symmetry

MDPI

Article

Symmetry in Scientific Collaboration Networks: A Study Using Temporal Graph Data Science and Scientometrics

Breno Santana Santos [1,2,*], Ivanovitch Silva [1,*] and Daniel G. Costa [3]

[1] Postgraduate Program in Electrical and Computer Engineering, Federal University of Rio Grande do Norte, Natal 59078-970, Brazil
[2] Information System Department, Federal University of Sergipe, Itabaiana 49506-036, Brazil
[3] INEGI, Faculty of Engineering, University of Porto, 4200-465 Porto, Portugal
* Correspondence: breno.santos.038@ufrn.edu.br or breno1005@hotmail.com (B.S.S.); ivanovitch.silva@ufrn.br (I.S.)

Abstract: This article proposes a novel approach that leverages graph theory, machine learning, and graph embedding to evaluate research groups comprehensively. Assessing the performance and impact of research groups is crucial for funding agencies and research institutions, but many traditional methods often fail to capture the complex relationships between the evaluated elements. In this sense, our methodology transforms publication data into graph structures, allowing the visualization and quantification of relationships between researchers, publications, and institutions. By incorporating symmetry properties, we offer a more in-depth evaluation of research groups cohesiveness and structure over time. This temporal evaluation methodology bridges the gap between unstructured scientometrics networks and the evaluation process, making it a valuable tool for decision-making procedures. A case study is defined to demonstrate the potential to provide valuable insights into the dynamics and limitations of research groups, which ultimately reinforces the feasibility of the proposed approach when supporting decision making for funding agencies and research institutions.

Keywords: graph data science; symmetry properties; machine learning; graph embedding; temporal analysis; scientometrics

check for updates

Citation: Santos, B.S.; Silva, I.; Costa, D.G. Symmetry in Scientific Collaboration Networks: A Study using Temporal Graph Data Science and Scientometrics. *Symmetry* **2023**, *15*, 601. https://doi.org/10.3390/sym15030601

Academic Editor: Domenico Labbate

Received: 9 February 2023
Revised: 22 February 2023
Accepted: 23 February 2023
Published: 27 February 2023

1. Introduction

Roughly speaking, science can be understood as the set of theories and methods, rigorously and systematically proven, contained in the so-called state of the art. In most cases, knowledge is generated by researchers, either alone or by their collaboration, from the addition of new concepts to a scientific field [1]. Since science making is a global activity performed by a great and highly diverse number of researchers, it is not surprising that several parameters associated with it have been the object of investigation in multiple ways.

It is quite common, during scientific activities, to have the creation of research groups composed of members that are proactively collaborating to advance the knowledge of science. Doing so, they intend to have greater chances to achieve relevant results, to increase citations, and to obtain better opportunities for research funding [1–3].

Considering the overall research macro-system, funding agencies and research institutions have used bibliometric/scientometric indicators and other collaboration metrics to assess the impact, relevance and cohesion of research groups. This is required to support the selection of the best candidates when granting funds [2,4,5], although this process can be complex and biased. Among the existing approaches to evaluate research groups, there are those that use the h-index and its derivatives due to its popularity and easy understanding, but there are some controversies as to its effectiveness in evaluating scientific entities [4–7]. Regarding cohesion analysis, network science techniques and metrics are commonly used because they capture well the various aspects of collaboration between scientific entities [2,8,9]. Most studies have focused on count-based metrics (e.g., h-index and its derivatives) for evaluating research groups [5–7], as well as cohesion analysis [2,10,11]. However, such works have not properly addressed the temporal aspects associated to the aforementioned research-related elements.

The combination of scientometrics and graph symmetry plays a crucial role in evaluating the evolution of research groups over time. While scientometrics measures science, technology, and innovation, graph symmetry analyzes the symmetry in graphs or networks. By merging these two fields of study, researchers can better develop innovative solutions to understand scientific networks' evolution and communities. For example, scientometric techniques can analyze the distribution of scientific publications and citations, while graph symmetry can study collaboration and communication patterns among scientists. Integrating scientometrics and graph symmetry offers researchers a comprehensive understanding of the science and technology landscape, providing valuable insights into the structure and evolution of research groups. In this work, we highlight a novel approach that evaluates and analyzes the cohesion of research groups while considering the time factors related to the analyzed publications.

As highlighted in [12,13], a systematic, reproducible and transparent approach is needed to extract knowledge from sources of scientific production, making it possible to identify the dynamics and/or growth of a given area or to support strategic decisions related to research entities (e.g., evaluations and analysis of cohesion of research groups). Thus, as illustrated in Figure 1, the proposed approach analyzes scientific productions in line with that problem since complex networks will be modeled from raw and unstructured data. Additionally, the more specific the analyzed scientific entity is (e.g., affiliations or researchers), the more insights and relevant information will be extracted from the combination of complex network analysis (CNA) and machine learning mechanisms.

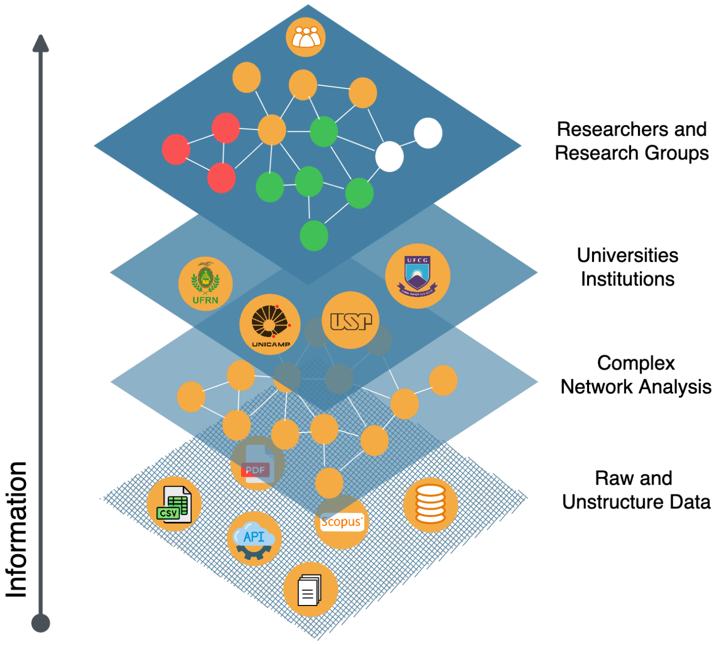


Figure 1. Proposed approach for scientific production analysis.

In order to improve the process of evaluation and cohesion analysis of research groups, a methodology was developed to extract patterns and knowledge from scientific publications of research groups, combining machine learning, complex network analysis and graph embeddings techniques. Furthermore, the proposed approach was experimentally validated through a case study involving a graduate program in electrical and computer engineering in a Brazilian university.

The expected results will demonstrate the feasibility of the proposed methodology indicating the main researchers, temporal behavior of research group's scientific collaboration, the alignment between scholars' research focus, and their patterns of collaboration, as well as the temporal cohesion of the research group. The presented approach has the potential to instrument and expand strategic and proactive decisions of research entities from the insights generated for the evaluation and analysis of the cohesion of research groups.

Thus, the significance of symmetry in scientific collaboration networks is highlighted through the threefold contributions of this work:

- A new methodology to perform scientific production analysis that considers symmetry in collaboration networks for evaluating and analyzing the cohesion of research groups;
- A temporal and experimental process that incorporates symmetry in network metrics and embedding features to cluster and investigate data, providing a more comprehensive analysis;
- A practical analysis on a dataset from a Brazilian graduate program in computer and electrical engineering, demonstrating the usefulness of considering symmetry in scientific collaboration networks in evaluating the scientific production of research groups.

