evaluation form [39, 40].

Data gathering

A significant contribution of this work is the creation of the database, because there is no open site that makes the data available in an integrated way. It was necessary to extract the data from some platforms, which are presented below. As a result, the data was integrated and stored in a single source.

The Lattes, besides data on the professional performance of researchers and their publications, their academic affiliation is not always made clear. To find out about the researchers that belong to an institution and graduate program, 89,255 Brazilian researchers' records were gathered from the Sucupira [36]. The Sucupira platform is an important platform to collect information, perform analysis and assessments and be the reference base of the National Post-Graduation System (SNPG) in Brazil. The Sucupira platform makes available the information, processes and evaluation procedures that CAPES performs openly. From this platform, the records of 89,255 Brazilian researchers were extracted.

The data from the Sucupira platform allowed creating a list of graduate programs and their respective researchers. This list was filtered to include only researchers in Computer Science and academic-oriented graduate programs, thus leaving out all other researchers from other areas or professional-oriented programs. A list of 1,644 Computer Science researchers' full names and their affiliations was produced with this process. Using this approach, even if an international collaboration has been carried out, if 2 or more researchers are linked to the programs, it will be considered in our analyses.

The scope of this work is the Brazilian Computer Science programs, however, the method applied in it can be extended and applied to other areas and researchers in further works.

Data parsing and storage

The data parsing started by obtaining the academic resumes of the researchers from the Lattes platform, which were identified in the list produced in the previous section. More specifically, the resumes were downloaded from Lattes platform in an XML format. The XML data were converted into structured data (DBMS). As a result, a SQL-based database management system was produced.

Regarding the publications available on the Lattes platform, there are several types, such as articles published in journals, full papers published in conference proceedings, abstracts published in conference proceedings, books, book chapters, among others. In this work, publications in journals and full papers in conference proceedings were used to study and analyze research collaboration among researchers associated with Brazilian graduate programs in computer science (both types of publications are the most common tangible output of a research collaboration).

Therefore, the information about the list of researchers, the respective graduate programs to which they are linked, their institution and their intellectual production were stored and reviewed in a database. In this way, CAPES evaluates each of the programs and assigns them a grade for each time interval. This information was entered and indexed to the graduate programs for further analysis. The generated database, as well as the algorithms and filters applied to the data (step by step) are freely available at: https://github.com/alexjrns/datamining_lattes_computer_science.

Graph generation

One challenge in building a co-authorship network is to extract the data from a source and correctly attribute the publications to the respective authors. Names may contain errors, for example, names written with distinct characters, with abbreviations, without accents, or the existence of homonyms, leading to unreliable relationships [41].

To circumvent this problem, all researchers' names were normalized (i.e., names were transformed to lowercase, no accents and no punctuation marks). The correct identification of the researchers was carried out through the approximate matching of normalized names using the Levenshtein's distance [42, 43]. The comparison between the names of the authors was performed to remove or reduce the number of ambiguities, leading to improvement in the quality of collaborative relationships [19, 44]. The adopted approach comprises analyzing two strings A and B and returning the number of operations required to transform string A into a string B. If the number of operations is less than or equal to 2, it is understood that the strings match. Only full names are stored in the Lattes and Sucupira platforms, allowing us to reach a high percentage of perfect matches. In fact, 1529 (93%) of the 1644 researchers were identified directly without any ambiguities. Using the data from Lattes to filter the matched names to include only researchers belonging to the field of computer science allowed us to clear many of the ambiguities and, in the end, only 3% of all researchers were not identified automatically.

The academic collaboration networks were produced considering the researchers as nodes in this network. However, the networks were analyzed considering the respective graduate programs that the researchers are associated with. Thus, the researchers are represented by nodes, and the articles published in collaboration between two researchers represent an edge in the produced graph. Thus, both the collaboration within the graduate program (subgraph formed only by the researchers of each program), and the collaboration between researchers belonging to different programs (collaboration between subgraphs) were considered in this work.

Another relevant aspect is the analysis of the dynamics of each graduate program. This work considers three evaluative periods by CAPES between 2007-2009, 2010-2012 and 2013-2016 and their respective CAPES grades. The evaluation between 2017-2020 is ongoing and not yet available from CAPES. Therefore, the analysis of dynamics helps to identify which properties, in terms of academic collaborations, lead a program to receive the same evaluation or better/worse evaluations in different time windows.

Graph analysis

Complex network theory has been successfully applied in many areas, particularly within representation of networks of different types, such as biological systems [45–48], computer vision [49–52], the electric power grid [53], the Internet [54], subway systems [55], and neural networks [56], to cite but a few. Another area in which they are also applied is the representation of friendship networks or collaborations between individuals.

The complex network measurements can represent and characterize specific topological models [57]. Therefore, these measurements can be applied in the characterization of topological patterns in the networks. In this context, 42 complex networks measurements [58] were considered in this work, such as Number of Nodes, Number of Edges, Betweenness Centrality, Cluster Coefficient, Average Path Length, among others in order to explore and analyse the relationships of the topological patterns and the CAPES grades. Among the measures considered are measures of complex network structure (topology), analysis of vulnerability measures and measures that assess the position in which the researcher's name is among the authors of the publication.

In order to analyze the topological patterns of co-authorship networks, a feature matrix is generated by composing the complex networks measurements and the respective CAPES grades. More specifically, the measures were organized as feature vectors with size $n + 1$, where n is the total number of features, and CAPES grade (cg) was added in the last column referring to the analyzed program and the observed period of evaluation. Thus, feature vectors were produced for each graduate program and each period of time evaluation and its respective CAPES grade and arranged into a feature matrix M_f as follows:

$$M_{\mathrm{f}} = \begin{bmatrix} f_{1,1} & f_{2,1} & f_{3,1} & \cdots & f_{\mathrm{n},1} & cg_1 \\ f_{1,2} & f_{2,2} & f_{3,2} & \cdots & f_{\mathrm{n},2} & cg_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ f_{1,\mathrm{m}} & f_{2,\mathrm{m}} & f_{3,\mathrm{m}} & \cdots & f_{\mathrm{n},\mathrm{m}} & cg_{\mathrm{m}} \end{bmatrix}$$

Classification

From the feature matrix, classification algorithms applied to identify which features lead to an adequate classification and generalization considering the respective CAPES grade. The framework Weka [59] and the Random Forest (RF) algorithm [60] with default parameters were adopted. RF is a decision tree algorithm and allows a direct interpretation of its results by recovering the rules applied in the classification process. Thus, random forest allows the retrieval of the adopted features by the classifier, which was one goal of this work and the estimating the importance of these features. The 10-fold cross-validation was adopted as the validation method as described in [61]. As a result, it is possible to recover the importance of each feature in the classification of CAPES grade, i.e. it is possible to observe which feature has relevance for the correct classification.

