Uncovering Collaboration Patterns in Brazilian Computer Science Graduate Programs Through Network Embeddings

Icaro Luiz Lage Vasconcelos[1], Augusto Ferreira Guilarducci[1], Jadson Castro Gertrudes[1], Gladston Juliano Prates Moreira[1], Vander Luis de Souza Freitas[1], Eduardo Jose da Silva Luz[1]

[1]Federal University of Ouro Preto
35400 – 000 – Ouro Preto – MG – Brazil

icaro.vasconcelos@aluno.ufop.edu.br

Abstract. *The Science of Science (SciSci) is crucial for understanding scientific production. However, there remains a gap in analyzing how geographic and productivity factors influence collaboration. This study explores collaboration networks in Brazilian Computer Science graduate programs. Using embeddings, clustering techniques, and Decision trees, we identified a larger group representing average scientific production, while smaller high-performing clusters often include researchers from prestigious institutions. Geographic disparities also highlight regional differences, with the South and Southeast regions dominating distinct groups. These findings emphasize the interplay between location, productivity, and impact, offering insights into collaboration dynamics.*

1. Introduction

The Science of Science (SciSci) is an interdisciplinary approach that offers a quantitative lens to understand the dynamics of scientific production and knowledge dissemination [Fortunato et al. 2018]. An important element of SciSci is the analysis of academic collaboration networks, which reveal the complex interactions among researchers, institutions, and the broader scientific community. Understanding these networks is important to expose conditions that either foster or hinder creativity and innovation, among other processes related to doing science [Fortunato et al. 2018]. This is particularly relevant for evaluating and improving the effectiveness of research ecosystems, such as Graduate Programs.

The analysis of scientific collaboration networks presents several challenges. Science operates as a self-organizing, evolving, and multiscale system, where knowledge flows through complex interactions encoded in academic artifacts such as publications and citations [Fortunato et al. 2018]. In the Brazilian context, for example, disparities in citation practices across subfields of Computer Science can lead to unfair evaluations when uniform criteria are applied [Druszcz and Vignatti 2024]. To address these intricacies, recent studies have explored collaboration networks from multiple perspectives, emphasizing both structural and contextual dimensions.

For instance, [FitzGerald et al. 2023] analyzed the hiring and collaboration history in the US mathematics field over seven decades, revealing systemic inequalities in academic mobility and the influence of institutional prestige, while [Bai et al. 2021] combined co-authorship and citation data to build heterogeneous networks that capture the interplay between collaboration and scientific impact.

On the other hand, [Liu et al. 2021] introduced Shifu2, a network representation learning model that identifies advisor-advisee relationships by jointly modeling the structural and semantic features of co-authorship networks across multiple disciplines. Similarly, [di Bella et al. 2021] analyzed internal and external co-authorship networks within the Italian Institute of Technology (IIT), showing that research productivity is closely related to centrality within the institutional network, and that internal collaboration tends to be denser and more persistent than external ties. Their findings reinforce the importance of considering institutional boundaries and the role of long-term affiliation in shaping collaborative behavior.

This research takes the previous papers as inspiration, and addresses key questions about the structure of the Brazilian Graduate Programs in Computer Science (BRGPCS) co-authorship network. We investigate the primary factors that determine the formation of collaboration ties under some hypotheses. **Hypothesis 1 - Geographic influence**: researchers in the same region are more likely to collaborate; **Hypothesis 2 - Productivity Influence**: academic productivity metrics, such as citations and h-index, are strong predictors of an author's position within the network; **Hypothesis 3 - Institutional Influence**: researchers from highly-ranked institutions are more likely to cluster; **Hypothesis 4 - Clusterability**: node2vec embeddings [Grover and Leskovec 2016] capture sufficient relational information uncover meaningful groups.

The data comes primarily from OpenAlex [Priem et al. 2022], a comprehensive dataset of interconnected academic entities. The node2vec algorithm generates vector representations (embeddings) for each author, encapsulating their structural position and relationships within the network. These embeddings are then analyzed using clustering techniques (K-Means [MacQueen 1967] and Wards hierarchical clustering [Ward Jr 1963]) to identify groups of authors with similar collaborative behaviors. We also integrate decision trees with the clustering results [Gabidolla and Carreira-Perpiñán 2022] to enhance interpretability.

The results indicate that geographic location and academic productivity metrics influence collaboration ties. The largest group represents authors with average scientific production, and smaller groups highlight high-performing researchers, the latter often affiliated with top institutions or distinguished by high-impact metrics (e.g., citations, prolific publication records, earlier PhD completion). Geographically, the South and the Southeast dominate some of these groups, with the contributions of Pernambuco and Ceará, whose institutions are of international excellence, according to Capes [CAPES 2021]. These findings underscore the interplay of regional excellence and individual achievement in shaping collaboration networks, exposing disparities, and offering insights into Brazil's academic landscape.

2. Method

In the following sections, we elaborate on each phase of the methodological process, guided by the flowchart presented in Figure 1.