The methodology provides a novel approach to study the evolution of research groups over time and offers valuable insights into the structure and cohesion of scientific communities. The practical analysis results demonstrate the proposed methodology's effectiveness in evaluating and analyzing the scientific production of research groups.

The remainder of this article is organized as follows. Section 2 describes some related works. Section 3 describes the proposed methodology to support the process of evaluation and cohesion analysis of research groups. In Section 4, we detail the experimental evaluation of our approach. Section 5 discusses the obtained results. Section 6 details the threats to the validity of our study. Finally, in Section 7, the conclusions are presented.

2. Related Works

The extensive use of graph theory to model data structures has been widely studied in the literature. However, recent applications, such as database structuring and knowledge extraction, bring new and exciting challenges. For instance, recent works have demonstrated how concept networks could be used to understand the growth of scientific knowledge, revealing non-randomly organized networks that expand outward and inward [14]. Furthermore, these works provided a mathematical formulation of historical data that can guide scientific progress for individuals and funding agencies while identifying novel contributions by underrepresented groups. In another example, a recent paper provided a bibliometric overview of graph neural networks (GNNs), analyzing publication trends, identifying impactful researchers and institutions, and highlighting exciting research directions, such as the explainability of GNN models [15]. These studies showcase the enormous potential of graph theory and graph-based approaches for modeling complex data structures and extracting valuable insights from them. As such, graph theory has become a powerful and versatile tool for numerous fields and will undoubtedly continue to play an essential role in modeling and understanding data in the future.

Within this perspective, graph-based approaches have shown great potential in scientometric evaluations, enabling the modeling and analysis of complex relationships between researchers, publications, and institutions. Recent works have demonstrated that such approaches can provide a more comprehensive and nuanced evaluation of research groups, their dynamics, and their impact. Many studies now use graphs for scientometric applications, contributing to a better understanding of various research entities, including researchers, groups, institutions, departments, and universities. These studies have focused on specific aspects, disciplines, or contexts and have significantly contributed to the field. In this section, we review some of the most relevant works that have proposed methodologies for the analysis and evaluation of research groups using graph-based approaches, including [2,4–7,10,11].

In [2], based on articles about rare diseases published between 2000 and 2013, the authors conducted a cohesion analysis of the CIBERER (Biomedical Research Networking Centers) research groups. They computed the global and local measures after modeling the collaboration networks for each year. Their conclusions after the case study indicated that the global network metrics were steady during the period, and the local and subgroup metrics suggested that the research groups were becoming more cohesive. Another interesting result was establishing a strongly connected component and reducing the number of communities. Moreover, assortativity measurements revealed that, following an early phase in which subject affinity and a common geographical location aided group cooperation, the collaboration was ultimately driven by other factors and complementarities.

Next, using various Hirsch-based measures, ref. [4] presented a systematic strategy for comparing academic research groups within the same topic. In addition, the authors may additionally represent the groups' bibliometric standing in the scholarly community. They only conducted a specialized investigation of Italian researchers in Production Technology and Manufacturing Systems. Following the conclusion of their study case, the results indicated that their systematic strategy was validated by empirical data and could be expanded to research groups linked with other scientific areas.

In addition, ref. [10] discovered and represented scholar cooperation patterns using ego nets embedding-based strategy and validated whether the scholars' research performance was associated with their cooperation practices. Additionally, they categorized scholars based on their cooperation approaches and compared the scientometric and CNA measures of the academics for each cluster discovered using their method. Moreover, the results of the investigations revealed that the proposed strategy was insufficient to confirm the correlation hypothesis. However, the patterns of collaboration may indicate the research approaches of experts.

In the sequence, ref. [6] introduced the aH-index as a new author-assessment metric and provided intuitive insights into its characteristics and interpretations. They examined communities of authors who cited publishing scientists using the theories underpinning the h-index. They also tested their suggested metric on scholars with high h-index who work in computer science—contained in AMiner Network (https://www.aminer.cn/aminernetwork, accessed on 20 February 2023)—and exhibited their measure's qualities. In addition, the findings of the experiments suggested that the aH-index might supplement the h-index with a new type of information representing the effectiveness of a researcher's citation response. The aH-index may give a more effective assessment of scientific activities regarding citation response.

In [7], the authors developed a new metric called the MZE-index was normalized by the number of manuscripts and ranged from -1 to 1. This measure allowed for comparing the relative position of a research group, institution, or author to those of his/her peer groups. Moreover, the authors also compared the MZE-index to the h-index of countries. The results of the experiments revealed that the MZE-index identified differences among countries that are above or below the average. Furthermore, the manuscripts of countries with a positive MZE index were more relevant or visible than those from other countries. Their approach also allowed for a less skewed comparison of researchers, journals, universities, or countries based on any given combination of the h-index and scientific output.

The work in [5] focused on the conventional h-index to highlight the issue of group h-index to overall research-quality assessment in a straightforward approach. They proved that randomly rearranging authentic scientometric data (changing the number of citations) among institutions of varied sizes did not influence their departmental h-index, while retaining the volume of their scientific manuscripts. In other words, their findings revealed that the relative position in evaluations based on the group h-index was driven not just by the quality/impact of specific research outputs but also by the production volume. In addition, the conventional departmental h-index was questionable as a foundation for the comparison of research groups, institutions, or journals.

In [11], the authors conducted a methodical assessment of the important university-level ResearchGate (RG) measures, including overall RG score, number of manuscripts, number of associated profiles, and Academic Ranking of World Universities (ARWU). Additionally, they used the ARWU rating system as a basis for comparison. Ten different ranks were created and compared to the baseline rating. Thus, the results of the experiments highlighted the potential of utilizing CNA measures to evaluate universities as a supplementary way of research unit evaluation. The distance between ARWU universities, determined by RG measures, is greater for the top of the list.

Finally, in Table 1, a comparative overview of the aspects of the studies mentioned above is provided. This allows visualizing their fundamental similarities and differences, with each row representing a different manuscript and each of the four columns identifying themes (or characteristics) and how they were addressed in a specific study. The analyzed features are analysis focus, metrics, temporal, and DVC (data version control).

Table 1. Summary of related works.

Study \ Features	Analysis Focus	Metrics	Temporal	DVC
[2]	Cohesion Analysis of groups	CNA metrics	Yearly	No
[4]	Evaluation of groups	h-index and its derivatives	No	No
[6]	Evaluation of groups	h-index and its derivatives	No	No
[10]	Evaluation of groups	Embeddings	No	No
[7]	Evaluation of groups	h-index and MZE-index	No	No
[5]	Evaluation of groups	h-index and its derivatives	No	No
[11]	Evaluation of groups	CNA metrics	No	No
Proposed approach	Evaluation and Cohesion Analysis of groups	CNA metrics and Embeddings	Yearly	Yes

All of the studies in Table 1 have different features, and none of them cover all of the predefined categories. Indeed, as mentioned earlier, there is a preponderance of studies focused on evaluating research groups in the literature (through the analysis focus feature), and those focused on cohesion analysis are still underexplored, in contrast to the suggested study, which considers both scenarios. Similarly, the metrics feature reveals that most studies investigate the count-based measures for evaluating research teams and CNA indicators for cohesion analysis. In addition, few recent studies have employed graph embedding techniques for any of these types of assessments. In terms of the temporal characteristic, it is worth noting that few works investigate the temporal component of their analyses. When they do, they mostly utilize the isolated annual strategy, as opposed to this work, which employs the isolated, windowed, and cumulative annual approaches. Finally, for the last comparison feature, no other work has employed DVC to automate the analytic process and make it auditable and reproducible.