The overview of this process are presented in Fig 2.

Author order

The order in which authors are listed in a publication can provide information that can be analysed and discussed. This order follows distinct patterns, depending on the research field or country. A straightforward way is the alphabetical sorting of the authors. However, this is not

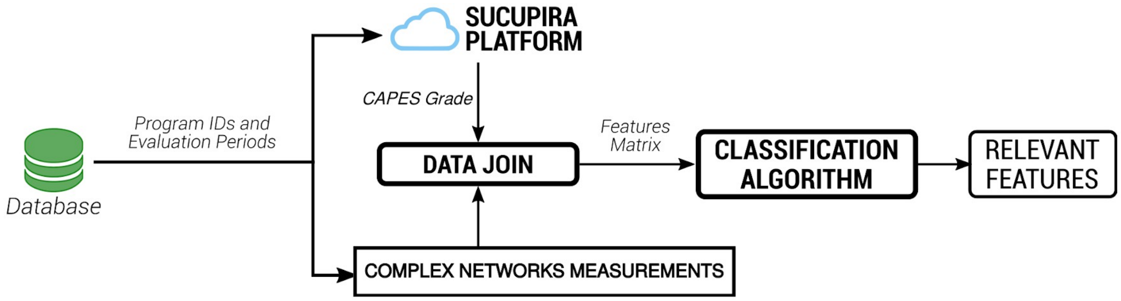

Fig 2. Process flow for topological metrics analysis. Note that the data saved in the database is used as a filter for downloading CAPES grade from the internet and identifies each program with its topological measurements. This data is gathered into a feature matrix and the algorithms performed.

https://doi.org/10.1371/journal.pone.0261200.g002

typical behavior in computer science, and in other research areas, in which authors are ordered according to their contribution [62–64]. First author is the one that used a more considerable effort, defined the materials, methods, and objectives of the work and realizes the final analysis of the results. Last author quoted commonly is the supervisor of the work and the project leader. The authors who are not in first or last authorship are co-authors who contributed to the work, but specifically in certain points. The analyses of these co-authorship orders can be used as qualitative measures of a group of authors' publications.

The analysis on researchers' performance considering the effects of seniority, their respective genders, and their geographic positions was performed [65]. More specifically, academic publications and citations were analyzed based on the Scopus, Web of Science and European Research Council (ERC) collaboration network, which covered 355 Life Science scholars in the period from 2007 to 2009. The authors considered 2 types of researchers, the first being those who were called juniors who are the researchers who are starting or consolidating their research team. They also considered the so-called senior researchers, who are those with a significant research history in the last 10 years. With the results, it was possible to observe that although all researchers had an increase in the size of the collaboration networks and the number of sub-communities during the time analyzed, the growth in juniors was greater. It was also possible to see that in both groups, the collaboration network was enlarged from the grant award to the researchers.

In order to investigate this context, this work proposes 3 qualitative indexes to evaluate the order of citation of the authors. The indexes are the *First Author Index*, which is the proportion of all the publications in which the authors of the graduate program are the first. The *Collaboration Index* is the proportion of publications in which the researcher is cited at the middle (neither first nor last) and the *Seniority Index* is the proportional of publications in which the researchers are the last author in a publication. The proposal is to compare these indexes with the CAPES grade and analyse its relation with the proposed indexes.

Results and discussion

The first step was to perform a normalization throughout the feature matrix in order to adjust each feature to the range from 0 to 1. It is an important issue so that classification algorithms did not suffer interference by the range of the feature values.

Another important issue is the unbalanced dataset. The produced dataset contains 171 computer science graduate programs in which 75 programs with CAPES grade 3, 58 programs with grade 4, 14 programs with grade 5, 9 programs with grade 6, and 15 programs with grade

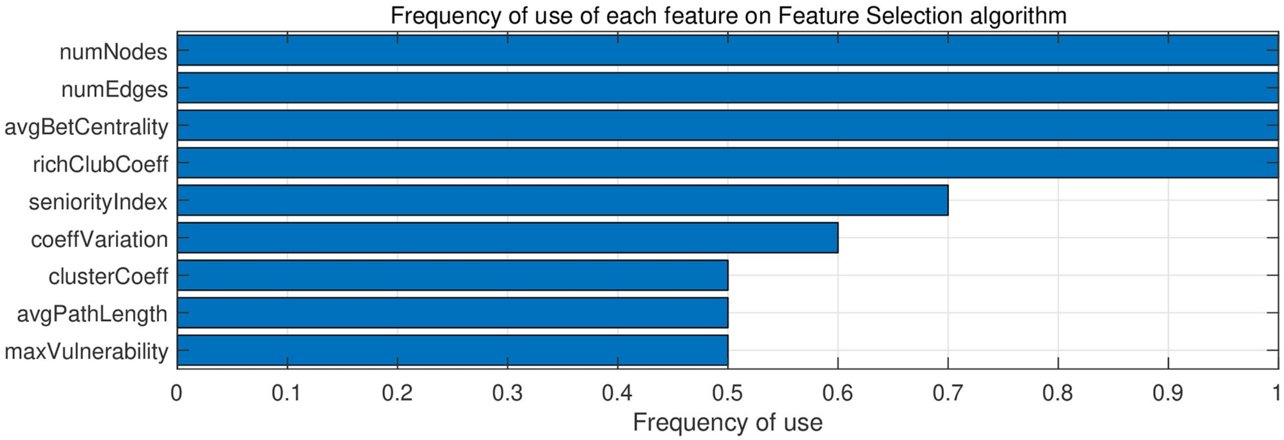

Fig 3. Importance of each feature for the classification of CAPES grade regarding the feature selection algorithm.

https://doi.org/10.1371/journal.pone.0261200.g003

7. Thus, the different number of samples per class (CAPES grade) can influence the classification algorithms. In this context, subsets of 15 graduate programs were created considering the graduate programs with grades 3 and 4 equally distributed. All experiments were performed for each subgroup, and the average of the subsets results was considered. It is important to note that the results refer to the last 3 evaluated periods.

