Scientific Collaboration Network Views: A Brazilian Computer Science Graduate Programs Case

Aurelio Ribeiro Costa, Vanessa Tavares Nunes, Célia Ghedini Ralha

Department of Computer Science, University of Brasilia, Brasília, Brazil
arcosta@gmail.com, vanunes@gmail.com, ghedini@unb.br

Abstract. Scientific collaboration networks can present different views of researchers' interactions. This work presents SCI-synergy, an online navigable artifact aiming to promote mechanisms and views of scientific collaboration networks. The artifact focuses on the researchers' interaction in the co-authorship of publications considering intra- and inter-program relationships. SCI-synergy is developed upon the design science research paradigm using scientific publication data available on the large Digital Bibliography & Library Project (DBLP) repository. Official data from the Sucupira repository of six Brazilian graduate program members including Federal University of Minas Gerais (UFMG), State University of São Paulo (USP), Federal University of Rio Grande do Norte (UFRN), Federal University of Amazonas (UFAM), University of Brasília (UnB), and University of Vale do Rio dos Sinos (UNISINOS) is used. Data from these graduate programs illustrate the artifact usage regarding the scientific collaboration network of each program, how each researcher cooperates, and what relationship patterns exist in intra- and inter-programs views. We advocate that, even though it is necessary to consider data from each program's history and current contextualization regarding politics, economics, and administration, the collaboration network views provided by SCI-synergy might help to understand collaboration network patterns.

Categories and Subject Descriptors: Information Systems [**Information Systems Applications**]: Information and Data Management

Keywords: design science research, digital libraries, scientific interactions, social networks

1. INTRODUCTION

Researchers frequently collaborate internally, with their institution members, and externally. Information sharing, joining of competencies, and the union of researchers' efforts to pursue common goals have boosted the production of knowledge [Zhang et al. 2018; Hu et al. 2020], and the association of different points of view may generate new perspectives in research [Coxon 1979]. Then, combining technology and connection facilities might increase the number of collaborations among individuals geographically dispersed. These factors contribute to the current appreciation for researchers and universities capable of forming productive scientific collaboration networks. As the number of studies and scientific publications increases, besides other specificities the co-authoring of scientific publications is a strong indicator of collaboration [Katz and Martin 1997; Newman 2004; Cronin 2005; Hu et al. 2020].

Co-authorship studies cover many collaboration characteristics by using different disciplines and considering researchers from several institutions or countries [Glänzel 2001; Tomizawa and Shirabe 2002; Yoshikane and Kageura 2004; Kim 2006; Cugmas et al. 2020; Hu et al. 2020]. These studies confirm that authors' collaboration has increased in science and demonstrate that it has substantial impact and visibility what leads to increasing researchers' productivity [Glänzel 2002]. Regarding inter-institutional collaborations, it boosts the flow of knowledge and technology between researchers of an innovation system [Chesbrough et al. 2006; Park and Nam 2016].

This work focuses on Computer Science graduate programs case associated with Brazilian universities. Brazilian Computer Science has long studied their graduate programs as presented in a seminal work [Laender et al. 2008] and posterior works [Arruda et al. 2009; Digiampietri et al. 2014; Linden et al. 2017]. Considering collaboration networks, the work of [Lopes et al. 2012] uses a quantification method based on the Gini coefficient to analyze scientific collaboration in research networks. Our previous works ([Costa and Ralha 2015; 2019]) present a collaboration network analysis as a management tool using publication data available in the Lattes Platform from one Computer Science graduate program in Brazil, including a recommendation module for scientific partnerships. Besides, many international works are leading to compelling findings as in [Franceschet 2011] that explores how collaboration in Computer Science evolved since 1960, measuring influences of scientist groups based on multiple types of collaboration [Jiang et al. 2017], and a cross-disciplinary research work presenting a survey of scientific teamwork collaboration [Yu et al. 2019]. Nevertheless, there is still space for contribution in the Brazilian Computer Science graduate programs concerning intra- and inter-scientific collaboration analysis, without focusing on evaluating research or ranking research groups.

In this context, we present SCI-synergy online navigable artifact that allows scientific collaboration network analysis from the point of view of universities and researchers. SCI-synergy is established over the Design Science Research (DSR) paradigm as an important methodology for the development of information systems. The source data used comes from six Computer Science graduate programs linked to Brazilian universities. The Brazilian Sucupira platform assembles information about the graduate program members. The co-authorship relation in publications was indexed by the Digital Bibliography & Library Project (DBLP) [Ley 2002]. The graduate programs used as cases are located in each geographic region in Brazil illustrating the variety of characteristics of the country's research in the area: Federal University of Minas Gerais (UFMG), State University of São Paulo (USP), Federal University of Rio Grande do Norte (UFRN), Federal University of Amazonas (UFAM), University of Brasília (UnB) and University of Vale do Rio dos Sinos (UNISINOS).

The main goal of this work is to allow different views of the scientific collaboration network focusing on the researchers' and universities' interactions. Therefore, the following research questions (RQ) were defined: *RQ1*) How is the scientific collaboration network of each graduate program? *RQ2*) How cooperative is each researcher in the graduate program? *RQ3*) What are the intra- and inter-program collaboration patterns? We use social network metrics as *average degree*, *closeness* and *betweenness centralities*, and *connected components* (union-find or community detection) to highlight the intra- and inter-scientific collaboration among these networks providing mechanisms to answer these questions.

The remainder of the paper includes the research approach section with aspects of the design science research paradigm. The scientific collaboration network analysis section presents the answers to the research questions with the Brazilian Computer Science graduate programs case. The literature review is described in the related work section, and the final remarks close this paper.