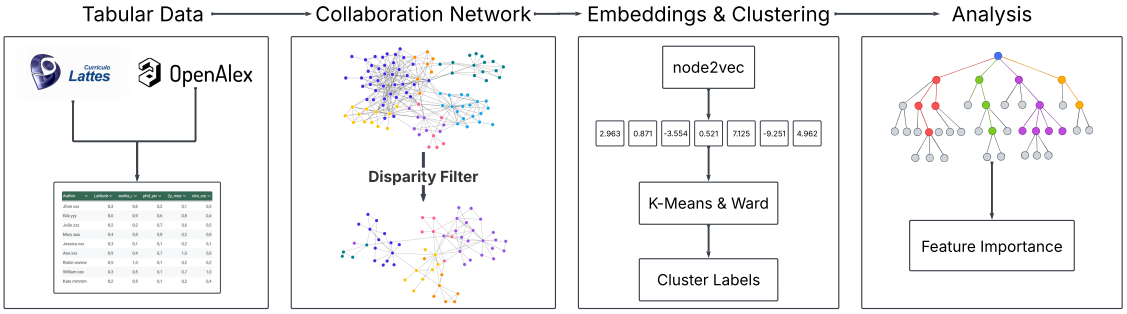

Figure 1. A systematic methodology of our study.

2.1. Data Sources

The dataset was compiled from two key sources: the Lattes Platform[1], managed by the National Council for Scientific and Technological Development (CNPq), and the OpenAlex[2] database. The Lattes Platform is a comprehensive academic and professional information repository in Brazil. At the same time, OpenAlex contains data relating to works, authors, institutions, and other entities. The list of Professors of BRGPCSs was acquired from the 2022 CAPES Evaluation Report on Brazilian Graduate Programs via Sucupira [3], and later cross-referenced with their Lattes and OpenAlex profiles.

2.2. Tabular Dataset

We developed a tabular dataset to facilitate the understanding of relationships derived from the collaboration network (described in subsequent sections). This dataset is constructed by combining data from the Lattes Platform and the OpenAlex database, resulting in a comprehensive set of 28 attributes for each author. These attributes include unique identifiers from both databases (Lattes ID and OpenAlex ID), productivity metrics (e.g., h-index, works count, citation count), and geographic information (latitude and longitude).

The dataset encompasses 1,491 researchers affiliated with 71 graduate programs. Each researcher is characterized by attributes representing their academic background, institutional affiliations, and research output. These attributes comprise basic information, academic background details, OpenAlex identifiers, productivity metrics, and geographic details. The complete dataset and further information are available in our GitHub repository[4].

2.3. Collaboration Network

A collaboration network was constructed based on co-authorship data extracted from OpenAlex, while the consolidated tabular dataset was used to enrich the analysis with additional author attributes. The process involved collecting information on all research publications over 20 years (starting from 2004). From this data, a network was formed by linking authors through their co-authorship relationships in published works.

[1] https://lattes.cnpq.br
[2] https://openalex.org
[3] https://sucupira.capes.gov.br/sucupira
[4] https://github.com/IcaroLu1z/BRGPCSGraph

We considered only publications with up to 20 authors to minimize noise and prevent the network from being artificially inflated by a few works involving hundreds of authors.

The resulting original network denoted as G, consisted of 105,168 nodes (authors) and 531,648 edges (co-authorship connections). Given the large scale of G, a backbone network G_B was extracted using the disparity filter [Ángeles Serrano et al. 2009]:

$$p_{ij} = 1 - \left(1 - \frac{w_{ij}}{s_i}\right)^{(k_i - 1)} \tag{1}$$

where $s_i = \sum_j w_{ij}$ is the total strength (sum of weights) of node i, k_i is the degree of node i (number of connections), and p_{ij} represents the probability that the weight w_{ij} is compatible with the null hypothesis, i.e., that nodes i and j are connected by chance with such weight.

Edges are retained in the backbone network G_B if their significance score p_{ij} falls below a chosen threshold α, which controls the sparsity — that is, the proportion of edges that are preserved relative to the total in the original network G — of the resulting network. By applying this filter to the original network G, the backbone network G_B retains only the most significant co-authorship connections, effectively reducing noise and highlighting the essential structural features of the network.

The parameter α for the disparity filter was optimized across 100 values within $[0.01, 1]$ to satisfy the following conditions:

$$\alpha = \underset{\alpha}{\operatorname{argmax}} \left(|V_{BRGPCS}(P_\infty^{(B)})| \right) \quad \text{subject to} \quad 0.05 \leq \frac{L_B}{L_R} \leq 0.15 \tag{2}$$

where:

- $V_{BRGPCS}(P_\infty^{(B)})$ represents the number of faculty members from BRGPCSs present in the largest connected component of the backbone network $P_\infty^{(B)}$,
- L_R and L_B are the number of edges in the original network G and the backbone network G_B, respectively,
- The ratio $\frac{L_B}{L_R}$ was constrained to lie between 5% and 15% to ensure a meaningful reduction in network size while retaining key structural properties.

It is important to highlight that no faculty member from the original dataset was excluded from the network. To ensure this, the extraction of G adhered to the condition:

$$\forall v \in G, \quad k_v^{(B)} \geq \min(k_v, 20) \tag{3}$$

where each author v in the original network G retains at least their top 20 strongest connections k_v in the backbone network G_B. This threshold was chosen to align with the maximum number of authors per paper, ensuring consistency and preserving every researcher's most significant collaboration relationships.