Considering the re-scaled complex network measurements, and the balanced subsets of graduate programs, the feature selection algorithm was performed in order to analyse which features better describe the CAPES grade. The CfsSubSetEval feature selection and BestFirst as the search method available at Weka [59] with its default parameters adopted. The CfsSubSetEval [66] is a correlation-based feature selection, which evaluates the value of a subset of attributes by considering the individual predictive ability of each feature, along with redundancy between them. Thus, subsets of features that are highly correlated with the class, yet have low inter-correlation, are preferred. Thus, the results showed that some features have a high association with the CAPES grade. More specifically, 9 features have a significant correlation with the CAPES grade. Fig 3 presents the selected features with the percentage of times each feature was selected. An explanation about the selected features and its properties are discussed in the following.

- *Number of Nodes*: This measure deals with the number of nodes in each network, in this case, the number of researchers in each program. Thus, programs with the highest grade is the programs with the largest number of researchers, so this measure is consistent with the analysis, where the higher the number of nodes (researchers), the greater the CAPES grade;
- *Number of Edges*: This measure refers to the number of connections in the network, i.e. the number of publications among the researchers. Programs with larger CAPES grade are programs that have more works published in collaboration with their researchers. Therefore, the best-evaluated programs have greater internal collaboration compared to the others;
- *Average Betweenness Centrality*: This index deals with the nodes' centrality in a network because it analyses the nodes in the shortest path between two connected nodes. The higher the average of this measurement, the more researchers are taking part in shorter paths. More specifically, the betweenness centrality quantifies the relevance of a researcher in relation to co-authorships of the network, i.e. the more publications an author has in collaboration, the greater will be your betweenness. Thus, an author with a higher betweenness centrality

represents the greater participation in the publications of the program. As the aim of our work was to evaluate the graduate programs regarding the CAPES grade, it was adopted the average betweenness centrality [57] of each program;

- *Rich Club Coefficient*: This index measures the proportion in which the nodes with a significant number of connections (hubs) of a network are connected. This coefficient can evaluate the robustness of a network, because the higher the value, the more strongly connected, which shows that if one of these hubs is removed, the lower the impact on the network structure. In this work, this measure evaluates the tolerance to changes of a program if a researcher is randomly removed, programs with lower CAPES grade have a greater dependence on their researchers, in case a vital researcher is removed from the program, the structure of the program will be strongly affected, different from the best-evaluated programs, that have less individual dependence of researchers;
- *Seniority Index*: This measure is proposed in order to evaluate the percentage of publications in which the researcher has the last name in a publication. Thus, it is possible to qualify the publications of each researcher and therefore generalize this measure as the average of the Seniority Index of all researchers in the same graduate program. It is possible to notice that higher average seniority index refers to the graduate programs with higher CAPES grade.
- *Variation Coefficient*: The Coefficient of Variation measures the variation of a network; its mathematical equation results from the standard deviation of the values of the network divided by the average of these values. Regarding this work, this metric informs how much the adopted measures vary in each one of the graduate programs;
- *Cluster Coefficient*: This index shows a tendency in which the graph nodes have to group and form subsets. In social networks, these clusters are communities of individuals that share common features. In this work, a cluster is a group of researchers that have research projects in common, so the higher the value of this metric, the greater the number of publications among the researchers of this program (internal collaboration). Programs with higher CAPES grade have greater internal collaboration (more edges);
- *Average Path Length*: This measure performs the average path of the network as the average number of steps in the shortest paths for all nodes' pairs. The lower this indicator, the higher the efficiency in transporting information inside the network. In this work, this measure represents the average number of authors connecting an Author X to an Author Y, assuming that both do not have a direct connection, so the smaller the measure, the easier it is to connect 2 directly disconnected researchers. The results show that programs with lower evaluations of CAPES grade have a lower value in this measure. Better-evaluated programs have a higher complexity in the connection of their nodes. Since the number of nodes and the number of edges in better-evaluated programs are higher, there are more paths (co-authorship) and more nodes (researchers) in these networks. However, when this metric is considered relatively, dividing the value of the Average Path of the Network by the number of edges, it is possible to notice that this value is inverted. Therefore, it shows that although the networks have more researchers, they still efficient;
- *Swan Connectivity*: The Swan Connectivity is a measure of network vulnerability [67], which calculates the loss of connectivity when a vertex is removed from the network. As a result, measures the decrease in the number of relationships between each vertex of the network when one vertex or several are removed. In this work, this measure shows how vulnerable a network can be because when removing a researcher (vertex), the network loses connectivity, which is a particularity of graduate programs with lower CAPES grade.

Table 1. Area under the receiver operating characteristic (ROC) curve (AUC) for the Random Forest applied on feature matrix.

CAPES grade	AUC
3	0.828
4	0.701
5	0.741
6	0.838
7	0.929
Average	**0.787**

https://doi.org/10.1371/journal.pone.0261200.t001

The second experiment was performed in order to analyse the importance of features into the classification process. Weka [59] and its Random Forest classification algorithm with default parameters was performed by adopting the 10-fold cross-validation. A suitable way to assess the performance of classifiers algorithms is the area under the receiver operating characteristic (ROC) curve (AUC) [68]. Table 1 shows the classification results. It is possible to notice that the AUC was superior to 0.7 for all CAPES grade, achieving 0.828 for CAPES grade 3 and 0.929 for CAPES grade 7.

Fig 4 shows the percentage of times each feature was selected. It is possible to observe that the most relevant feature in this context is the number of isolated nodes, researchers that have no work in common with their colleagues in the graduate program, which was a particularity of the graduate programs with higher CAPES grade. The second important feature was researchers per publication, showing that is an important pattern that distinguish the graduate programs. The cluster coefficient, number of nodes, number of edges and average path length reinforce the importance of these features for the identification of CAPES grades.

Considering that graduate programs in Brazil evaluated with a CAPES grade of 3 or higher are recommended, which leads to five classes: 3, 4, 5, 6 and 7. In order to better investigate the behavior of the measures identified as important for the characterization of CAPES grades, Fig 5 shows the average of these measures considering the graduate programs classified by the same CAPES grade. It can be seen that all complex networks measurements have variations between CAPES grades. As expected, the number of researchers (numNodes), the number of isolated researchers (numIsolatedNodes) and the number of publications (numEdges) are important measures. However, it can be highlighted the average betweenness centrality