The significance and relationship between each of the exhibited studies and our suggested methodology were emphasized. The preceding analysis identified some areas that remain to be explored, which are addressed in this article.

3. Proposed Methodology

The proposed method to extract scientometric knowledge for evaluation and cohesion analysis of research groups is presented in Figure 2. Firstly, it is necessary to collect the researchers' data and their scientific production, available on an indexed bibliographic database. In this work, we chose Scopus because it is quite extensive and well-recognized by the research community. As the data collection process is not coupled to the DVC pipeline, it is possible to use any bibliographic database to collect researchers' data and their production, as long as these data have the same format as the required inputs.

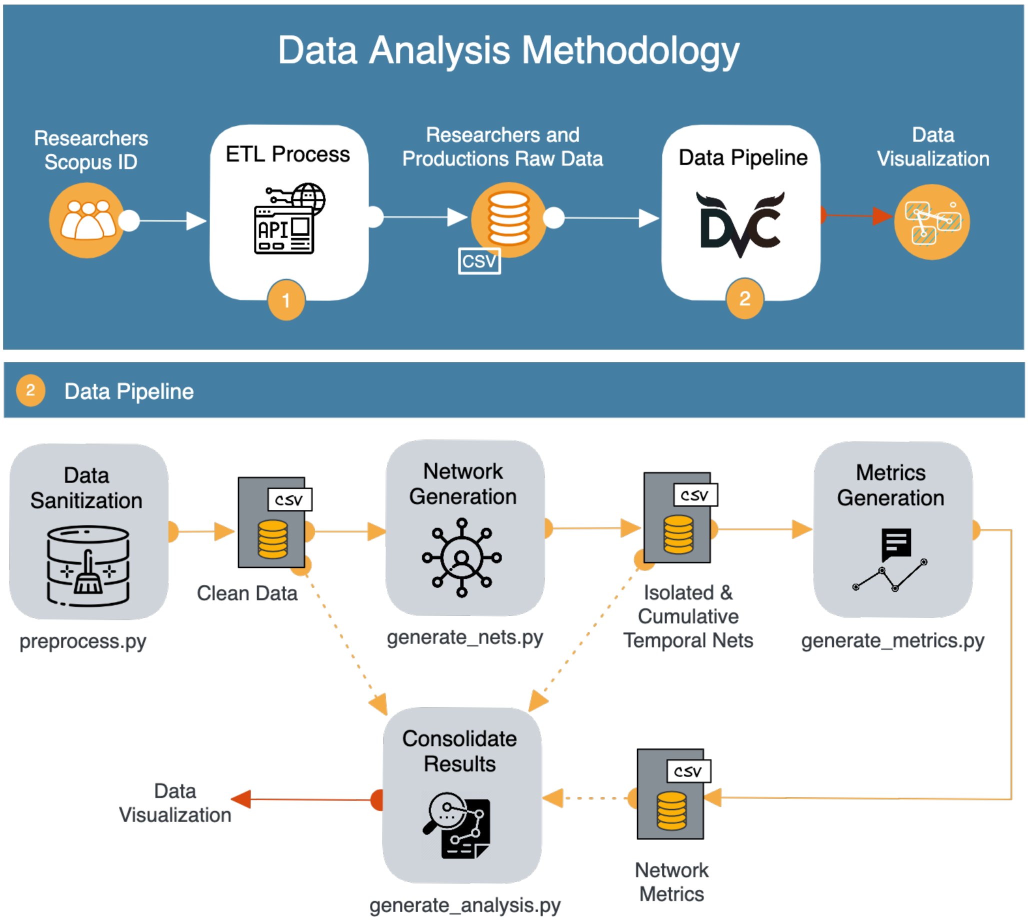


Figure 2. A brief account of the proposed methodology. In the first step, data are filtered and extracted from the Scopus API. The second step employs an innovative pipeline created with the aid of the DVC tool. The pipeline is responsible for executing data sanitization, generating data structures in network form, and computing evaluation metrics. The outcome of the pipeline provides a thorough analysis of the data, empowering effective decision making.

The considered data can provide an overview of a research group's scientific activities, research focus, members' and group's collaboration patterns, beyond the cumulative—i.e., cumulative sliding window and k-window—and isolated temporal cohesion analysis of the research group. Therefore, for this work, the collected datasets (i.e., researchers' data and their production) have researchers' updated stats and their works published until 14 October 2022 (date of retrieval).

With the input data properly collected, we start the data analysis pipeline. In the first stage, all raw data are pre-processed using the Pandas Data Wrangling tool, generating the Researchers and Group Production datasets used in this work. Next, from those datasets, using the NetworkX tool, the temporal networks are modeled and generated for each year specified in the "year" feature on the Group Production dataset. In addition, for each year, there are three versions of temporal co-authorship networks: (i) an isolated variant, which only considers the works published in a specific year; (ii) a version with cumulative sliding window, that considers the studies published until a certain year; and (iii) k-windowed, which considers studies whose year of publication is within the range $[year - k, year]$.

After the generation of the temporal nets, a dataset is created, containing the main CNA metrics for all the nodes in the third stage. In other words, it processes the group's members and partners of scientific collaboration. The used metrics in this stage are betweenness centrality, closeness centrality, clustering coefficient, degree centrality, eccentricity, eigenvector centrality, and number of cliques. This way, it is possible to compute the main statistics (mean, min, max and standard deviation) for each metric, both for the complete network and for research groups and subgroups.

For [8], the betweenness centrality of a node v is the sum of the fraction of all-pairs shortest paths that pass through v, i.e., it represents the level of mediation of the network nodes. The closeness centrality defines a ratio of the fraction of nodes in the net that are reachable, to the average distance from the reachable nodes [16]. According to [17], the degree centrality of a node v is the fraction of nodes it is connected to. The eigenvector centrality computes the centrality for a node based on the centrality of its neighbors, i.e., it represents the importance level of a node based on the value of this metric of its neighbors [8,16]. According to [8], the clustering coefficient of a node v is the fraction of possible triangles through that node that exists, while the eccentricity of a node v is the maximum distance from v to all other nodes in a network. Additionally, the number of cliques is the number of vertices in a maximum clique of a graph, where a maximal clique is a clique to which no more vertices can be added, as well as a clique being a subset of nodes all linked to each other [8,16,17].

In the last stage, all generated datasets and temporal nets are used in tasks of data analysis and visualization to extract insights hidden in the datasets, and the temporal nets are converted in vectorial structures through the graph and node embeddings algorithms [18–20]. These data along with the metric datasets are submitted to machine learning algorithms for pattern detection. So, all these techniques and used analyses are necessary tools to support the evaluation process and cohesion analysis of research groups.

As a result, the workflow for preprocessing, generating, and analyzing datasets and temporal nets is versioned with the DVC, making it simple to replicate and audit [21,22]. The data version control ensures the reliability and integrity of data science pipelines, and it is Git-compatible, with lock-free, local ramifications, and version control. Moreover, DVC is often used to version data and data workflows in the same way that source code is [22].

The methodology is interactive and iterative because it is possible to analyze and evaluate the extracted knowledge using data visualization techniques (within the data analysis stage), and if such patterns and information are insufficient, we can reapply or refine the data analysis process until the insights obtained are satisfactory. It is interactive because the researcher engages throughout the analysis process, and it is iterative because the phases in the data analysis process are repeated until suitable results are produced.

With the proposed approach properly presented, its empirical evaluation is detailed in the next section.

4. Materials and Methods

This section describes a case study for our approach, which was based on an experimental process such as that presented by [21,23,24]. The next subsections focus on the definition and planning of this empirical evaluation. The last subsection presents its operation process.