Through this optimization process, the value of α was determined to be 0.24. The final backbone network comprised 29,223 nodes (authors) and 67,544 edges (co-authorship connections), representing a significant reduction from the original network G. This corresponds to a reduction of 72.2% in the number of nodes and 87.3% in the

number of edges. Additionally, the largest connected component of the backbone network retained 1,463 faculty members, a slight decrease of 1.1% compared to the 1,479 faculty members present in the largest connected component of the original network.

This approach resulted in a streamlined network G_B that preserves the most relevant collaboration relationships while significantly reducing complexity. The final backbone network is the foundation for subsequent analyses.

2.4. Embedding Extraction

We used Node2vec to extract embeddings, mapping researchers into a lower-dimensional vector space that captures local and global structural properties, enriching relationship representations. The initial parameters were based on the experiments presented in the original node2vec paper [Grover and Leskovec 2016], which used 128 dimensions, a walk length of 80, 10 walks per node, 10 training epochs, a context window size of 10, and the bias parameters p and q both fixed at 1.

In this work, the walk length varied from 75 to 200 nodes, with increments of 25. Considering that the ratio between the walk length and the number of walks in the original work was 12.5%, the number of walks per node was adjusted to 25% of the walk length, that is, twice the ratio used originally.

In addition, for the extraction of embeddings with the word2vec algorithm [Mikolov et al. 2013], dimensions ranging from 2 to 32 were tested, with increments of 4 after 4 dimensions. The window size was fixed at 10, with 50 training epochs, an initial learning rate of 10^{-3}, and a final rate of 10^{-4}.

2.5. Cluster Analysis and Results Interpretation

To identify groups of researchers with similar collaboration patterns, the K-means [MacQueen 1967] and Ward's hierarchical clustering [Ward Jr 1963] algorithms were applied to the generated embeddings. Both algorithms are widely used for unsupervised learning and offer complementary approaches to cluster analysis: K-means partitions data into k clusters based on distance to centroids, while Ward's method builds a hierarchy of clusters based on minimizing variance within each cluster. The number of clusters k was systematically varied from 3 to 10 to explore different levels of granularity in the grouping of researchers. This range was chosen to allow for broad- and fine-grained analyses of the collaboration network.

To evaluate the quality of the clustering results, we employed three widely recognized metrics:

- **Silhouette Score [Rousseeuw 1987]:** Measures how similar an object is to its own cluster compared to other clusters. It ranges from -1 to 1, where higher values indicate better-defined clusters. This metric emphasizes intra-cluster similarity and was the primary criterion for assessing clustering performance.
- **Calinski-Harabasz Index [Caliński and Harabasz 1974]:** Evaluates cluster separation by computing the ratio between-cluster variance to within-cluster variance. Higher values indicate better separation between clusters. While useful, this metric was secondary in our analysis due to its focus on separation rather than intra-cluster cohesion.

- **Davies-Bouldin Index [Davies and Bouldin 1979]:** Measures the average similarity ratio of each cluster with the most similar cluster. Lower values indicate better clustering, as clusters are more distinct and compact.

The Silhouette Score was used as the main metric for cluster quality, while the Calinski-Harabasz Index and Davies-Bouldin Index were used as supporting metrics.

Due to the large volume of results with extremely low values, only those with a Silhouette score exceeding 0.5 were considered for further analysis. This threshold ensured that only meaningful and interpretable clusters were retained for subsequent exploration.

After labeling each node with its corresponding cluster, decision trees were also utilized to assess feature importance, following the methodology proposed by [Gabidolla and Carreira-Perpiñán 2022]. These methods were chosen for their ability to identify the most relevant features for distinguishing between different groups, providing clear insights into the factors influencing cluster formation.

Three main metrics were calculated to assess the contribution of each attribute in separating the groups. The first was *Feature Importance*, which measures the relative weight of each feature in the model, highlighting the most impactful variables in differentiating the clusters. The second was *Permutation Importance*, which evaluates the effect of randomly shuffling each feature on the model's predictions, allowing us to identify which features have the greatest influence on overall performance. Finally, the SHAP Score (*Shapley Additive Explanations*) [Lundberg and Lee 2017] was used, which provides a marginal importance measure for each feature, enabling a detailed and localized analysis of the individual impact of each attribute on the clusters.

3. Results and Discussion

3.1. Clustering Results

The results show that the K-means algorithm achieved the highest scores across all metrics. However, some of these values were identified as outliers, casting doubt on their reliability. In contrast, the Ward method exhibited more consistent performance with no significant outliers. While Ward's method scored slightly lower on the Calinski-Harabasz Index — a metric focused on cluster separation — this aspect is less relevant for analyzing author relationships. On the other hand, the Silhouette score emphasizes intra - cluster similarity, which aligns better with our study's objectives and thus serves as a more suitable metric for evaluating clustering outcomes.

To ensure robustness and consistency in our analysis, we selected four models generated using the Ward method. These models correspond to cluster configurations with 3, 4, 5, and 6 groups, each chosen based on achieving the highest Silhouette scores within their respective group sizes as shown in Table 1. These models will serve as the foundation for exploring relationships between authors and graduate programs in Brazil.