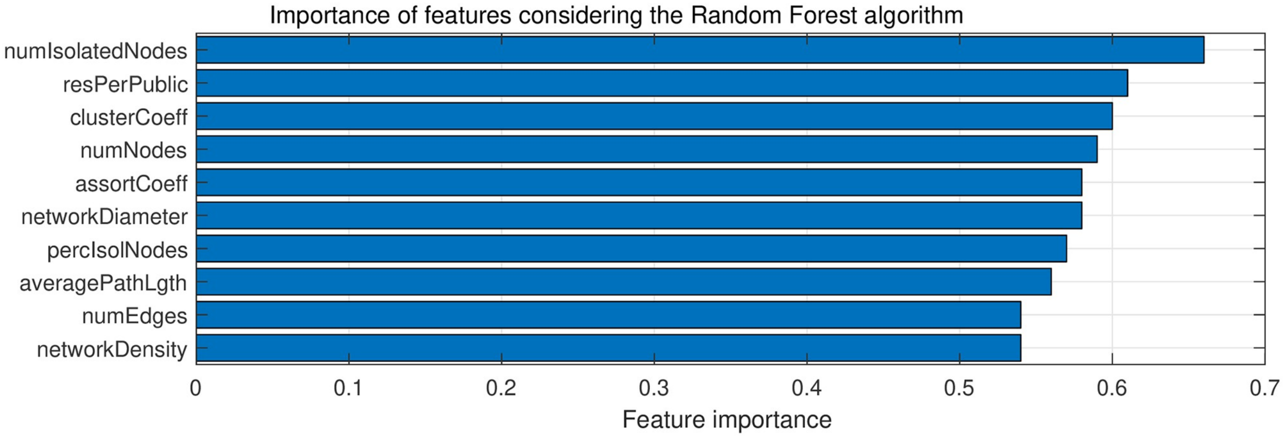

Fig 4. Importance of each feature for the classification of CAPES grade regarding the Random Forest algorithm.

https://doi.org/10.1371/journal.pone.0261200.g004

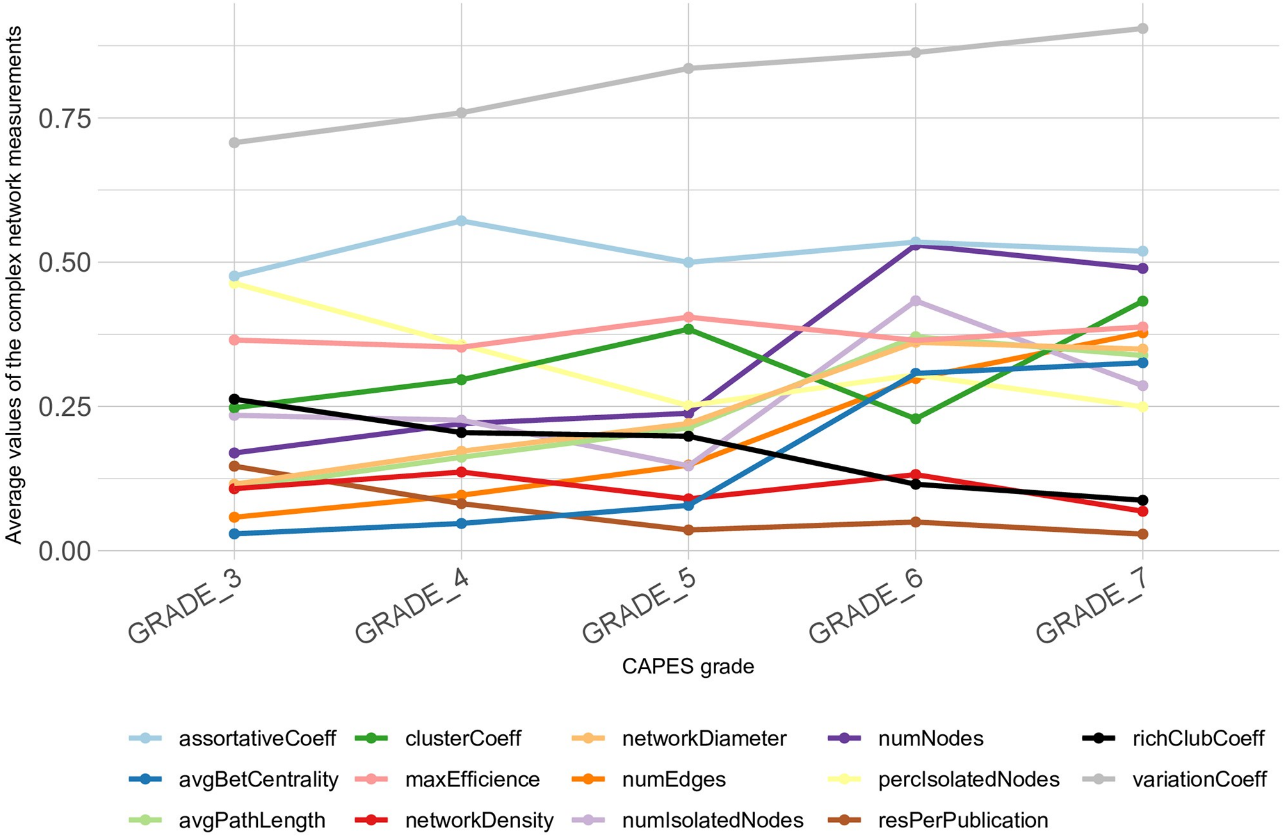

Fig 5. Average of the most important complex network measurements regarding the CAPES grades.

https://doi.org/10.1371/journal.pone.0261200.g005

(avgBetCentrality) which presents lower values for programs with CAPES grades 3, 4 and 5 and a significant increase in its value occurs for programs with CAPES grades 6 and 7. With similar behavior, but with less intensity, also stand out the average path length (avgPathLength) and network diameter (networkDiameter), which reinforces the increasing distance between researchers in programs with higher CAPES grades. The measure of rich club coefficient (richClubCoeff) can also be highlighted with decreasing behavior as the CAPES grades increase, showing that programs with higher CAPES grades have less individual dependence of researchers. On the other hand, the coefficient of variation (variationCoeff) shows a clear increasing behavior with the increase of CAPES grades. This measurement quantifies the heterogeneity of the vertices in the network, so the more different the vertices are, such as their degree and other measurements, the greater the variation, indicating that better evaluated graduate programs have a greater diversity of researcher profiles.