4.1. Goal Definition

The main goal of this study is to evaluate the suitability and feasibility of the proposed methodology for the process of evaluation and cohesion analysis of research groups. This goal is formalized using the GQM (goal–question–metric) template proposed by [25] and presented by [26]: **analyze** the proposed methodology **with the purpose of** evaluating it **with respect to** the feasibility and analytical capacity **from the point of view of** researchers, affiliations and research groups that work with studies associated with the evaluation and cohesion analysis of research groups **in the context of** the published peer-reviewed manuscripts of a research group.

4.2. Planning

This subsection details the design of the case study.

4.2.1. Participant and Artifact Selection

After the goal definition of this empirical evaluation, the process of selecting participants and objects was started. Firstly, for convenience, aiming to better control the sampling of manuscripts, and consequently facilitating the data preprocessing process, the chosen research group was composed of researchers/professors of the graduate program in Computer and Electrical Engineering at Federal University of Rio Grande do Norte (UFRN), referred as PPgEEC. This choice is due to the high familiarity in precisely determining the set of manuscripts to be analyzed and also for being one of the largest UFRN programs in terms of active students and, consequently, in scholar productivity. This scholar group has different research lines, as shown in the Table 2.

Table 2. Group characterization.

ID	Research Line	Number of Members
RL_1	Automation and Systems	10
RL_2	Computer Engineering	11
RL_3	Telecommunication	5

Note. ***RL***: Research Line.

The analyzed period (2010–2022) was chosen, which corresponds to four Coordination of Improvement of Higher Education Personnel (CAPES) global evaluations for Brazilian graduate programs, wherein the last one (2021–2024) is in progress. Particularly, CAPES is a foundation of the Ministry of Education (MEC) in Brazil that is responsible for evaluating graduate programs in universities, among other functions [27].

One of the available resources is the CAPES evaluation system that serves as an instrument for the university community when searching for an academic excellence standard for the national master's and doctoral programs. Such evaluations serve as a basis for the formulation of policies in this area, as well as for the dimensioning of funding actions (scholarships, financial and personal grants/support) [27,28].

Finally, for legal reasons, we will not use the names of the research group's members in this work. Letters and/or numbers will be used to identify each individual.

4.2.2. Research Questions

After the definition of the participants and artifacts, the research questions we want to explore in this work are as follows:

- Is it possible to evaluate a research group based on the centrality metrics of co-authorship networks?
- Based on centrality metrics, is it possible to assess the evolution of a research group over time?
- Do isolated temporal networks tend to have a large number of subgroups isolated from the strongly connected component of a collaborative network, and consequently low cohesion indices?
- Can the number of connected components be considered as a metric for the cohesion analysis of research groups?
- Can the chosen research group be considered cohesive or is it becoming more cohesive?
- Can the use of graph embeddings help in the detection of collaboration patterns in temporal networks?

- Can the use of node embeddings help in the detection of patterns of collaborations in a research group?
- Is there any correlation between the group members' lines of research and their patterns of scientific collaboration?
- Is the proposed methodology suitable to support the tasks of evaluation and cohesion analysis of research groups?

The metrics to evaluate these questions are: (1) number of manuscripts; (2) number of connected components; (3) number of nodes in a connected component; (4) indicators of CNA discussed earlier in Section 3; (5) accuracy; and (6) F1 score.

4.2.3. Instrumentation

The materials and/or resources used in this work are as follows (accessed date (20 February 2023)):

- Python Data Science ecosystem (pandas (https://pandas.pydata.org), NumPy (https://numpy.org), Matplotlib (https://matplotlib.org), seaborn (https://seaborn.pydata.org), scikit-learn (https://scikit-learn.org) and others), provided by Anaconda platform (https://www.anaconda.com) or Google Colab (https://colab.research.google.com);
- Anaconda's Jupyter Lab;
- NetworkX (https://networkx.org) e Gephi (https://gephi.org), library and tool, respectively, for modeling, visualization, analysis and manipulation of complex networks;
- Node2Vec (https://snap.stanford.edu/node2vec/) and Graph2Vec (https://github.com/benedekrozemberczki/karateclub) libraries for conversion of graphs into vector structures;
- Data Version Control;
- The researchers' stats, scholarly production associated to research group chosen, and the proposed methodology, both discussed in Section 3;
- The Jupyter Notebooks that contain all source code for performing the data analysis, which are available in a GitHub repository (https://github.com/breno-madruga/evaluation-research-groups).

4.3. Operation

This subsection describes the preparation and execution of this empirical evaluation. The operation process was performed initially with the configuration of the environment for the case study and planning of data collection.

Firstly, the scholarly data related to the researchers' stats and scientific production of research group were collected. Next, the analysis pipeline was defined, as detailed in Section 3. Finally, the analysis process earlier discussed was performed (see Section 3), with the required artifacts (see Section 4.2.3).

After the execution, the analyses results were obtained, which were based on the metrics previously discussed (see Section 4.2.2). It is worth mentioning that these results are used to answer the research questions of this work.

The results related to this empirical evaluation are presented in the next section.

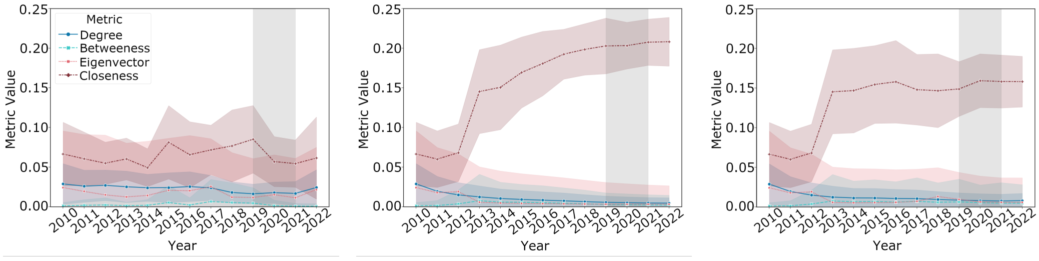
5. Results and Discussion

This section presents the results of the empirical evaluation that answers the research questions of this work. It is divided into three parts: (i) the conventional analysis; (ii) the temporal nets' embeddings analysis; and (iii) the research group's members analysis. We would like to stress that the methodology and all results can be reproduced from the source code available in a public repository previously mentioned (https://github.com/breno-madruga/evaluation-research-groups, accessed on 20 February 2023).

5.1. Conventional Analysis

As mentioned earlier by [10,11], the indicators extracted from CNA techniques are correlated with the research performance of scholars, and it is possible to compare the performance of research groups based on these types of indicators. Hence, since the centrality metrics can estimate the significance of scholars in their scientific communities, a temporal evaluation of the chosen research group was carried out based on the main centrality metrics in different temporal scenarios.

As can be seen in Figure 3, when analyzing the cumulative scenarios (Figure 3b,c), there are some small variations in the betweenness, eigenvector and degree centralities, with a decreasing trend for the period 2010–2014. However, there is a noticeable stabilization of these metrics from year 2015 with small variations too. This behavior could be an indication of the maturation of the group, as well as the consolidation of previously established partnerships. In addition, this fact is in line with the change of some old members by new senior ones (remaining until today), acting significantly for the evolution of the team scientific activities.