3.2. Feature Importance

When applying decision tree models to analyze the clusters, it was observed that two attributes consistently stood out as the most important: latitude and the two-year mean citedness (`2_yr_mean_citedness`). These attributes were repeatedly identified as the primary factors responsible for separating and defining the groups, as shown in Figure 2.

Table 1. Parameters of the models selected for analysis. The columns Dim, Walk Lenght, # Walks correspond to node2vec parameters, and C-H and D-B are the Calinski-Harabasz and Davies-Bouldin clustering indexes.

Clusters	Dim	Walk Length	# Walks	Silhouette	C-H Index	D-B Index
3	2	100	25	0.72	3291	0.42
4	2	150	38	0.70	3267	0.52
5	2	200	50	0.61	2990	0.56
6	2	175	44	0.52	3182	0.61

These findings indicate a strong relationship between geographic location and academic impact on researchers' collaboration behavior. We present the attributes used for analysis in Table 2.

Table 2. Description of attributes used in the analysis, grouped by production metrics and geographic attributes.

Attribute	Description
works_count	Number of publications authored by the researcher.
cite_count	Number of citations received across all publications.
2yr_mean_citedness	Average number of citations per publication over the last two years.
i10_index	Number of publications with at least 10 citations.
h_index	A metric that measures both the productivity and citation impact of a researcher's publications.
phd_year	Year in which the author completed their PhD.
productivity_grant	Whether the author holds a productivity grant.
productivity_grant_type	Type of productivity grant held by the author, if any.
gp_score	Graduate Program score assigned by CAPES.
latitude	Geographic coordinate indicating the north-south position of the author's institutional affiliation.
longitude	Geographic coordinate indicating the east-west position of the author's institutional affiliation.
region	The broader geographic region where the author's institution is located.
state	The Brazilian state where the author's institution is located.

Additionally, other variables, such as the total number of publications (works_count) and the year of PhD completion (phd_year), also frequently appeared among the most relevant attributes, albeit with less importance compared to latitude and 2_yr_mean_citedness. These results suggest that researchers' academic trajectories and productivity also influence group formation, although to a lesser extent. The combination of these characteristics reflects the complex dynamics shaping collaboration networks, highlighting that, in addition to geographic proximity, academic output, and citation-based impact play central roles in structuring these networks.

The analysis of Permutation Importance revealed significant consistency in the most important attributes for separating the groups. Across all evaluated models, the four

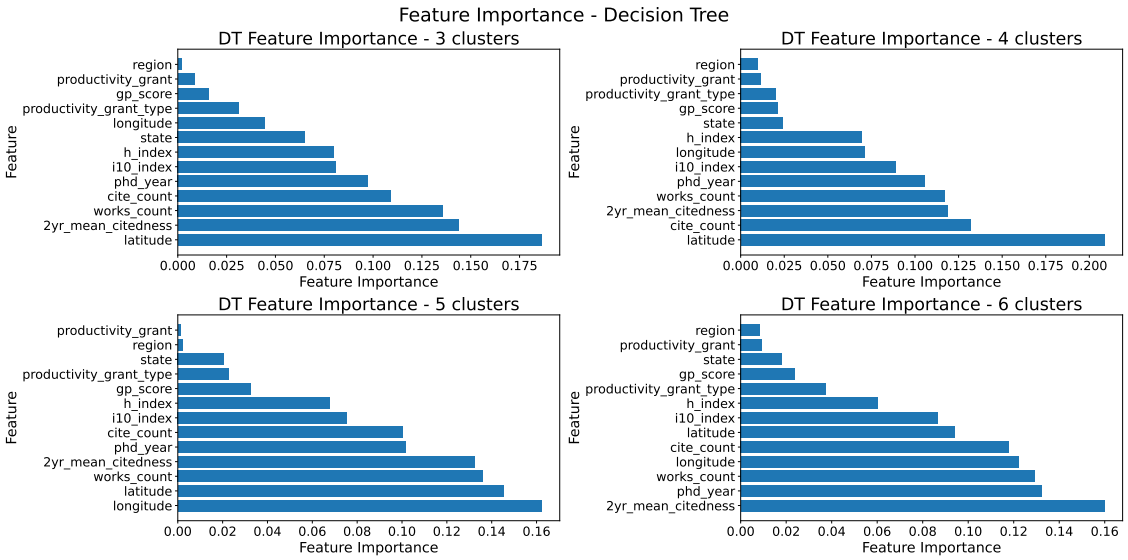

Figure 2. Feature Importance based on Decision Trees.