2. RESEARCH APPROACH

This work was conducted under the DSR paradigm [Weber 2010]. DSR is derived from the engineering discipline and depicts the science of the artificial that bridges practice and theory through a conceptual framework to research in information system [Simon 1996]. DSR has proven to produce practically relevant research results aiming to ensure rigor and relevance in prototyping research projects [Hevner and Chatterjee 2010]. In the following sections, we present the environment where this work was conducted, followed by the knowledge base describing the fundamentals and social metrics used. The final section describes the artifact construct, including the architecture, the conceptual data model, and the operationalization process.

2.1 Environment

The environment in the DSR paradigm defines where the information system research is being applied and to whom it wants to provide results and benefits. This work focuses on graduate programs of Brazilian universities. Therefore, the researchers' information was captured from the Sucupira platform and linked to the publications' authorship in the DBLP digital repository.

Although this work focuses on collaboration, it is worth noticing that in Brazil, every four years, the graduate programs have their research and education quality assessed by a public agency within the Brazilian Ministry of Higher Education named Brazilian Coordination for the Improvement of Higher Education Personnel (CAPES).[1] Thus, CAPES is responsible for graduate programs' expansion and consolidation in the country. Some of the regular assessment metrics are related to the intellectual production and intra- and inter-program collaboration used to compose each program's level (from three to seven). Therefore, one of the Brazilian programs' research investments focuses on increasing those numbers considering the high impact of the publication. Regarding the Computer Science graduate programs selected as study case, Table I presents six universities in different regions and states of Brazil, considering four CAPES levels. The intention was to cover the socio-economic differences in the country's regions with a possible impact on the level of courses offered. Also, more traditional programs in Computer Science were included, such as the master's course at UFMG created in 1974 and doctorate in 1990, and MSc/Ph.D. in 1997 at USP. In contrast to these traditional programs, more recent doctoral courses were included at UFRN/UFAM in 2008, UnB in 2010, and UNISINOS in 2015. Considering the researchers in programs (see Table II), we selected programs with level five in CAPES, i.e., UFAM - 28 researchers, UFRN - 29 researchers, and UnB - 31 researchers, with similar size to USP with 34 researchers (level six) forming collaborative networks with interesting aspects for analysis. This information is public on the Sucupira platform.[2]

Table I: The Computer Science graduate programs selected as study case.

Brazilian Region	Brazilian State	University Name	CAPES Level
South	Rio Grande do Sul	University of Vale do Rio dos Sinos (UNISINOS)	four
Southeast	Minas Gerais	Federal University of Minas Gerais (UFMG)	seven
Southeast	São Paulo	State University of São Paulo (USP)	six
Central-West	Federal District	University of Brasília (UnB)	five
North	Amazonas	Federal Univeristy of Amazonas (UFAM)	five
Northeastern	Rio Grande do Norte	Federal University of Rio Grande do Norte (UFRN)	five

2.2 Knowledge Base

Knowledge base relates to the definition and discussion of the fundamentals and methodologies adopted in the research presented in the sequence.

2.2.1 *Social Networks.* The scientific study of networks is broadly interdisciplinary, including computer networks, social networks, and biological networks [Newman 2010]. Social networks have received enormous interest in the last few years using technology in platforms such as Facebook, Instagram, and others, having breakthroughs in the social sciences field.

According to graph theory, a social network can be represented by a graph $G = (V, E)$, where V is the set of vertices that denotes individuals under consideration, and E are the edges corresponding to a set of existing relationships between these individuals (e.g., friendship, parenthood or professional

[1] http://www.capes.gov.br/

[2] https://sucupira.capes.gov.br/sucupira/public/consultas/coleta/programa/quantitativos/quantitativoIes.xhtml?areaAvaliacao=2&areaConhecimento=10300007

collaboration). Relationships can have different intensities that reflect the strength of social connections. The relationship intensity is usually represented by a function $w(e)$ with $e \in E$ that associates a weight as an edge property in the graph. In essence, scientific collaboration networks are social networks where relationships represent a scientific interaction type. In this work, we are interested in a specific class of network, in which vertices correspond to authors of scientific publications and the edges (with weights) represent the intensity of collaboration between two authors through the number of articles written together. There are online platforms that use this representation, such as ResearchGate, including over 15 million registered researchers and 118 million publications. But such platforms do not provide answers to the RQs that might interest Brazilian Computer Science graduate programs specifically considering views of the scientific collaboration network of each graduate program or the intra- and inter-program collaboration patterns.

2.2.2 *Network Metrics.* With a network, one can compute a range of measures to assess particular features of the network structure. Some social network metrics first introduced assessing aspects of the network relationships are now in wide use in many other areas [Newman 2010]. In this work, we assume the network centrality concepts according to [Bonacich 1987], and we use the following well-known metrics to support the analysis of co-authorship relationships.

Average degree or *degree of vertices* computes the average number of relationships of each node in the network by the sum of adjacent edges divided by the number of vertices. It highlights how connected the network is. A low *average degree* may denote many isolated nodes, while, a high degree tells that there are many relationships among nodes (i.e., high collaboration). Although there are no conclusive studies, in [Martin and Niemeyer 2021] the authors present empirical evidence demonstrating that average degree can explain the robustness of networks regardless of their sizes. Specifically, they tested that average degree is a more robust centrality measure when compared to Eigenvector and PageRank, even considering different levels of errors (e.g., duplicated nodes, erroneous removal of nodes or vertices).

Connected components or *union-find* describes disjoint sets of connected nodes in the network detecting smaller research groups or communities. This measure gives us a picture of how big and how many communities exist in the network. These communities can be inside a graduate program (intra-program view) or among graduate programs from different universities (inter-program view).