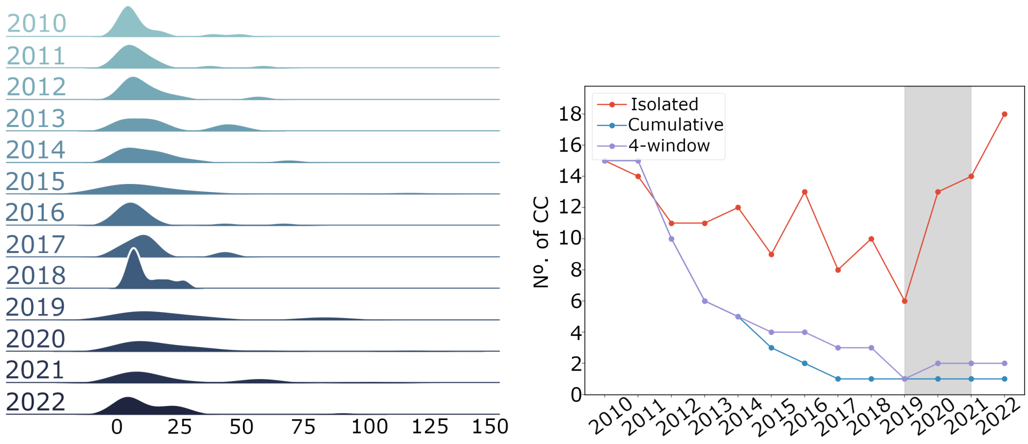
(**a**) Isolated (**b**) Cum. sliding window (**c**) Cum. with 4-window

Figure 3. Temporal comparative analysis of PPgEEC centrality metrics.

Moreover, this same behavior can be seen in the isolated scenario (Figure 3a), with a positive advance of the metrics being noticeable for the year 2015, in line with the fact of the change of researchers, with some small oscillations over the years. Still in this same scenario, during the COVID-19 pandemic (2019–2020, the area highlighted in gray in the Figure 3), there was an increase in the eigenvector and degree metrics, while an opposite behavior was identified for the betweenness and closeness indicators. A possible reason for this effect is that, during the pandemic period, new and relevant scientific collaborations were formed with new academic partners for the task force against COVID-19. On the other hand, these new partnerships could have possibly inhibited collaborations between team members.

Therefore, as it was already discussed by [2,10,11], despite the difficulty in understanding the non-trivial patterns detected from the centrality metrics, it was still possible to verify and monitor the evolution and maturity level of the research group chosen from these indicators.

Continuing the data analysis, the levels of team cohesion were evaluated over the chosen period (see Figure 4). Firstly, it was analyzed the distribution of the connected components' order (number of nodes) for each temporal scenario. For isolated nets, as can be seen in Figure 4a, there is a considerable amount of connected components (CC) with small numbers of nodes. Thus, analyzing each year, there is a strong incidence of subgroups collaborating in isolation, which may be an indication of low cohesion. In addition, the distributions of both cumulative scenarios are not presented because they did not show representative results.



(**a**) Distribution of No. of nodes in CC in isolated nets.

(**b**) Time Series of Number of CC

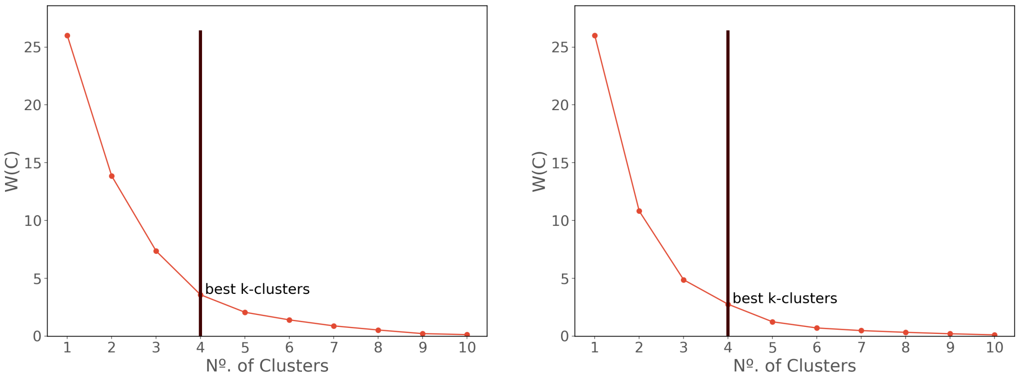
Figure 4. Temporal comparative analysis of connected components.

Moreover, comparing Figure 4a,b, considering the publications history, for the cumulative scenarios, it is evident that, until 2016, there was a concern of team members to collaborate with each other, increasing the level of cohesion. After that, it reached an excellent level of cohesion from 2017, having a maximum of two connected components.

5.2. *Temporal Networks' Embeddings Analysis*

In order to investigate the collaboration patterns of the chosen research group from its temporal nets, the graph embeddings techniques were combined with unsupervised machine learning to extract collaboration patterns and/or similarities from these temporal networks. For each temporal scenario, vector representations were generated with two dimensions of each network by using the Graph2Vec library [18,20]. Aiming to facilitate the plotting on 2D plane and, as [29–31] highlighted, it is possible to use a low number of dimensions for small real-world networks. From those 2D vector representations, a widely known clustering algorithm for patterns detection, K-means, was used [32–34].

The main parameter of K-means is the number of clusters to be detected, widely known as *k-clusters*. This way, in order to define the best value of this parameter, the elbow method was used [32,34]. After the execution of this method, as can be seen in Figure 5, the best value of *k-clusters* is 4 for both isolated and k-window scenarios, respectively (see Figure 5a,b). Interestingly, this value also corresponds to the number of existing CAPES evaluations within the chosen period.



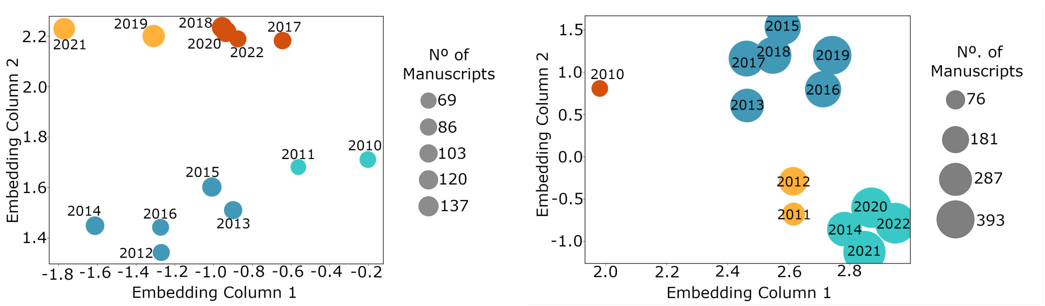
(**a**) Isolated

(**b**) Cumulative with 4-window.

Figure 5. Discovering the optimal k-value for K-means for isolated and 4-window cumulative nets, respectively.

Once the parameter *k-clusters* was defined, it was possible to apply the K-means algorithm on the embeddings set to detect insights from the temporal networks in each scenario, as shown in Figure 6. For isolated nets (see Figure 6a), it is possible to identify the similarities between the collaboration patterns of temporal networks; additionally, their number of manuscripts can apparently have some correlation with their respective clusters, although this feature was not used in the clustering process. This can be an indication that the patterns of collaborations were balanced over the years. In addition, it is not possible to confirm that there is any correlation between the defined clusters with the CAPES evaluations, and further investigations are necessary.

For cumulative nets with 4-window (see Figure 6b), the intersection of collaborations between temporal windows was an almost predominant factor for the formation of the clusters, especially the cluster with nets whose labels were 2013, and 2015 to 2019. Additionally, their number of manuscripts could apparently have some correlation with their respective clusters, although this feature was not used in the clustering process too. Possibly with the change of members in 2015, collaboration patterns became very similar, as most temporal networks are in the same group. Additionally, apparently, the constancy of collaboration patterns has been affected by the pandemic, as the periods 2020 to 2022 are in a different group. Again, it is not possible to confirm that there is any correlation between the defined clusters with the CAPES evaluations, and further investigations are necessary.