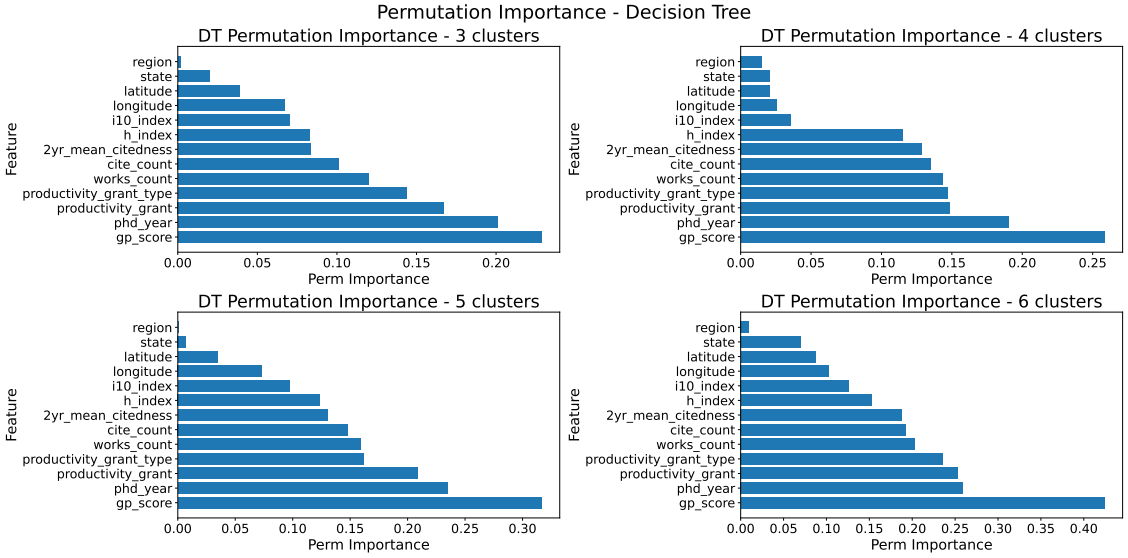

Figure 3. Permutation Importance based on Decision Trees.

most relevant attributes were the graduate program score (`gp_score`), the PhD completion year (`phd_year`), whether the author holds a productivity grant (`productive_grant`), and the type of productivity grant (`productive_grant_type`), as shown in Figure 3.

In particular, the `gp_score` stood out with an importance margin up to 0.15 higher than the second most relevant attribute, `phd_year`. This substantial difference highlights the predominant influence of the level of the Graduate Program the author is affiliated with in structuring the groups within the collaboration network. The other three attributes—related to academic productivity and types of funding—also demonstrated consistent importance, reinforcing the idea that academic trajectory and financial support are central factors in shaping collaborative patterns.

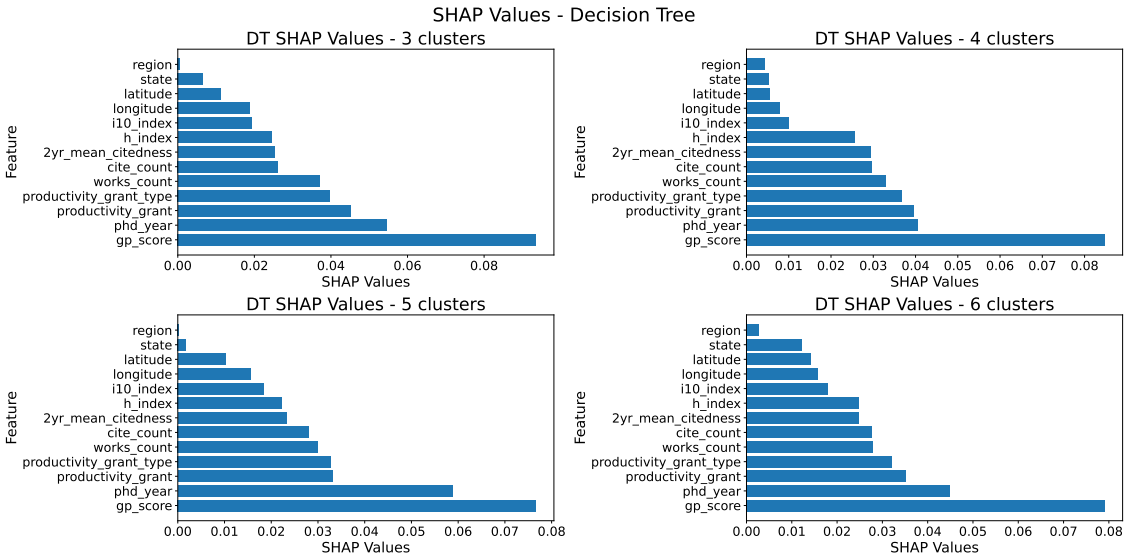

Figure 4. SHAP values based on Decision Trees.

The analysis of SHAP Values for decision tree models further supports the findings of the Permutation Importance analysis. The values confirm the significant impact of academic productivity metrics and institutional factors. Features like `gp_score`, `phd_year`, `productivity_grant`, and `productivity_grant_type` show substantial importance. The `gp_score`, in particular, stands out as a dominant factor, with its SHAP values indicating a strong influence on cluster formation. This consistency across both SHAP and Permutation Importance analyses underscores the pivotal role of graduate program scores in characterizing researchers within the collaboration network.

3.3. Quantitative Analysis

Based on the results from decision trees, as well as attribute importance analyses, we identified the factors that most significantly influence the definition of the groups. Using this information, we quantitatively analyzed these attributes to understand better how the groups were formed, revealing patterns and characteristics that guided the clustering process.

The analysis of attribute averages within clusters, presented in Table 3, reveals the presence of a significantly larger group compared to the others, suggesting the existence of a "base" class that encompasses most of the authors analyzed. This group can be interpreted as representative of average scientific production, where attributes such as the average number of works and two-year citation metrics reflect typical productivity levels and impact for these authors.