Betweenness centrality calculates the shortest path through a network between two vertices (geodesic path). It shows researchers that can act as "bridges" between nodes in a network, which is a way to find members who can influence the flow around a system. This centrality measure is not to be used isolated from other contexts but might point to specific nodes (researchers) of the network considered influencers or some kind of authority over social relations among researchers.

Closeness centrality calculates the shortest paths between all nodes. Nodes with a high closeness score have the shortest distances to all other nodes being useful to find researchers best placed to influence the entire network quicker. This measure calculates how close each person is to the other people in the network. A low closeness centrality means that a person is directly connected or just "a few steps" from most nodes in the network. In contrast, vertices in very peripheral locations may have high closeness centrality scores, indicating the high number of steps or connections they need to take to connect to distant others in the network [Hansen et al. 2020].

2.2.3 *Name Disambiguation.* In the scientific social networks context, one challenge is to check, automatically, whether authors correspond to the same individuals [Kang et al. 2009; Ferreira et al. 2020]. We applied traditional techniques available at the DBLP that include homonyms and synonyms treatment. In DBLP, different authors with the same name (including special characters and punctuation) are homonyms. But in DBLP, authors are assigned a unique key, and their names are distinguished in the database by a unique numerical suffix. At the moment, the splitting of existing

DBLP author publications is either triggered by authors requests or if the DBLP team can prove that there are several persons behind an entry.

Although DBLP team has been working on this for years, unfortunately, in some cases, homonyms remain undetected, needing further investigation. There are many reasons why several author names are considered to be synonymous with a particular author: name changes, nicknames, sporadic use of middle names, missing or abbreviated name parts, or even pseudonyms. Different spelling, misspelling, or mistranslation are also causes of author name synonym [Tang and Walsh 2010]. When multiple versions of a name are frequently used on publications, they may be included as aliases to the DBLP data set. Thus, we used those aliases in the SCI-synergy artifact. Since there is not a significant amount of problems within name disambiguation in our dataset, some were solved manually, and we rely on DBLP efforts to treat the others, in general.

2.3 Artifact Construct

SCI-synergy architecture is presented in Figure 1. It is composed of three modules: (1) data source (input) with researchers' information (authors name) from CAPES Sucupira,[3] and authors' publications from the DBLP data source eXtensible Markup Language (XML); (2) database storage, which consolidates all data collected from the other two repositories (Sucupira and DBLP); and (3) collaboration network visualization (output), which presents the scientific collaboration network views with intra- and inter-program, researchers' information with collaboration chart and publication histogram, and also presents social network metrics based on [Brandes and Pich 2007].

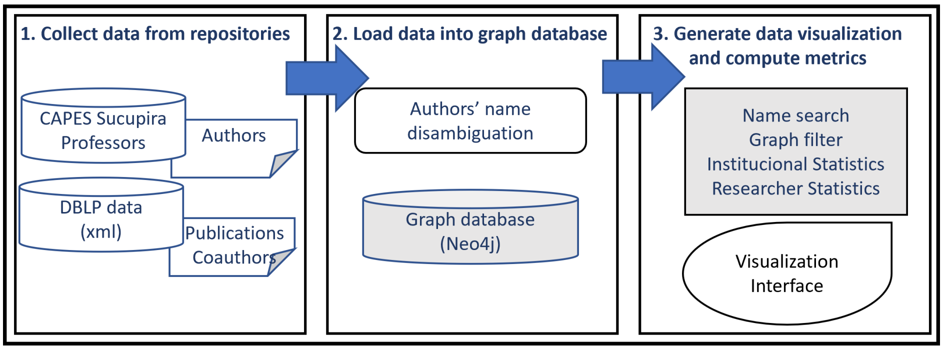

Fig. 1: The SCI-synergy architecture.

The operationalization process is presented in Figure 2. The process begins with researchers' data collected from CAPES's Sucupira platform in a CSV file as input. This initial data (containing only the author's name) is augmented with the synonyms used by DBLP.[4] These initial authors are called "seeds" since we start from their publications to identify their co-authors. The "seeds" and their related data are loaded onto the graph database according to the conceptual data model presented in Figure 3. One important aspect to note is that the researcher's name in the Sucupira platform should match any name used by this researcher in DBLP. However, regardless of all efforts to automate the process, some manual verification must be done to identify mismatches in the researchers' names.

In parallel, publications data is retrieved from the DBLP repository. A DBLP compressed XML archive is obtained through the DBLP URL.[5] This data source originated from a database dump weekly produced by the DBLP. Thus, publications are loaded onto the graph database if at least

[3]Available at `https://dadosabertos.capes.gov.br/dataset`
[4]An illustration of how DBLP handles synonyms can be found in `https://dblp.org/faq/How+does+dblp+handle+homonyms+and+synonyms.html`
[5]`http://dblp.uni-trier.de/xml`

one author was loaded before, worthing the "seeds" use. The DBLP authors' name disambiguation is carried out, and the scientific collaboration network is defined.

Finally, the graph visualization of the scientific collaboration networks is generated, and the social network metrics computed. Subsequently, the scientific social network graph is ready to be visualized through the online interface. The graph presentation's purpose is to allow the visualization of the relationships between the researchers and how connected (the edge weight) they are with other groups. With the final database, several information and statistics focusing on researchers' collaboration are collected and presented.