The similar analysis was also performed by the proposed indexes, thus the average values of each CAPES grade were performed for each proposed index. Fig 6 shows the average values for the 3 proposed indexes considering the CAPES grades. It is possible to observe that the seniority was pointed out by the feature selection algorithm, present a monotonically increasing behavior, i.e. as the CAPES grade of the graduate programs increases, the index also increases. The first author index presents the inverse behavior, with higher value for graduate programs with grade 3 and lower value for graduate programs with grade 7, indicating a pattern of research composition for the graduate programs. Regarding the collaboration index, it

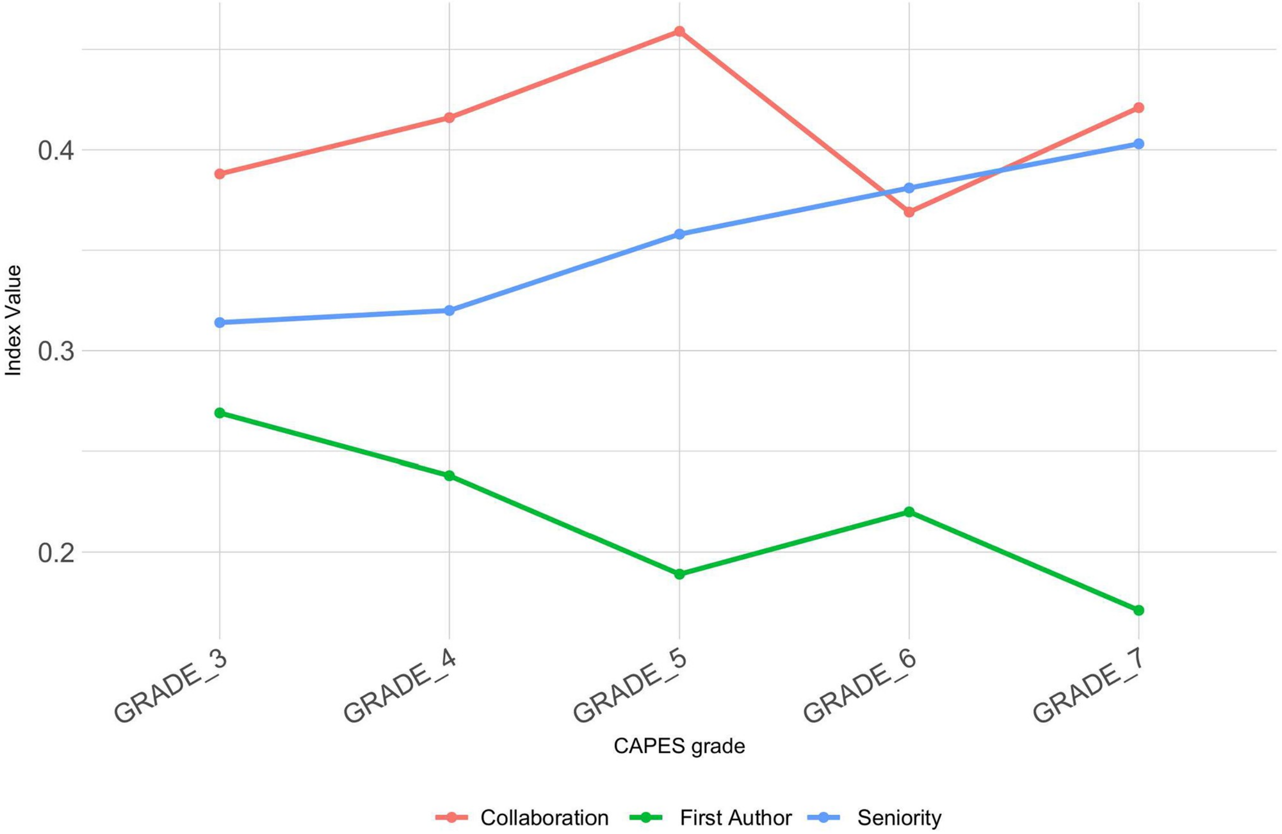

Fig 6. Average of the proposed indexes in graduate programs regarding their CAPES grades.

https://doi.org/10.1371/journal.pone.0261200.g006

is possible to notice that there is an increasing pattern among programs 3, 4 and 5, however a decrease of this index for programs with grades 6 and 7.

The random forest algorithm pointed the number of isolated nodes out as an important feature. In order to better investigate this feature, the average values of researchers per publication were adopted. The average values of the graduate programs for each CAPES grade are shown in Fig 7. It can be seen that there is a clear variation in this index between the programs, in which graduate programs with higher CAPES grade have a lower average number of researchers per publication. Thus, it can be noticed that programs with higher CAPES grades are more efficient than the programs with lower CAPES grades, regarding the platforms Lattes and Sucupira provide identification and publications from the Brazilian researchers and graduate programs. Therefore, international researchers are not part of the co-authorship networks adopted in this study.

Conclusions

This work proposes a complex network approach to analyse and characterize the co-authorships of Brazilian computer science graduate programs. Considering that the Brazilian graduate programs are evaluated relatively by CAPES within each area of knowledge for the attribution of the grade, the computer science graduate program co-authorship networks were analysed, identifying the most relevant network measurements to characterize these graduate programs regarding their quality by considering the CAPES grade. It was analysed 62 Brazilian