In contrast, smaller groups emerge as clusters of authors who stand out, generally in a positive manner, with higher averages in both the number of works and the `2yr_mean_citedness` metric. This difference suggests that these authors achieve higher academic performance, potentially due to factors such as affiliation with prestigious research institutions or involvement in high-visibility fields of study. Thus, the analysis not only identifies patterns in the distribution of authors but also allows us to infer the existence of a spectrum of academic performance, ranging from a base class to

Table 3. Average of Main Attributes for each cluster N.

N	Size	2yr_m_cite	gp_score	phd_year	works_count	latitude	longitude
Model 1 - 3 clusters - Silhouette Score: 0.72							
0	1156	1.26	4.89	2007.69	100.54	-17.11	-43.67
1	150	1.59	5.15	2007.37	136.81	-24.05	-47.98
2	183	1.80	5.16	2004.79	119.47	-20.43	-45.52
Model 2 - 4 clusters - Silhouette Score: 0.70							
0	1117	1.22	4.94	2007.52	101.65	-17.30	-43.92
1	123	1.59	5.13	2008.06	102.00	-25.25	-48.17
2	165	2.05	4.84	2005.46	112.79	-15.48	-41.66
3	84	1.53	5.11	2006.82	165.45	-25.45	-49.37
Model 3 - 5 clusters - Silhouette Score: 0.61							
0	1011	1.23	4.98	2007.45	101.42	-18.57	-44.97
1	211	1.55	4.76	2007.14	103.94	-11.84	-39.28
2	101	1.71	5.09	2006.70	154.81	-25.64	-48.52
3	64	2.27	4.69	2006.20	143.19	-12.78	-40.00
4	102	1.29	5.13	2007.37	91.56	-23.92	-46.98
Model 4 - 6 clusters - Silhouette Score: 0.52							
0	813	1.23	4.93	2007.35	98.57	-18.87	-44.83
1	148	1.40	5.52	2006.28	101.13	-23.05	-47.06
2	136	1.55	4.73	2007.18	112.19	-10.06	-38.16
3	220	1.43	4.70	2008.29	111.11	-12.49	-41.52
4	80	1.53	4.89	2005.29	149.18	-23.60	-47.30
5	92	1.81	5.18	2008.09	128.95	-25.72	-48.69

those who excel in the production and impact of their research.

As previously discussed, geographic data played a significant role in forming the clusters. This influence becomes evident when analyzing the mean and median values of the authors' geographic coordinates, as shown in Figure 5. The averages of latitude and longitude exhibit substantial variations across the generated groups, reinforcing the central role of geographic location in group segmentation and highlighting relevant regional differences for the analysis.

Furthermore, when examining the predominant states and regions within each group, as shown in Figure 6, it was observed that the South and Southeast regions are often grouped into distinct clusters. Additionally, in models with a higher number of groups, part of the Northeast region clusters with the Southeast, particularly authors located in the states of Pernambuco and Ceará. This characteristic may be attributed to the fact that the Federal University of Pernambuco is currently an international reference in graduate programs in computing in Brazil, alongside other universities in the Southeast, such as USP, UFMG, UFRJ, and Unicamp.

By integrating the analysis of Table 3 with Figure 6, intriguing interactions between the formed groups become apparent. Specifically, in both Model 2 and Model 3,

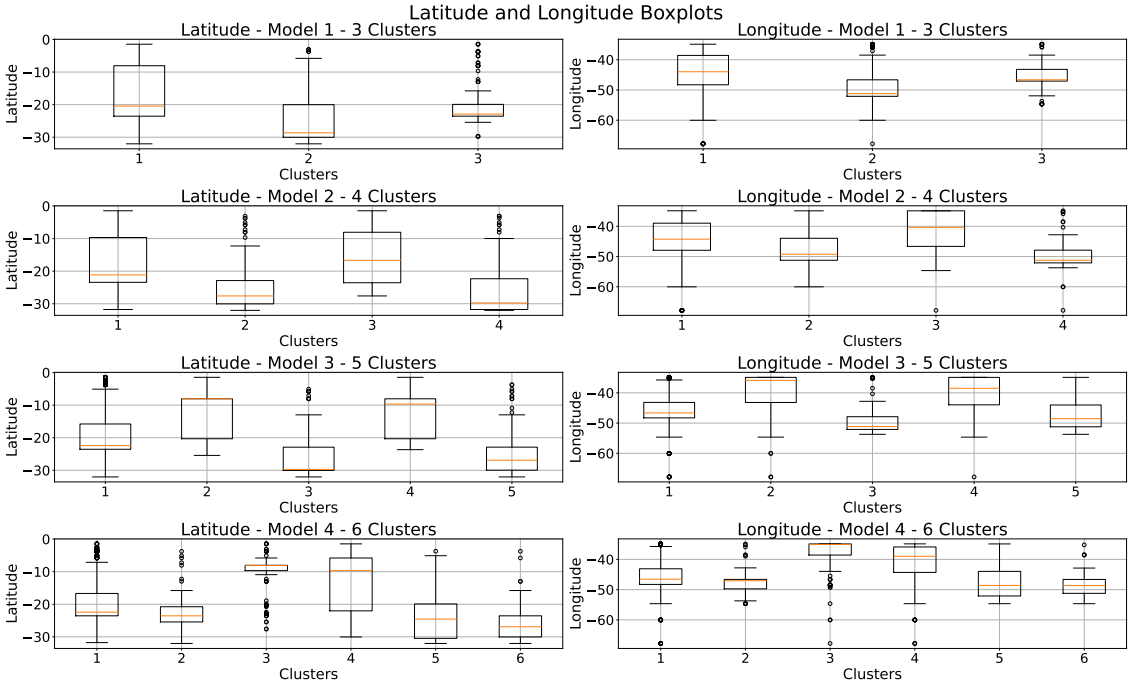

Figure 5. Boxplots for Latitude and Longitude in each cluster.