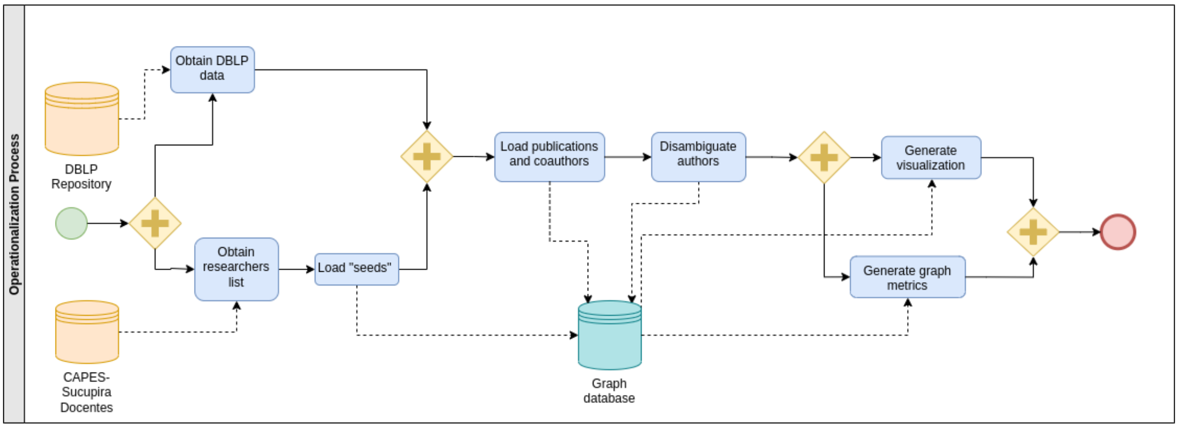

Fig. 2: Operationalization process.

The conceptual data model used in the graph NoSQL database is depicted using a directed graph in Figure 3. The centered vertex «Author» represents the authors of each published article that uses the *Write* edge to relate to their respective articles («Article» vertex). This uses the *Publish* edge to relate to the «Journal» vertex. The author relationship with their university was modeled through the «Institution» vertex through the *Associated* edge. Its relation to the graduate program («Program» vertex) to which the author is bound using the *Has* edge when the author is a researcher from one of the six universities used in the case. Authors are connected to other authors through *Co-Authoring* edge.

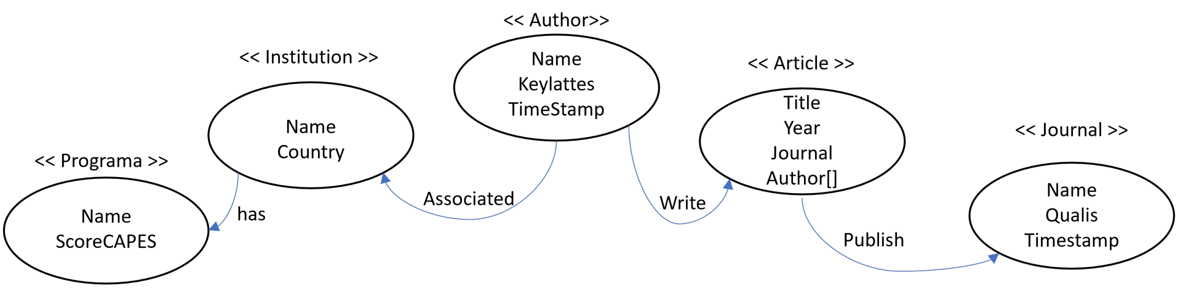

Fig. 3: Conceptual data model.

Technologies Summary. The artifact constructs with different technologies have to overcome challenges. One challenge was the collaboration network data storage in a NoSQL database. Neo4j is the option for the best query performance in graph-based data models [Holzschuher and Peinl 2013]. To execute analytical calculations and the generation of graphs, the python library Networkx was used [Hagberg et al. 2008]. The message exchange between modules uses JSON data format for structured data representation, frequently used in Web services development to the detriment of the XML format.

The DBLP publications data were obtained in the form of a gzip compressed XML file of 660 MB, and the collection of researchers data from the Sucupira platform with a CSV file of 46 MB.[6] When extracted the XML file consumed about 3.3 GB that was handled by a Simple API for XML (SAX) XML parser. For the authors' name disambiguation, the Lucene open-source Java-based search library is used through a Neo4j full-text index [Gospodnetic et al. 2010]. The SCI-synergy code is available in the GitLab InfoKnow research group project repository,[7] with the online artifact.[8]

3. SCIENTIFIC COLLABORATION NETWORK DISCUSSION

There are 196 authors associated to the six Brazilian Computer Science graduate programs distributed as follows (decreasing order): 62 from UFMG, 34 from USP, 31 from UnB, 29 from UFRN, 28 from UFAM and 12 from UNISINOS. From these 196 researchers ("seeds"), publications and co-authors were added, resulting in $10,883$ authors in the scientific collaboration network. We proposed three RQ to shed light on social interaction between researchers associated to graduate programs. RQ are analyzed, based on the dataset collected (March 2022), but updates in the graph database will modify the results presented in the RQs answers.

RQ1. How is the scientific collaboration network of each graduate program?

To perform this analysis, one can observe the network itself, as in Figure 4 filtered by UnB. Also the *average degree* of the entire network is computed by $Avg = 1/2\frac{\sum_{i=1}^{n} deg(i)}{N}$, where $N = |V|$ is the number of nodes, and $deg(i)$ is the number of edges of node i. In SCI-synergy artifact this metric is implemented using a *Cypher* query in the Neo4j database, as follows.

```
MATCH
 (s:Institution {name: 'UNB'})-[:ASSOCIATED_TO]-(a:Author)-[r:COAUTHOR]-(n)
WHERE
 id(a) < id(n)
RETURN
 s.name AS Program,
 toFloat(count(r))/count( distinct a) AS avg_degree
```

Intuitively, $deg(i)$ is computed through the graph traversal starting from institution UnB represented by s to every co-author node n. The query includes filter a not to compute twice the same author in the graph. The result of the *Cypher* query considering the UnB is 2.33 (Table II).