(**a**) Isolated

(**b**) Cumulative with 4-window

Figure 6. Clustering of PPgEEC nets' graph embeddings. In both figures, the bubbles' color represents the cluster defined by K-means algorithm and their size is the number of manuscripts.

5.3. Group's Members Analysis

Finally, aiming to investigate a possible correlation between the collaboration patterns with the research lines of chosen group, a pipeline was developed to perform this analysis type, named multi-level clustering, as can be seen in Figure 7. More specifically, the use of CNA metrics was compared against the node embeddings method in order to verify which one would be the most appropriate to identify research lines from the collaboration patterns.

The processing pipeline has three stages: clustering, labeling, and evaluation. In the clustering step, initially, the team members' unlabeled data are generated from the graph feature engineering substep. This substep can create this dataset from two approaches: (i) extracting the features from the nets metrics data by the proposed methodology; or (ii) generating new features from the temporal networks by using the Node2Vec algorithm for node embeddings [18,19].

For the approach of CNA metrics, it was computed the mean, min, max and standard deviation of each metric used, totaling 28 features for each member of the chosen group. As for the embeddings approach, it was generated ten new features by Node2Vec. As the number of dimensions (d) was also the main parameter of the node embedding algorithms, the value 10 was chosen for this parameter based on the study performed by [30] because the representation obtained by Node2Vec at $d \simeq 10$ was already able to capture the structural features in small real-world nets.

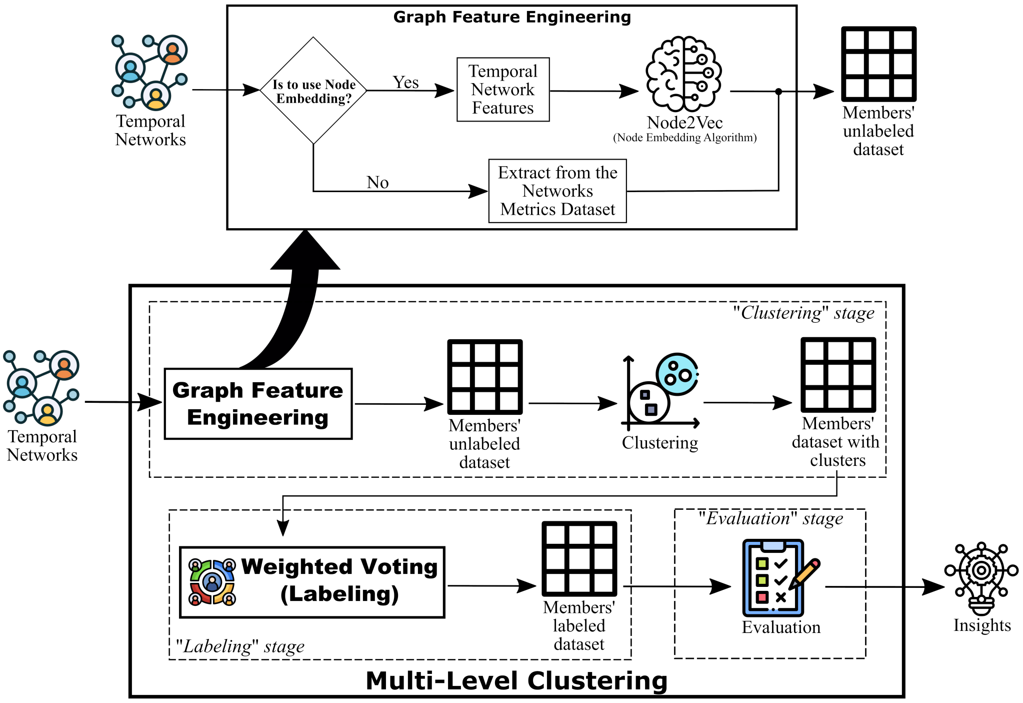
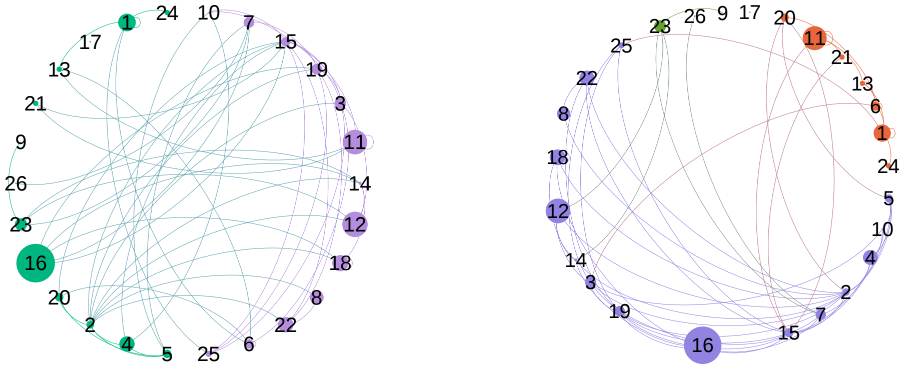

Graph Feature Engineering
Temporal Networks
Is to use Node Embedding?
Yes
No
Temporal Network Features
Node2Vec
(Node Embedding Algorithm)
Extract from the Networks Metrics Dataset
Members' unlabeled dataset
"Clustering" stage
Graph Feature Engineering
Members' unlabeled dataset
Clustering
Members' dataset with clusters
Weighted Voting (Labeling)
Members' labeled dataset
"Labeling" stage
"Evaluation" stage
Evaluation
Insights
Multi-Level Clustering

Figure 7. Multi-level clustering for definition of research lines.

In the sequence, a clustering process was performed on the unlabeled data, generating a new dataset where the researchers were grouped into three distinct clusters. For this, it was used the K-Means algorithm with *k-clusters* equal to 3 because this value is the same quantity of thematic lines belonging to the research team. After clustering the data, it was proceeded to the next stage, labeling.

In the labeling step, the real research line of each member was initially included into the dataset, based on the chosen group's website (https://sigaa.ufrn.br/sigaa/public/programa/equipe.jsf?lc=pt_BR&id=103 (accessed on 20 February 2023)). Next, the weighted voting process was performed according to Algorithm 1. In summary, this voting process determines the research line of each cluster, based on the count of thematic lines within a cluster, normalized by the total number of researchers in each research line. In addition, in the case of a tie, the resulting research line of a cluster is defined by the thematic line with the greatest number of manuscripts. For example, in Figure 8, the result of the labeling stage for the period 2017–2020 of CAPES evaluation is presented, where the predicted research lines are, respectively, 1 (green), 2 (purple) and 3 (orange).



(**a**) Networks Metrics (**b**) Embeddings

Figure 8. Clustering of PPgEEC members (2017–2020).