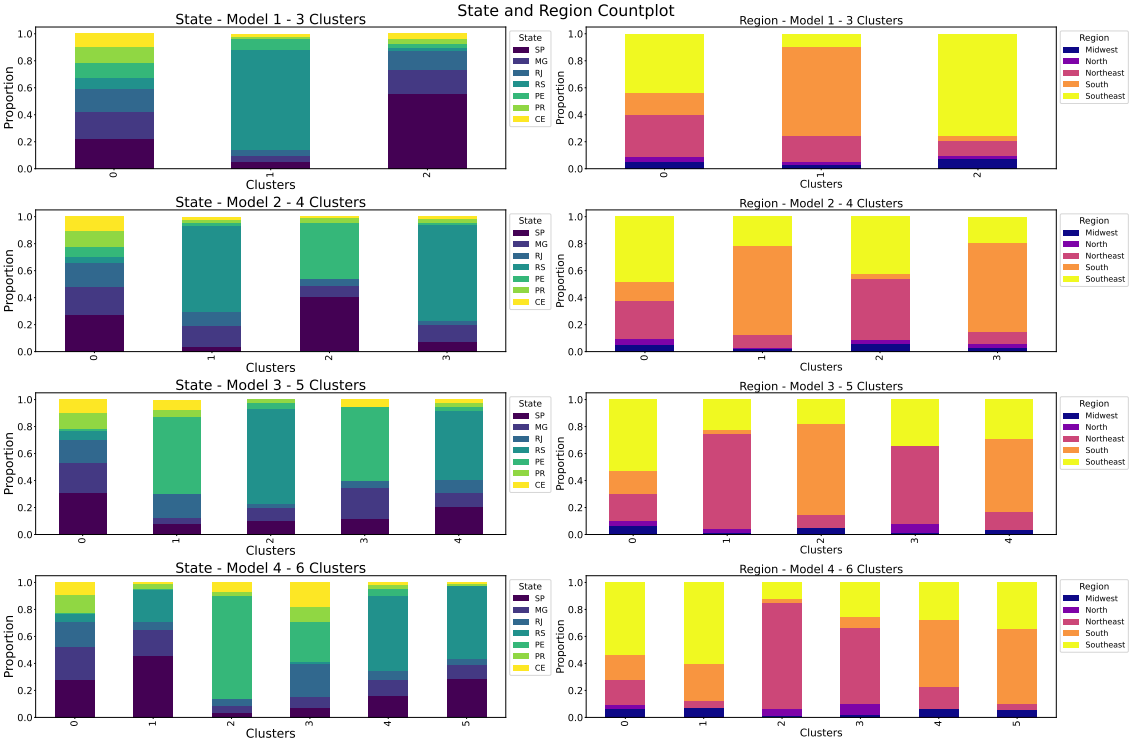

Figure 6. Predominant country states and regions within each cluster.

the groups with the highest averages of `2yr_mean_citedness` show a significant presence of authors from Pernambuco. This suggests that researchers from this region may produce work attracting greater attention and recognition within the scientific community.

Analyzing the number of publications, it is also notable that the groups with the highest averages are predominantly composed of authors from Rio Grande do Sul. This discrepancy indicates a possible specialization or higher productivity among researchers in this region, setting them apart from others.

3.4. Discussion

To comprehensively address the hypotheses outlined in the introduction, this study employed a systematic methodology leveraging network embeddings, clustering techniques, and interpretability methods such as decision trees and SHAP values.

For Hypothesis 1, which posits that geographic proximity significantly influences collaboration patterns, we analyzed geographic attributes such as latitude, longitude, and regional affiliations using feature importance metrics. The results consistently confirmed that geographic location plays a central role in shaping collaboration networks. As shown in Figure 2, latitude and longitude were among the most important features driving cluster separation. Furthermore, Figures 5 and 6 illustrate substantial variations in geographic coordinates across clusters, underscoring regional disparities and hubs of excellence.

Hypothesis 2 explores the impact of academic productivity metrics on group formation within the collaboration network. We found that high-performing researchers were consistently clustered based on their superior productivity by examining attributes such as citation counts, h-index, and works count. Table 3 reveals that smaller groups exhibited higher averages in metrics `2yr_mean_citedness` and `works_count`, suggesting that academic output and citation-based impact are critical for distinguishing high-performing researchers. Decision trees and SHAP analyses further corroborated these findings, as shown in Figures 3 and 4, where `2yr_mean_citedness` emerged as one of the most influential features.