The *average degree* is presented in Section 2.2.2 as a basic social network measure. A high score means there are many relationships among nodes, while isolated nodes present a low score. This metric was applied to each graduate program in the dataset as it presents the overall profile of each program, focusing on how connected or collaborative researchers are in the scientific network. The first-order neighborhood calculates the *average degree* of each program. The co-authorship relation between the allied authors and their co-authors is considered for this metric.

[6] CSV file named BR-CAPES-COLSUCUP-DOCENTE-2020-2021-11-10, available at https://dadosabertos.capes.gov.br/dataset/2017-a-2020-docentes-da-pos-graduacao-stricto-sensu-no-brasil
[7] https://gitlab.com/InfoKnow/SocialNetwork/aureliocosta-sci-synergy
[8] http://165.227.113.212

The total network amounts including the number of researchers, co-authors, co-authorship relations, scientific production, and *average degree* of each graduate program is shown in Table II. The number of researchers was captured from Sucupira platform 2020-2021. The researchers publication with co-authoship relations include all data available in DBLP (up to March 2022). The *average degree* of each program's collaboration network varies from 2.31 to 2.84, meaning that on average each author is connected to more than 2 authors in each collaboration network. Note that, regardless of the others numbers presented, the programs *average degree* contained in this dataset have close values. Nevertheless, considering the total amount of researchers, co-authors, co-authorship relations, scientific production of the six graduate programs together, the *average degree* of the whole network is 6.93 (as presented in `http://165.227.113.212/statistics`).

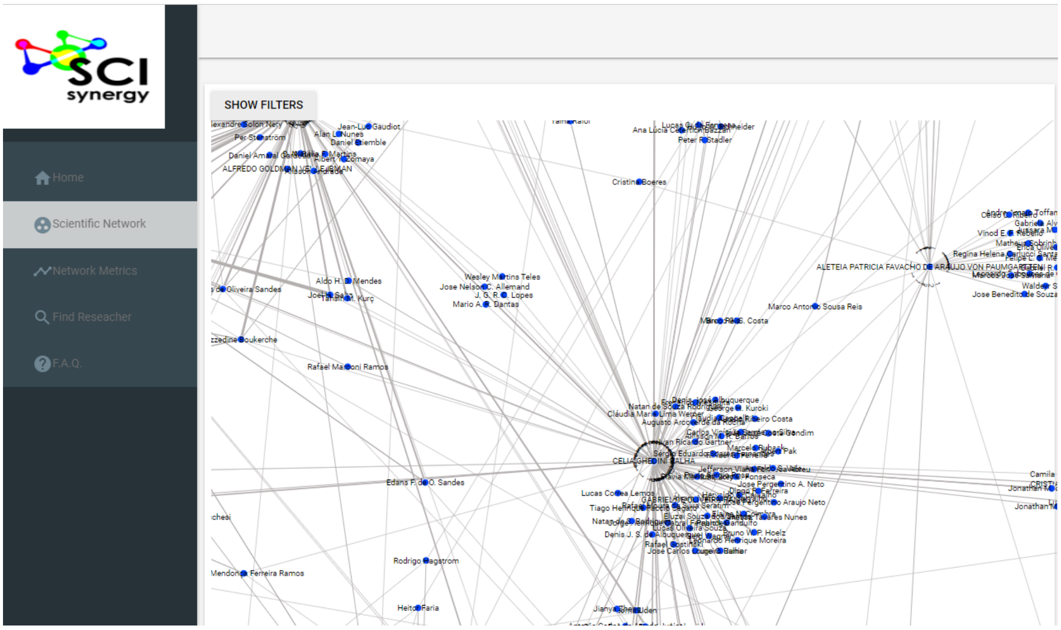

Fig. 4: Partial view of the UnB scientific network.

Table II: The *average degree* of the Computer Science graduate programs cases.

University Name	CAPES Level	Researchers	Co-authors	Co-authorship Relations	Total Publications	Average Degree
UFMG	seven	62	3,375	10,050	4,331	2.79
USP	six	34	1,610	3,921	1,990	2.31
UnB	five	31	1,791	4,344	1,797	2.33
UFRN	five	29	1,301	3,865	1,550	2.84
UFAM	five	28	1,365	3,683	1,503	2.55
UNISINOS	four	12	916	2,368	816	2.48

Although other variables must be considered, UFMG Computer Science graduate program is the largest and oldest one, which could imply the highest *average degree* of the researchers network. However, a high number of authors do not necessarily reflect the amount of collaboration in publications. Even though the year of the first publication of UFMG and USP is almost the same, 1981 and 1980, respectively, the co-authors' number, co-authorship relations, and the total of publications are almost two times bigger in UFMG, resulting in USP's *average degree* of 83% of UFMG. It is interesting to note that the *average degree* of UFRN (2.84) is the highest among of the three level 5 programs, even with the least number of co-authors $(1,301)$ and the total number of publications $(1,550)$, meaning there is a well-connected network in the program with many collaborative researchers.

RQ2. How cooperative is each researcher in the graduate program?

We chose two centrality measures to analyze how cooperative each researcher is in the graduate program: *betweenness centrality* and *closeness centrality*. The *betweenness centrality* allows finding the members that act as "bridge" between researchers in the collaboration network or the researchers who influence the flow of knowledge in the scientific network. Considering the dataset, Table III presents the ten first researchers with the highest *betweenness centrality* score, their program association, and CNPq researcher scholarship level. To compute this metric, we extract the graph projection containing only vertices with label *Author* and edges of type *COAUTHOR* of the whole network. The Neo4j graph data science package that provides the *betweenness centrality* function is used in the implementation based on [Brandes and Pich 2007].