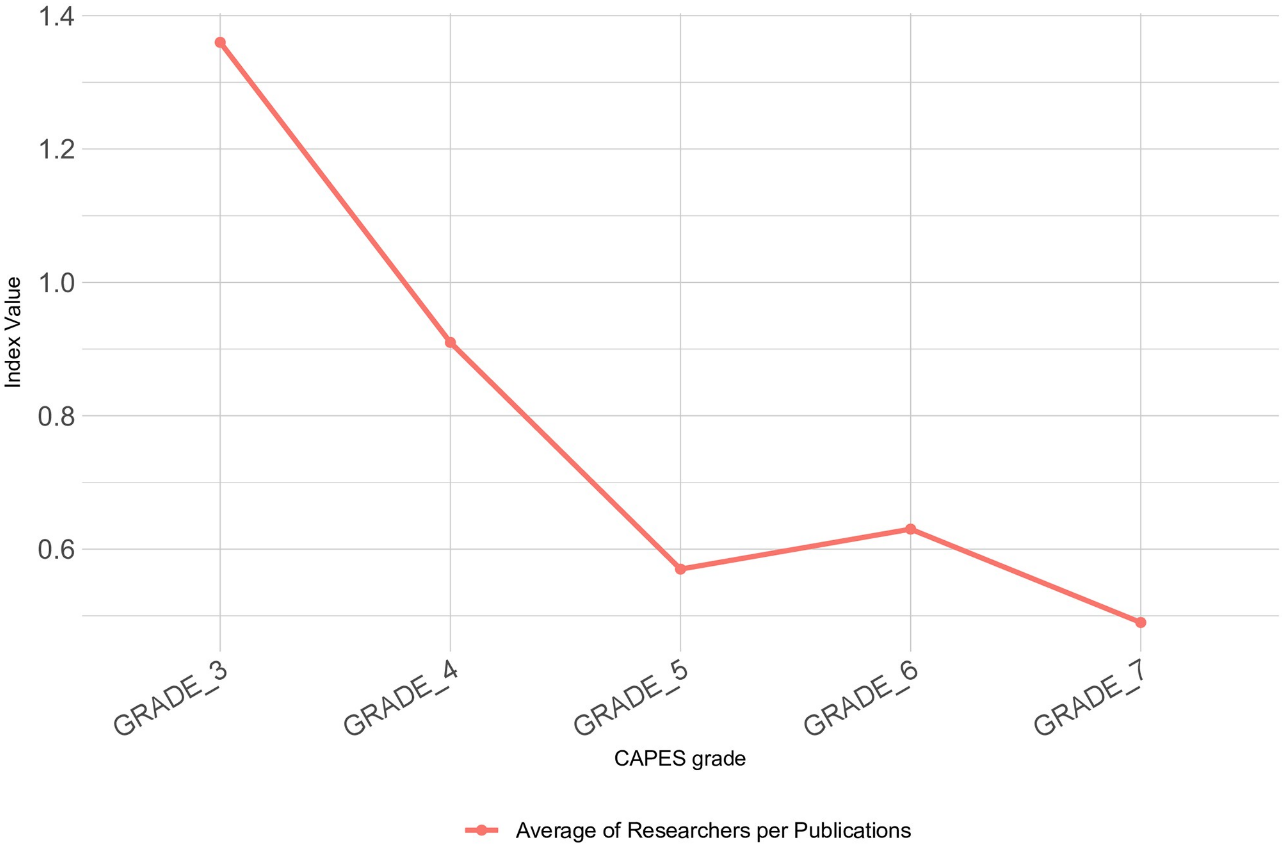

Fig 7. Average number of researchers per publication in graduate programs regarding their CAPES grades.

https://doi.org/10.1371/journal.pone.0261200.g007

graduate programs in computer science, with about 1,644 researchers observed in three CAPES evaluation periods from 2007 to 2016. A dataset was produced after the pre-processing the data from Lattes platform. Thus, the produced dataset allows the analyses of graduate programs by considering 42 complex networks measurements regarding their CAPES grades.

The adopted measurements that considering the size of the networks (graduate programs) were the most significant. Thus, the larger the program, either in the number of researchers and in the number of publications, higher are the CAPES grades. However, the higher number of researchers per program must be combined with more publication. Thus, the aim of this work is not only to point out the important features but also to explain how these features can act together in order to explain the CAPES grade.

The feature selection algorithm pointed the measures of centrality (importance) in the networks out. It was observed that better-evaluated graduate programs have more elements with higher centrality, so these programs have more researchers of greater influence. Vulnerability (Rich Club Coefficient and SWAN Connectivity) measures also yielded relevant results, with better-rated programs being less vulnerable than lower-rated ones, i.e. when a researcher is randomly removed, the program structure undergoes fewer changes, which is not the case with lower-rated programs that are more highly rated vulnerable.

The random forest classifier algorithm pointed the number of isolated nodes out as an important feature. Thus, it was possible to observe that graduate programs with higher CAPES grades are more productive because they have more publications per researcher, therefore, more efficient.

Three qualitative measures of collaborative evaluation among the researchers are proposed based on the author order of co-authorship regarding the publications: first, middle or last. The results for such analysis leads to interesting patterns, which graduate programs with lower CAPES grade have a higher first author index than the others. Programs with intermediate grades have a higher collaboration index than others, while the highest-rated programs have a higher seniority index than the others, with a monotonically increasing behavior as the programs evaluation increases.

In summary, this work points out some important patterns to be analysed that lead to the characterization of the graduate programs related to CAPES grade can bring information for the Brazilian computer science community to analyse and to adopt strategies that can lead to the improvement of these patterns and, improve the assessment of the graduate programs. Investigations concerning how the CAPES grade is related to other qualitative indexes available in the literature are needed and suggested as further work.

It is important to highlight that the measures raised in this work reveal the reality of these networks in which the scope of work. When evaluating other networks such as international collaborations or interaction between other programs, these measures may present different values, as in these cases they will be other scenarios and therefore future work can be carried out using the same approach as this one.

Author Contributions

Conceptualization: Jesús Pascual Mena-Chalco, Fabrício Martins Lopes.

Data curation: Alex Nunes da Silva, Junior, Matheus Montanini Breve.

Formal analysis: Matheus Montanini Breve.

Funding acquisition: Fabrício Martins Lopes.

Investigation: Jesús Pascual Mena-Chalco, Fabrício Martins Lopes.

Methodology: Alex Nunes da Silva, Junior, Matheus Montanini Breve, Jesús Pascual Mena-Chalco.

Project administration: Fabrício Martins Lopes.

Software: Alex Nunes da Silva, Junior, Matheus Montanini Breve.

Supervision: Jesús Pascual Mena-Chalco, Fabrício Martins Lopes.

Validation: Alex Nunes da Silva, Junior, Matheus Montanini Breve.

Writing – original draft: Alex Nunes da Silva, Junior, Matheus Montanini Breve, Jesús Pascual Mena-Chalco, Fabrício Martins Lopes.

Writing – review & editing: Jesús Pascual Mena-Chalco, Fabrício Martins Lopes.

References

1. Boissevain J. Network Analysis: A Reappraisal. Current Anthropology. 1979; 20(2):392–394. https://doi.org/10.1086/202277
2. Freeman LC. The Development of Social Network Analysis: A Study in the Sociology of Science. Empirical Press; 2004.
3. Barnes JA. Graph Theory and Social Networks: A Technical Comment on Connectedness and Connectivity. Sociology. 1969; 3(2):215–232. https://doi.org/10.1177/003803856900300205
4. Watts DJ, Strogatz SH. Collective dynamics of'small-world' networks. Nature. 1998; 393(6684):440–442. https://doi.org/10.1038/30918 PMID: 9623998

5. Barabási AL, Albert R. Emergence of scaling in random networks. Science. 1999; 286(5439):509–512. https://doi.org/10.1126/science.286.5439.509 PMID: 10521342

6. Albert R, Barabási AL. Statistical mechanics of complex networks. Rev Mod Phys. 2002; 74:47–97. https://doi.org/10.1103/RevModPhys.74.47

7. Newman MEJ. The Structure and Function of Complex Networks. SIAM Review. 2003; 45(2):167–256. https://doi.org/10.1137/S003614450342480