Algorithm 1 Weighted voting process.

```
Require: Data $D$, and dictionary of number of members per research line $rl$
Ensure: Data $D$ updated with the research line determined for each cluster
get the unique list of clusters' identifiers $X \subset D[cluster]$
for all $c_k \in X$ do
    $score \leftarrow$ an empty dictionary
    get the unique list of research line $R \subset D[real\_rl \mid D[cluster] = c_k]$
    for all $r_k \in R$ do
        $score[r_k] \leftarrow \mid \{r_k \in D[real\_rl \mid D[cluster] = c_k]\} \mid$
        $score[r_k] \leftarrow \frac{score[r_k]}{rl[r_k]}$
    end for
    $main\_line \leftarrow \arg\max_{r_k} score[r_k] = \{r_k \mid \max_x score[x]\}$
    if $\mid main\_line \mid = 1$ then
        $D[pred\_rl \mid D[cluster] = c_k] \leftarrow main\_line[0]$
    else if $\mid main\_line \mid > 1$ then
        $score \leftarrow$ an empty dictionary
        for all $r_k \in main\_line$ do
            $M \leftarrow D[number\_manuscript \mid D[cluster] = c_k, D[real\_rl] = r_k]$
            $score[r_k] \leftarrow \sum_{x=1}^{|M|} m_x$
        end for
        $main\_line \leftarrow \arg\max_{r_k} score[r_k] = \{r_k \mid \max_x score[x]\}$
        $D[pred\_rl \mid D[cluster] = c_k] \leftarrow main\_line[0]$
    end if
end for
return $D$
```

With the actual research lines and those predicted by the voting process, it is possible to verify which approach is most accurate in determining the research lines based on the collaboration patterns. thus, in the evaluation stage, the main quality metrics of supervised learning were used. They were (i) accuracy, the proportion of scholars that are correctly classified, and (ii) F1 score, called the harmonic mean of the measures precision and recall [33,35]. Precision is the proportion of researchers classified as belonging to a specific research line that were really from this line, while recall is the proportion of members of a specific research line that were correctly classified as that line.

From the results obtained in the evaluation stage, according to Table 3, the node embedding approach was the best in terms of accuracy, as well as being more successful in grouping researchers from research lines 2 and 3. Furthermore, for RL_1, there was a slight difference between the F1 scores of the approaches evaluated, however, in favor of the CNA metrics. In relation to the relatively low values of accuracies, this fact could be an indication that the collaborations are based on the multidisciplinary factor. Despite the clear separation between the thematic lines, it is possible that this separation is becoming more tenuous due to such multidisciplinary behavior.

With the results properly presented, the threats to the validity of this study are explained in the next section.

Table 3. Results of clustering of members.

Period	Network Metrics				Node Embedding			
	Accuracy	$F1_{RL_1}$	$F1_{RL_2}$	$F1_{RL_3}$	Accuracy	$F1_{RL_1}$	$F1_{RL_2}$	$F1_{RL_3}$
2010–2012	52%	0%	67%	57%	56%	0%	69%	67%
2013–2016	58%	61%	67%	0%	69%	50%	77%	80%
2017–2020	58%	61%	67%	0%	69%	57%	77%	67%
2021–2022	58%	50%	71%	0%	69%	67%	73%	67%

Note. ***RL***: Research Line; ***F*1:** *F*1 Score.

6. Threats to Validity

The threats to the validity of this work are as follows:

- **Collection bias:** A risk associated with the collecting procedure is the possibility of relevant manuscripts being published after the date of retrieval, in addition to failing to consider additional sources of supplementary data that may contain relevant works from the evaluated research group. Thus, the main production base extensively used by the research world was employed to mitigate this problem.
- **Indexing bias:** This threat could not be attenuated since the data contained in the utilized datasets are classified and maintained by external parties, which is beyond the limits of the suggested solution.
- **Ethical approval:** Given that this was a metadata analysis of published manuscripts, no ethics committee permission was necessary.
- **External validity:** An evaluation and cohesion analysis of a research group was carried out for the period from January 2010 to 14 October 2022. However, it is possible that some relevant work was not yet indexed or is indexed on bases other than the employed one. Therefore, it is impossible to generalize the conclusions obtained to validate the complete effectiveness of the proposed methodology. However, the results are pretty relevant to outline future research group evaluation and cohesion analysis investigations.

7. Conclusions

The process of cohesion analysis and evaluation of research groups is essential for determining potential collaborators, as well as for promoting research in this area. Thus, the development of new approaches for this purpose is necessary because the current approaches do not contemplate both types of evaluation/analysis, and they also do not consider the temporal factor of co-authorship networks.

In order to fulfill the discussed research gaps in the literature, a graph data science-oriented methodology was developed, combining machine learning, complex network analysis and graph embeddings techniques to evaluate research groups concerning temporal metadata, in addition to temporal cohesion analysis. Moreover, a case study was considered to evaluate the effectiveness of the proposed approach.

The achieved results suggested the feasibility of the proposed methodology indicating the main researchers, temporal behavior of the research group's scientific collaboration, the alignment between scholars' research focus and their patterns of collaboration, and the temporal cohesion of the research groups. The presented approach has the potential to instrument and expand strategic and proactive decisions of research entities when the performed analyses are properly considered.

Future work will address specific open problems, including discovering similar and complementary groups, which is strategically crucial to scientific funding agencies. This enables identifying and capturing talent for collaborative projects and finding individuals with specific expertise. However, such knowledge is dynamic and may contain latent properties that vary over time, with periods of low production followed by high activity due to external stimuli or anomalous scenarios, such as the impact of the COVID-19 pandemic. Another area of interest is identifying individual talent and normalizing metrics to facilitate their detection. Additionally, exploring patterns that incorporate the textual aspect of research, such as abstracts, may further enhance the evaluation of research groups. Finally, different issues associated to graph labeling will be also investigated, especially when concerning its relation to educational data [36,37] These open problems represent exciting directions for further research, demonstrating the potential of this methodology for identifying and capturing talent and enhancing the evaluation of scientific collaboration networks.

Author Contributions: Conceptualization, B.S.S. and I.S.; data curation, B.S.S. and I.S.; formal analysis, B.S.S. and I.S.; investigation, B.S.S. and I.S.; methodology, B.S.S. and I.S.; project administration, I.S.; resources, I.S.; software, B.S.S.; supervision, D.G.C. and I.S.; validation, B.S.S. and I.S.; visualization, B.S.S., D.G.C. and I.S.; writing—original draft preparation, B.S.S.; writing—review and editing, B.S.S., D.G.C. and I.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data generated and used during this research are openly available on Mendeley Data [38]. In addition, the code and pipeline are hosted in a public repository on GitHub (https://github.com/breno-madruga/evaluation-research-groups (accessed on 20 February 2023)).

Conflicts of Interest: All authors declare that they have no conflict of interest.