Table III: The ten highest *betweenness centrality* scores.

Researcher	Graduate Program	CNPq Fellow	Score
Jussara Marques de Almeida	UFMG	PQ-1C	106,862
Renato Antonio Celso Ferreira	UFMG	PQ-1D	69,862
Alba Cristina Magalhães Alves de Melo	UnB	PQ-1C	68,624
Frederico Araújo da Silva Lopes	UFRN	—	67,440
Alessandro Fabricio Garcia	PUC-Rio	PQ-1B	65,661
Carlos José Pereira de Lucena	PUC-Rio	PQ-1A	63,068
Eduardo Figueiredo	UFMG	PQ-2	61,997
Wagner Meira Junior	UFMG	PQ-1B	58,253
Leandro Aparecido Villas	Unicamp	PQ-1D	50,375
Nelson Luis Saldanha da Fonseca	Unicamp	PQ-1A	49,550

One can observe that the participation of UFMG is 40%. We added information about these researchers as their CNPq fellow level varying from PQ 1A (the highest level) to PQ 2. It might indicate them as a "bridge" of knowledge or maybe some sort of authority promoting collaboration in the network. It is also interesting to note that some researchers are co-authors from universities outside the dataset collected, as PUC-Rio and Unicamp.

Another centrality measure used to check how cooperative researchers are is *closeness centrality*. It tells us the researcher that is closer to a major number of other collaborators in the network and can influence the entire network quicker. The Neo4j graph data science package with the same graph projection already presented is used. Table IV presents the ten first researchers with the highest scores with their program association and CNPq researcher level. The researcher with the highest score is from UFAM. Also, an interesting fact is the presence of six researchers without the CNPq scholarship and the researcher's better-ranked PQ-2 score. This is a trait of *closeness centrality* and could indicate that these researchers are tied to big influencers of the entire network.

RQ3. What are the intra- and inter-program collaboration patterns?

Regarding the dataset, Sci-Synergy intra- and inter-program research collaboration network views of the six graduate programs are illustrated in Figure 5. In this figure, the edge width represents the amount of collaboration between the institutions, and the total amount appears as the edge label. For better readability, Sci-Synergy does not use self-loops and represents intra-program collaboration as a subgraph with two vertexes and one edge. Considering the inter-program relationships greater than ten, note that there are 218 relationships between UFMG–UFAM, followed by UFRN–UnB (31), UnB–UNISINOS (21), UFMG–UnB (15) and UFAM–USP (13). Thus, UFMG presents a greater total of

Table IV: The ten highest *closeness centrality* scores.

Researcher	Graduate Program	CNPq Fellow	Score
Tayana Uchôa Conte	UFAM	PQ-2	8,992
George Luiz Medeiros Teodoro	UFMG	PQ-2	6,743
Sarajane Marques Peres	EACH-USP	——	6,706
Routo Terada	USP	——	6,348
Fatima de Lourdes dos Santos Nunes Marques	USP	PQ-2	5,985
Sandro Jose Rigo	Unicamp	DT-2	5,859
Everton Ranielly de Sousa Cavalcante	UFRN	——	5,571
Mauricio Roberto Veronez	UNISINOS	PQ-2	5,206
Rafael Kunst	UNISINOS	PQ-2	5,186
Maristela Terto de Holanda	UnB	——	5,038

inter and intra-collaborations among the six Computer Science graduate programs that could indicate this research program is in a higher collaboration ranking and have good inter-program cooperation partners.

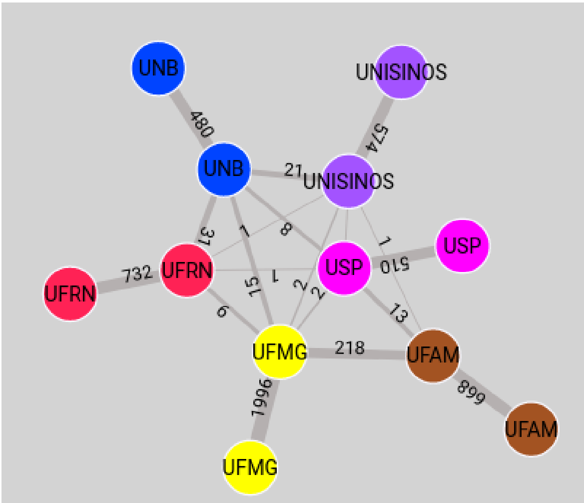

Fig. 5: Intra- and inter-program views.

An interesting aspect is an absence of cooperation between some programs as UFAM–UnB, and UFAM–UFRN. Also considering the intra-program view the UFMG holds the highest collaboration score (1, 996), followed by UFRN (732), UFAM (668), UNISINOS (574), USP (510) and UnB (480). One can observe that it is not necessarily a rule that programs with the highest number of researchers have the biggest intra-program relationships.

Considering the topological aspect of the network, we applied the *connected components* metric to check how big and how many communities there are in the network and validate whether the small world phenomenon happens [Milgram 1967]. The identification component represents each research group in the network. Table V presents the first ten groups applying the *connected components* metric to the network formed by the six graduate programs. To compute this metric disconnected sub-graphs are formed by traversing the graph through edges type *COAUTHOR*.

Note that component 0 is formed by 10, 847 inter-program members, representing the biggest group in the network and almost the only component with more than one member. The *connected components* metric indicates that most of the researchers are connected to the bigger group, being 168 of 196

Table V: First ten groups of *connected components* with two database updates.