8. Katz JS, Martin BR. What is research collaboration? Research Policy. 1997; 26(1):1–18.

9. Kong X, Shi Y, Yu S, Liu J, Xia F. Academic social networks: Modeling, analysis, mining and applications. Journal of Network and Computer Applications. 2019; 132:86–103. https://doi.org/10.1016/j.jnca.2019.01.029

10. Garfield E. From the science of science to Scientometrics visualizing the history of science with HistCite software. Journal of Informetrics. 2009; 3(3):173–179. https://doi.org/10.1016/j.joi.2009.03.009

11. Lužar B, Levnajić Z, Povh J, Perc M. Community structure and the evolution of interdisciplinarity in Slovenia's scientific collaboration network. PLOS ONE. 2014; 9(4):e94429. https://doi.org/10.1371/journal.pone.0094429 PMID: 24728345

12. Glänzel W, Leta J, Thijs B. Science in Brazil. Part 1: A macro-level comparative study. Scientometrics. 2006; 67(1):67–86. https://doi.org/10.1007/s11192-006-0055-7

13. Mena-Chalco JP, Cesar RM Junior. scriptLattes: an open-source knowledge extraction system from the Lattes platform. Journal of the Brazilian Computer Society. 2009; 15(4):31–39. https://doi.org/10.1590/S0104-65002009000400004

14. Mena-Chalco JP, Digiampietri LA, Lopes FM, Cesar RM Junior. Brazilian bibliometric coauthorship networks. Journal of the Association for Information Science and Technology. 2014; 65(7):1424–1445. https://doi.org/10.1002/asi.23010

15. Salinero MC, Michalski F. Implications of scientific collaboration networks on studies of aquatic vertebrates in the Brazilian Amazon. PloS one. 2016; 11(6):e0158413. https://doi.org/10.1371/journal.pone.0158413 PMID: 27352247

16. Galuppo Azevedo F, Murai F. Evaluating the state-of-the-art in mapping research spaces: A Brazilian case study. PLOS ONE. 2021; 16(3):1–27. https://doi.org/10.1371/journal.pone.0248724 PMID: 33735233

17. Leifeld P, Wankmüller S, Berger VT, Ingold K, Steiner C. Collaboration patterns in the German political science co-authorship network. PloS one. 2017; 12(4):e0174671. https://doi.org/10.1371/journal.pone.0174671 PMID: 28388621

18. Çavuşoğlu A, İlker Türker. Scientific collaboration network of Turkey. Chaos, Solitons & Fractals. 2013; 57:9–18. https://doi.org/10.1016/j.chaos.2013.07.022

19. Digiampietri LA, Mena-Chalco JP, Vaz de Melo POS, Malheiro APR, Meira DNO, Franco LF, et al. BraX-Ray: An X-Ray of the Brazilian Computer Science Graduate Programs. PLOS ONE. 2014; 9(4):1–12. https://doi.org/10.1371/journal.pone.0094541 PMID: 24728179

20. Cerinšek M, Batagelj V. Network analysis of Zentralblatt MATH data. Scientometrics. 2015; 102(1):977–1001. https://doi.org/10.1007/s11192-014-1419-z

21. Newman MEJ. Scientific collaboration networks. I. Network construction and fundamental results. Physical Review E. 2001; 64(1):016131. https://doi.org/10.1103/PhysRevE.64.016131 PMID: 11461355

22. Newman MEJ. Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. Physical Review E. 2001; 64(1):7. https://doi.org/10.1103/PhysRevE.64.016132 PMID: 11461356

23. Gui Q, Liu C, Du D. Globalization of science and international scientific collaboration: A network perspective. Geoforum. 2019; 105:1–12. https://doi.org/10.1016/j.geoforum.2019.06.017

24. Cantner U, Rake B. International research networks in pharmaceuticals: Structure and dynamics. Research Policy. 2014; 43(2):333–348. https://doi.org/10.1016/j.respol.2013.10.016

25. Abbasi A, Jaafari A. Research impact and scholars' geographical diversity. Journal of Informetrics. 2013; 7(3):683–692. https://doi.org/10.1016/j.joi.2013.04.004

26. Carson L, Bartneck C, Voges K. Over-competitiveness in academia: A literature review. Disruptive science and technology. 2013; 1(4):183–190. https://doi.org/10.1089/dst.2013.0013

27. Hicks D, Wouters P, Waltman L, de Rijcke S, Rafols I. Bibliometrics: The Leiden Manifesto for research metrics. Nature. 2015; 520(7548):429–431. https://doi.org/10.1038/520429a PMID: 25903611

28. Jappe A, Pithan D, Heinze T. Does bibliometric research confer legitimacy to research assessment practice? A sociological study of reputational control, 1972-2016. PLoS ONE. 2018; 13. https://doi.org/10.1371/journal.pone.0199031 PMID: 29902239

29. Parish AJ, Boyack KW, Ioannidis JPA. Dynamics of co-authorship and productivity across different fields of scientific research. PLOS ONE. 2018; 13(1):e0189742. https://doi.org/10.1371/journal.pone.0189742 PMID: 29320509

30. Levitt JM, Thelwall M. Long term productivity and collaboration in information science. Scientometrics. 2016; 108(3):1103–1117. https://doi.org/10.1007/s11192-016-2061-8

31. Uddin S, Hossain L, Abbasi A, Rasmussen K. Trend and efficiency analysis of co-authorship network. Scientometrics. 2011; 90(2):687–699. https://doi.org/10.1007/s11192-011-0511-x

32. Li EY, Liao CH, Yen HR. Co-authorship networks and research impact: A social capital perspective. Research Policy. 2013; 42(9):1515–1530. https://doi.org/10.1016/j.respol.2013.06.012

33. Linden R, Barbosa L, Digiampietri L. "Brazilian style science"—an analysis of the difference between Brazilian and international Computer Science departments and graduate programs using social networks analysis and bibliometrics. Social Network Analysis and Mining. 2017; 7:44. https://doi.org/10.1007/s13278-017-0463-0

34. Lopes GR, Moro MM, Da Silva R, Barbosa EM, de Oliveira JPM. Ranking Strategy for Graduate Programs Evaluation. In: 7th International Conference on Information Technology and Applications (ICITA 2011); 2011. p. 59–64.

35. Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). Plataforma Lattes; 2021. Available from: http://lattes.cnpq.br/.

36. Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES). Plataforma Sucupira; 2021. Available from: https://sucupira.capes.gov.br/sucupira/.

37. Coordenação de Aperfeicoamento de Pessoal de Nível Superior (CAPES). Sobre a Avaliação; 2021. Available from: https://www.gov.br/capes/pt-br/acesso-a-informacao/acoes-e-programas/avaliacao.