References

1. Sugimoto, C.R.; Larivière, V. *Measuring Research: What Everyone Needs to Know*; Oxford University Press: Oxford, UK, 2018.
2. Amat, C.B.; Perruchas, F. Evolving cohesion metrics of a research network on rare diseases: A longitudinal study over 14 years. *Scientometrics* **2016**, *108*, 41–56. [CrossRef]
3. Vinkler, P. *The Evaluation of Research by Scientometric Indicators*; Chandos Publishing: Oxford, UK, 2010.
4. Franceschini, F.; Maisano, D. Structured evaluation of the scientific output of academic research groups by recent h-based indicators. *J. Inf.* **2011**, *5*, 64–74. [CrossRef]
5. Mryglod, O.; Holovatch, Y.; Kenna, R. Big fish and small ponds: Why the departmental h-index should not be used to rank universities. *Scientometrics* **2022**, *127*, 3279–3292. [CrossRef]
6. Kudelka, M.; Plato, J.; Krömer, P. Author evaluation based on H-index and citation response. In Proceedings of the 2016 International Conference on Intelligent Networking and Collaborative Systems (INCoS), Ostrava, Czech Republic, 7–9 September 2016; pp. 375–379. [CrossRef]
7. Montazerian, M.; Zanotto, E.D.; Eckert, H. A new parameter for (normalized) evaluation of H-index: Countries as a case study. *Scientometrics* **2019**, *118*, 1065–1078. [CrossRef]
8. Menczer, F.; Fortunato, S.; Davis, C.A. *A First Course in Network Science*; Cambridge University Press: Cambridge, UK, 2020.
9. Wang, D.; Barabási, A.L. *The Science of Science*; Cambridge University Press: Cambridge, UK, 2021.
10. Jeon, H.J.; Lee, O.J.; Jung, J.J. Is performance of scholars correlated to their research collaboration patterns? *Front. Big Data* **2019**, *2*, 1–10. [CrossRef] [PubMed]
11. Wiechetek, Ł.; Pastuszak, Z. Academic social networks metrics: An effective indicator for university performance? *Scientometrics* **2022**, *127*, 1381–1401. [CrossRef]
12. Camargo, L.S.d.; Barbosa, R.R. Bibliometria, Cientometria e um possível caminho para a Construção de Indicadores e Mapas da Produção Científica. *PontodeAcesso* **2018**, *12*, 109–125. [CrossRef]
13. Moral-Munoz, J.A.; López-Herrera, A.G.; Herrera-Viedma, E.; Cobo, M.J. Science Mapping Analysis Software Tools: A Review. In *Springer Handbook of Science and Technology Indicators*; Glänzel, W., Moed, H.F., Schmoch, U., Thelwall, M., Eds.; Springer International Publishing: Cham, Switzerland, 2019; pp. 159–185. [CrossRef]
14. Ju, H.; Zhou, D.; Blevins, A.S.; Lydon-Staley, D.M.; Kaplan, J.; Tuma, J.R.; Bassett, D.S. Historical growth of concept networks in Wikipedia. *Collect. Intell.* **2022**, *1*. [CrossRef]
15. Keramatfar, A.; Rafiee, M.; Amirkhani, H. Graph Neural Networks: A bibliometrics overview. *Mach. Learn. Appl.* **2022**, *10*, 100401. [CrossRef]
16. Zweig, K.A. *Network Analysis Literacy*; Springer: Berlin/Heidelberg, Germany, 2016.
17. Zinoviev, D. *Complex Network Analysis in Python: Recognize-Construct-Visualize-Analyze-Interpret*; Pragmatic Bookshelf: North Carolina, NC, USA, 2018.

18. Grohe, M. Word2vec, Node2vec, Graph2vec, X2vec: Towards a Theory of Vector Embeddings of Structured Data. In Proceedings of the 39th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, Portland, OR, USA, 14–19 June 2020; PODS'20. Association for Computing Machinery: New York, NY, USA, 2020; pp. 1–16. [CrossRef]
19. Grover, A.; Leskovec, J. Node2vec: Scalable Feature Learning for Networks. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; KDD '16. Association for Computing Machinery: New York, NY, USA, 2016; pp. 855–864. [CrossRef]
20. Narayanan, A.; Chandramohan, M.; Venkatesan, R.; Chen, L.; Liu, Y.; Jaiswal, S. graph2vec: Learning Distributed Representations of Graphs. *arXiv* **2017**, arXiv:1707.05005.
21. Santos, B.S.; Silva, I.; Lima, L.; Endo, P.T.; Alves, G.; Ribeiro-Dantas, M.d.C. Discovering temporal scientometric knowledge in COVID-19 scholarly production. *Scientometrics* **2022**, *127*, 1609–1642. [CrossRef] [PubMed]
22. Kuprieiev, R.; Skshetry; Petrov, D.; Rowlands, P.; Redzyński, P.; da Costa-Luis, C.; Schepanovski, A.; Gao; de la Iglesia Castro, D.; Shcheklein, I.; et al. DVC: Data Version Control-Git for Data & Models. Zenodo. February 2023. Available online: https://doi.org/10.5281/zenodo.3677553 (accessed on 20 February 2023).
23. Santos, B.S.; Júnior, M.C.; da Paixão, B.C.; Santos, R.M.; Nascimento, A.V.R.P.; dos Santos, H.C.; Filho, W.H.L.; de Medeiros, A.S.L. Comparing Text Mining Algorithms for Predicting Irregularities in Public Accounts. In Proceedings of the XI Brazilian Symposium on Information Systems SBSI 2015, Goiania, Goias, Brazil, 26–29 June 2015; Brazilian Computer Society: Porto Alegre, Brazil, 2015; pp. 667–674.
24. Santos, B.S.; Silva, I.; Melo, E. Metodologia orientada a ciência de dados em grafos para avaliação de PPGs. In Proceedings of the XV Simpósio Brasileiro de Automação Inteligente (SBAI 2021), Virtual, 17–19 October 2021; Sociedade Brasileira de Automática: Rio Grande, Rio Grande do Sul, Brazil, 2021; pp. 1998–2005. [CrossRef]
25. Basili, V.R.; Weiss, D.M. A Methodology for Collecting Valid Software Engineering Data. *IEEE Trans. Softw. Eng.* **1984**, *SE-10*, 728–738. [CrossRef]
26. van Solingen, D.R.; Berghout, E.W. *The Goal/Question/Metric Method: A Practical Guide for Quality Improvement of Software Development*; McGraw-Hill: New York, NY, USA, 1999.
27. CAPES. CAPES—Institutional Page. 2022. Available online: https://www.gov.br/capes/pt-br/acesso-a-informacao/institucional/historia-e-missao (accessed on 18 October 2022).
28. CAPES. CAPES—Quadrennial Evaluation. 2022. Available online: https://www.gov.br/capes/pt-br/acesso-a-informacao/acoes-e-programas/avaliacao/sobre-a-avaliacao/avaliacao-o-que-e/sobre-a-avaliacao-conceitos-processos-e-normas/conceito-avaliacao (accessed on 18 October 2022).
29. Cai, H.; Zheng, V.W.; Chang, K.C.C. A comprehensive survey of graph embedding: Problems, techniques, and applications. *IEEE Trans. Knowl. Data Eng.* **2018**, *30*, 1616–1637. [CrossRef]
30. Gu, W.; Tandon, A.; Ahn, Y.Y.; Radicchi, F. Principled approach to the selection of the embedding dimension of networks. *Nat. Commun.* **2021**, *12*, 1–10. [CrossRef] [PubMed]
31. Longa, A. Graph Embedding in 2D. Master's Thesis, Università degli Studi di Trento, Trento, Italy, 2019.
32. Bonaccorso, G. *Hands-On Unsupervised Learning with Python*; Packt Publishing Ltd.: Birmingham, UK, 2019.
33. Müller, A.C.; Guido, S. *Introduction to Machine Learning with Python: A Guide for Data Scientists*; O'Reilly Media: Sebastopol, CA, USA, 2016.
34. Patel, A.A. *Hands-On Unsupervised Learning Using Python: How to Build Applied Machine Learning Solutions from Unlabeled Data*; O'Reilly Media: Sebastopol, CA, USA, 2019.
35. Bramer, M. *Principles of Data Mining*; Springer: Berlin/Heidelberg, Germany, 2016.
36. Zhou, S.; Yuan, P.; Liu, L.; Jin, H. MGTag: A Multi-Dimensional Graph Labeling Scheme for Fast Reachability Queries. In Proceedings of the 2018 IEEE 34th International Conference on Data Engineering (ICDE), Paris, France, 16–19 April 2018; pp. 1372–1375. [CrossRef]
37. Agrawal, G.; Deng, Y.; Park, J.; Liu, H.; Chen, Y.C. Building Knowledge Graphs from Unstructured Texts: Applications and Impact Analyses in Cybersecurity Education. *Information* **2022**, *13*, 526. [CrossRef]
38. Santos, B.; Silva, I.; Costa, D.G. Research Group Dataset. Dataset Version 2, Mendeley Data. 2022. Available online: https://doi.org/10.17632/rwfd6p6xsd (accessed on 20 February 2023).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.