Identification Component	Total Members	Program Members
0	10,847	168
16	2	1
46	1	1
44	1	1
65	1	1
48	1	1
25	1	1
70	1	1
78	1	1
14	1	1

researchers associated with one of the six programs. Figure 6 presents the participation of each program member in the biggest component, as an example, UFMG has 62 total members, and 44 of them are in the biggest component representing 70.97% of its members. Besides, there are four programs with high participation in the biggest groups, namely: USP, UnB, UFAM, and UNISINOS (decreasing order). Those newer small programs can easily have their members in the biggest component. For instance, UFAM had a collaborative graduate program with UFMG. UnB researchers studied for master's and Ph.Ds. courses in older programs such as USP and UFMG. Nevertheless, further investigation is necessary to drive conclusions related to the *connected components* composition.

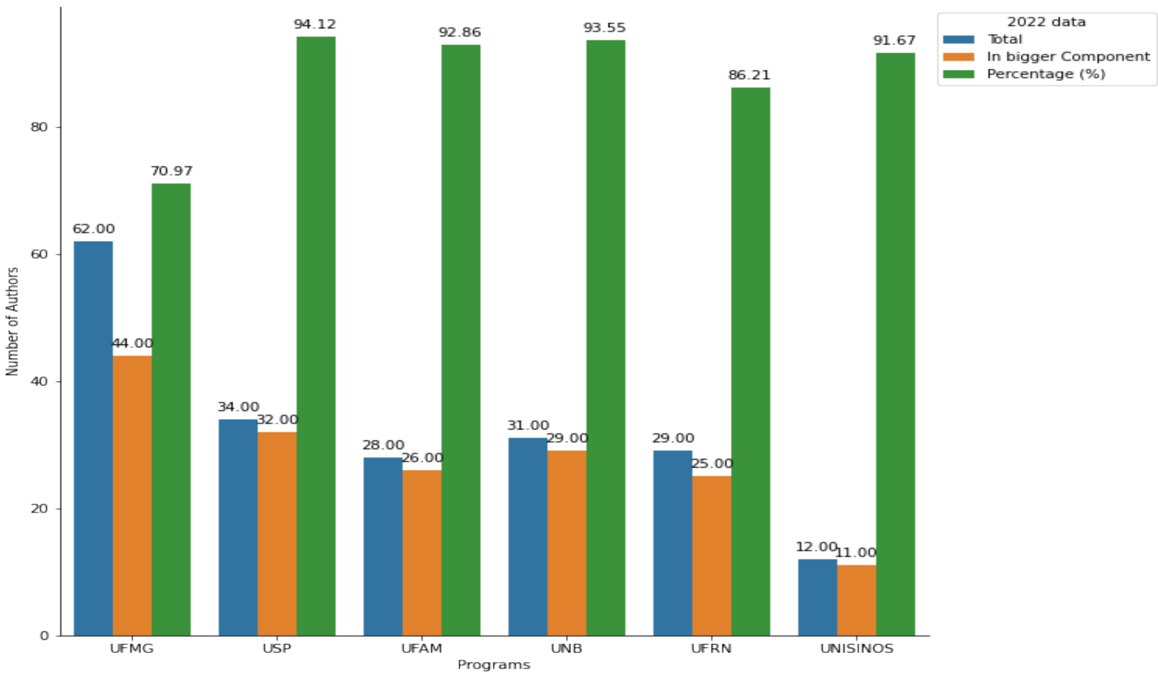

Fig. 6: Researchers' participation of the biggest component.

Considering the researchers' cooperation in the graduate programs many aspects can be analyzed using intra- and inter-program views of a scientific collaboration network. For example, the network results of Table V present many individual groups which might reflect a specific characteristic of the graduate programs analyzed. But further investigation with other Computer Science programs can verify whether this is a national reality. However, we can confirm that this network has a small world aspect with more than 99% of members in the biggest component or 10,847 members of 10,883 total researchers.

Regarding the case dataset used, SCI-synergy presents interesting results to discuss the research questions. An overview of the previously metrics calculated for each program is presented in Table VI. In the next section, we concentrate providing the related work that inspired the present study.

Table VI: Overview of scientific collaboration network analysis.

Program	Average Degree	Betweenness Centrality	Closeness Centrality	Intra & Inter Collaboration	Connected Components
UFMG	2.79	40%	10%	1,996 246	44/62 70.97%
USP	2.31	0%	20%	510 25	32/34 94.12%
UFAM	2.55	0%	10%	668 232	26/28 92.86%
UnB	2.33	10%	10%	480 61	29/31 93.55%
UFRN	2.84	10%	10%	732 42	25/29 86.21%
UNISINOS	2.48	0%	20%	574 26	11/12 91.67%

4. RELATED WORK

Literature review highlight works dealing with collaboration aspects in scientific social networks or discussing the Brazilian Computer Science scenario. To analyze these works, we elected a set of aspects related to the goals of this research inspired by the subjects addressed in the profile analysis of the top Brazilian Computer Science graduate programs [Silva et al. 2017]. The aspects were classified into six categories: (1) focus of the study: studies concentrated on individual scholars, research groups or institutions; (2) geographic coverage: the scientific network coverage (a country or the world); (3) data source: DBLP digital library or other platforms/sites (Lattes, Jems); (4) adopted metrics: volume of publication, authors and co-authors, and social network metrics; (5) scope of the analysis: network properties, collaboration patterns, temporal evolution and intra- and inter-program collaboration; and (6) implementation: it is available online and the code accessible for anyone. SCI-synergy was compared to ten research papers summarized in Table VII (columns are sorted by crescent number of features) and discussed in the sequence.