For Hypothesis 3, which investigates the influence of institutional factors on collaboration, we focused on attributes such as `gp_score`, productivity grants, and grant types. Permutation importance and SHAP analyses demonstrated that `gp_score` was the most influential attribute, with an importance margin significantly higher than other features (Figure 3). Attributes related to productivity grants and grant types also consistently appeared as key factors, reinforcing that institutional support and funding play a central role in shaping collaboration patterns. As illustrated in Figures 3 and 4, these institutional factors were pivotal in distinguishing clusters.

Finally, Hypothesis 4 examines whether node2vec embeddings can effectively capture relational information to enable meaningful clustering. Using node2vec embeddings alongside clustering algorithms, we successfully identified distinct groups with high Silhouette scores. While K-means achieved the highest metric values, it exhibited outliers, raising concerns about reliability. In contrast, Ward's method provided consistent and interpretable results, making it the preferred choice for detailed cluster analysis. Table 1 summarizes the parameters and performance metrics of the selected models.

4. Conclusion

This study investigated the structure of collaboration networks within Brazilian Graduate Programs in Computer Science (BRGPCSs), leveraging network embedding

techniques, clustering algorithms, and interpretable machine learning methods. Our primary goal was to identify the key factors shaping collaboration patterns among researchers, focusing on the roles of geographic location, academic productivity (measured by citation metrics and publication counts), and institutional factors. Our findings show that these factors significantly influence the formation of distinct groups within the co-authorship network and that network embeddings (specifically, node2vec) effectively capture these relationships. Overall, our contributions are: (i) we demonstrate that network embeddings effectively capture the structural dynamics of co-authorship networks; (ii) we identify key factors that shape collaboration patterns — particularly academic productivity, institutional features such as graduate program ratings, and geographic location; and (iii) we show that high-performing researchers tend to form smaller, more isolated clusters, distinguishing themselves from the larger groups with average performance.

As natural extensions of this research, we propose exploring more advanced deep learning models, especially Graph Neural Networks (GNNs), which have shown great potential in modeling complex structures in graph data. Architectures such as Graph Convolutional Networks (GCN) [Zhou et al. 2020], Graph Attention Networks (GAT) [Zhou et al. 2020] and GraphSAGE [Hamilton et al. 2018] could be applied to better capture subtle and context-aware collaboration patterns that go beyond static embedding limitations. Future studies could also incorporate temporal information and research topics to enable dynamic and thematic analyses of collaboration behavior.

Funding

The authors would like to thank the Fundação de Amparo a Pesquisa do Estado de Minas Gerais (FAPEMIG, grants APQ-01518-21, APQ-01647-22), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, grants 307151/2022-0, 308400/2022-4), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) and Universidade Federal de Ouro Preto (PROPPI/UFOP) for supporting the development of this study.

References

Bai, X., Zhang, F., Li, J., Xu, Z., Patoli, Z., and Lee, I. (2021). Quantifying scientific collaboration impact by exploiting collaboration-citation network. *SCIENTOMETRICS*, 126(9):7993–8008.

Caliński, T. and Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27.

CAPES (2021). Relatório de avaliação 2017-2020: Quadrienal 2021. Acesso em: 20 out. 2023.

Davies, D. L. and Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227.

di Bella, E., Gandullia, L., and Preti, S. (2021). Analysis of scientific collaboration network of italian institute of technology. *Scientometrics*, 126(10):8517–8539.

Druszcz, F. F. and Vignatti, A. L. (2024). Citation analysis disparity between sub-areas of brazilian computer science. In *Brazilian Workshop on Social Network Analysis and Mining (BraSNAM)*, pages 24–34. SBC.

FitzGerald, C., Huang, Y., Leisman, K. P., and Topaz, C. M. (2023). Temporal dynamics of faculty hiring in mathematics. *Humanities and Social Sciences Communications*, 10(1):247.

Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., Milojević, S., Petersen, A. M., Radicchi, F., Sinatra, R., Uzzi, B., Vespignani, A., Waltman, L., Wang, D., and Barabási, A.-L. (2018). Science of science. *Science*, 359(6379):eaao0185.

Gabidolla, M. and Carreira-Perpiñán, M. (2022). Optimal interpretable clustering using oblique decision trees. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 400–410, New York, NY, USA. Association for Computing Machinery.

Grover, A. and Leskovec, J. (2016). node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864.

Hamilton, W. L., Ying, R., and Leskovec, J. (2018). Inductive representation learning on large graphs.

Liu, J., Xia, F., Wang, L., Xu, B., Kong, X., Tong, H., and King, I. (2021). Shifu2: A network representation learning based model for advisor-advisee relationship mining. *IEEE Transactions on Knowledge and Data Engineering*, 33(4):1763–1777.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In Cam, L. M. L. and Neyman, J., editors, *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. University of California Press.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. In Bengio, Y. and LeCun, Y., editors, *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Priem, J., Piwowar, H., and Orr, R. (2022). Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244.

Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., and Sun, M. (2020). Graph neural networks: A review of methods and applications. *AI Open*, 1:57–81.

Ángeles Serrano, M., Boguñá, M., and Vespignani, A. (2009). Extracting the multiscale backbone of complex weighted networks. *Proceedings of the National Academy of Sciences*, 106(16):6483–6488.