38. CAPES. Sobre as áreas de avaliação; 2021. Available from: https://www.gov.br/capes/pt-br/acesso-a-informacao/acoes-e-programas/avaliacao/sobre-a-avaliacao/areas-avaliacao/sobre-as-areas-de-avaliacao/sobre-as-areas-de-avaliacao.

39. CAPES. Ficha de Avaliação; 2019. Available from: http://www.capes.gov.br/pt/relatorios-.

40. CAPES. Ciência da Computação; 2020. Available from: https://www.capes.gov.br/avaliacao/sobre-as-areas-de-avaliacao/75-dav/caa3/4656-ciencia-da-computacao.

41. Savić M, Ivanović M, Jain LC. Complex Networks in Software, Knowledge, and Social Systems. vol. 148. Springer; 2019.

42. Levenshtein VI. Binary codes capable of correcting deletions, insertions and reversals. Soviet Physics Doklady. 1966; 10(8):707–710.

43. Doan A, Halevy A, Ives Z. Principles of data integration. vol. 1 of 1. 1st ed. Elsevier; 2012.

44. Digiampietri LA, Mena-Chalco J, Silva GS, Oliveira L, Malheiro A, Meira D. Dinâmica das Relações de Coautoria nos Programas de pós-Graduação em Computação no Brasil. Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2012)—Anais do XXXII Congresso da Sociedade Brasileira de Computação (CSBC 2012). 2012; p. 12.

45. Lopes FM, Cesar-Jr RM, Costa LdF. Gene Expression Complex Networks: Synthesis, Identification, and Analysis. Journal of Computational Biology. 2011; 18(10):1353–1367. https://doi.org/10.1089/cmb.2010.0118 PMID: 21548810

46. Lopes FM, Cesar DC Jr, Barrera J, Cesar RM Jr. A feature selection technique for inference of graphs from their known topological properties: Revealing scale-free gene regulatory networks. Information Sciences. 2014; 272(0):1–15. https://doi.org/10.1016/j.ins.2014.02.096

47. Ito EA, Katahira I, Vicente FF, Pereira LP, Lopes FM. BASiNET—BiologicAl Sequences NETwork: a case study on coding and non-coding RNAs identification. Nucleic Acids Research. 2018; p. gky462. https://doi.org/10.1093/nar/gky462 PMID: 29873784

48. Breve MM, Lopes FM. A Simplified Complex Network-Based Approach to mRNA and ncRNA Transcript Classification. In: Advances in Bioinformatics and Computational Biology. Cham: Springer International Publishing; 2020. p. 192–203.

49. Backes AR, Casanova D, Bruno OM. Texture analysis and classification: A complex network-based approach. Information Sciences. 2013; 219(0):168–180. https://doi.org/10.1016/j.ins.2012.07.003

50. de Lima GVL, Castilho TR, Bugatti PH, Saito PM, Lopes FM. A Complex Network-Based Approach to the Analysis and Classification of Images. In: Pardo A, Kittler J, editors. Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. vol. 9423 of Lecture Notes in Computer Science. Springer International Publishing; 2015. p. 322–330. Available from: http://dx.doi.org/10.1007/978-3-319-25751-8_39.

51. Piotto JGS, Lopes FM. Combining SURF descriptor and complex networks for face recognition. In: 2016 9th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI); 2016. p. 275–279.

52. de Lima GVL, Saito PTM, Lopes FM, Bugatti PH. Classification of texture based on Bag-of-Visual-Words through complex networks. Expert Systems with Applications. 2019; 133:215–224. https://doi.org/10.1016/j.eswa.2019.05.021

53. Albert R, Albert I, Nakarado GL. Structural vulnerability of the North American power grid. Physical review E. 2004; 69(2):025103. https://doi.org/10.1103/PhysRevE.69.025103 PMID: 14995510

54. Maslov S, Sneppen K, Zaliznyak A. Detection of topological patterns in complex networks: correlation profile of the internet. Physica A: Statistical Mechanics and its Applications. 2004; 333:529–540. https://doi.org/10.1016/j.physa.2003.06.002

55. Angeloudis P, Fisk D. Large subway systems as complex networks. Physica A: Statistical Mechanics and its Applications. 2006; 367:553–558. https://doi.org/10.1016/j.physa.2005.11.007

56. Kötter R, Stephan KE. Network participation indices: characterizing component roles for information processing in neural networks. Neural Networks. 2003; 16(9):1261–1275. https://doi.org/10.1016/j.neunet.2003.06.002 PMID: 14622883

57. Boccaletti S, Latora V, Moreno Y, Chavez M, Hwang DU. Complex networks: Structure and dynamics. Physics reports. 2006; 424(4):175–308. https://doi.org/10.1016/j.physrep.2005.10.009

58. Costa LdF, Rodrigues FA, Travieso G, Villas-Boas PR. Characterization of complex networks: a survey of measurements. Adv in Phys. 2007; 56(1):167–242. https://doi.org/10.1080/00018730601170527

59. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA Data Mining Software: An Update. SIGKDD Explor Newsl. 2009; 11(1):10–18. https://doi.org/10.1145/1656274.1656278

60. Breiman L. Random Forests. Machine Learning. 2001; 45(1):5–32. https://doi.org/10.1023/A:1010933404324

61. Duda RO, Hart PE, Stork DG. Pattern Classification. 2nd ed. John Wiley & Sons, 2012; 2012.

62. Costa MM, Gatz M. Determination of Authorship Credit in Published Dissertations. Psychological Science. 1992; 3(6):354–357. https://doi.org/10.1111/j.1467-9280.1992.tb00046.x

63. Venkatraman V. Conventions of Scientific Authorship. Science. 2010;.

64. Biagioli M, Galison P. Scientific authorship: Credit and intellectual property in science. Routledge; 2014.

65. Pina DG, Barać L, Buljan I, Grimaldo F, Marušić A. Effects of seniority, gender and geography on the bibliometric output and collaboration networks of European Research Council (ERC) grant recipients. PLoS ONE. 2019; 14(2):1–16. https://doi.org/10.1371/journal.pone.0212286 PMID: 30763395

66. Hall MA. Correlation-based Feature Subset Selection for Machine Learning. University of Waikato. Hamilton, New Zealand; 1999.

67. Lhomme S. NetSwan: Network Strengths and Weaknesses Analysis; 2015. Available from: https://github.com/cran/NetSwan.

68. Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern recognition. 1997; 30(7):1145–1159. https://doi.org/10.1016/S0031-3203(96)00142-2