4.1 Focus of the Study

The focus of this research is not to evaluate the researchers' profiles but to analyze their scientific collaboration network. The analysis focuses on the researchers' collaboration considering intra- and inter-program relationships. Thus, the resulting analysis can be used by distinct audiences, including individual scholars, research groups, institutions, and research funding organizations, among others.

The authors in [Trucolo and Digiampietri 2017] developed an approach to predict trend topics in Computer Science. They used social network metrics of authors (from the Lattes platform) based on the co-authorship network. Authors in [Brandão and Moro 2017] measure tie strength in co-authorship social networks proposing a new metric called *tieness* to better differentiate the degrees of strength with good results compared to metrics from digital libraries on Computer Science, Medicine, and Physics.

The authors in [Freire and Figueiredo 2011] worked with individual scholars and research groups to characterize the structural properties of the scientific collaboration network in the area of Computer

Science. The authors considered the global network formed by DBLP individuals with 750 thousand authors (collected in June 2009) and over 1.3 million publications associated with the Brazilian Network of individuals affiliated with Brazilian institutions (about 1, 600 authors). They established a direct comparison between these two networks, ranking individuals within a group in the network using their relationships in opposition to rank all individuals. Intuitively, the importance assigned to an individual by the proposed group-based metric is proportional to the intensity of their relationships with the group outsiders.

Table VII: SCI-synergy compared to related work.

	[Lima et al. 2015]	[Lopes et al. 2012]	[Lima et al. 2017]	[Brandão and Moro 2017]	[Wagner et al. 2019]	[Bazzan and Argenta 2011]	[Maia et al. 2013]	[Silva et al. 2017]	[Trucolo and Digiampietri 2017]	[Freire and Figueiredo 2011]	SCI-synergy
Focus of the Study											
individual scholars				✓					✓	✓	✓
research groups	✓	✓	✓		✓	✓	✓	✓		✓	✓
Geographic Coverage											
specific country	✓	✓	✓			✓	✓	✓	✓	✓	✓
the world				✓	✓	✓				✓	✓
Data Source											
DBLP		✓		✓				✓		✓	✓
other[a]	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓
Adopted Metrics											
publication volume	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
co-authoring counts		✓	✓	✓	✓	✓	✓	✓	✓		✓
social network metrics				✓					✓		✓
Scope of the Analysis											
network properties				✓		✓	✓		✓	✓	✓
collaboration patterns			✓		✓	✓	✓	✓		✓	✓
temporal evolution	✓	✓	✓		✓		✓	✓	✓		✓
intra & inter-program											✓
Implementation											
code & use availability											✓

[a] Lattes platform (http://lattes.cnpq.br), Journal and Event Management System of the Brazilian Computing Society (https://jems.sbc.org.br), graduate programs web sites, etc...

The authors in [Silva et al. 2017] present a statistical analysis comparing the top 25 Brazilian Computer Science graduate programs to understand their strengths and weaknesses. The authors compare the number of faculty members, the number of publications of each researcher (classifying journals and in proceedings), and CAPES graduate program ranking focusing on the groups. The authors in [Lima et al. 2015] worked on a research statistical assessment and researchers ranking. The study focused on the 406 top researchers from Brazilian universities according to the CNPq five ranks (1A, 1B, 1C, 1D, and 2) grouped by computer science research area and H-index ranks.

The authors in [Lima et al. 2017] investigated the Brazilian Symposium of Database (SBBD) co-authorship network from bibliographic data of 30 editions (1986 to 2015). The analysis includes several network metrics as the progress of the most engaged authors, the number of authors, institutions, published papers, and the evolution of the most frequent terms presented in the paper titles, also the influence, density, and impact of the prominent authors. This approach focuses on evaluating statistical averages and ranges. The authors in [Maia et al. 2013] used the social network defined by paper authors published in the Brazilian Symposium on Computer Networks and Distributed Systems (SBRC) from bibliographic data from 30 editions (1983-2012). The authors used temporal analysis of the network evolution to investigate the number of communities and their interrelation through time (also considering geographic distribution), the importance of nodes, network homophily, their degree distribution, and correlations.

The authors in [Lopes et al. 2012] use the Gini coefficient to evaluate weighted research networks from two perspectives. The first perspective analyzes the temporal evolution of research networks by the distribution of researchers who have co-authored publications. The Gini coefficient was used in the INWeb project to measure the homogeneity level of collaboration. The second perspective compares internal collaboration networks of graduate programs with the Gini coefficient to support the ranking task. Authors end up proposing the β-*index* that combines average and Gini to evaluate the collaboration network as an assessment criterion for ranking Brazilian graduate programs.

The authors in [Bazzan and Argenta 2011] investigated the properties of the social network defined by researchers acting as PC members in Brazilian Computer Science conferences. Like the two previous papers, this one focuses on evaluating statistical averages and ranges. The authors in [Wagner et al. 2019] investigated whether research papers produced through international collaboration are more novel or only highly cited.

4.2 Geographic Coverage and Data Source

The geographic coverage of the related work consider different data sources, but most of them counts on Brazilian data: (1) in [Silva et al. 2017] the authors are from the top 25 Brazilian computer science graduate programs; (2) in [Lima et al. 2015] the social network is formed by the top Computer Science researchers of the Brazilian universities; (3) in [Lima et al. 2017] the authors are from the SBBD; (4) in [Trucolo and Digiampietri 2017] the authors network is formed by the Ph.Ds. in computer science registered in Lattes platform; (5) in [Maia et al. 2013] the social network is defined by authors that published in the SBRC from 1983 to 2012; (6) in [Lopes et al. 2012] the collaboration network includes 27 researchers involved in the Brazilian National Institute of Science and Technology for the Web (INWeb) from UFMG, UFRGS, UFAM and CEFET-MG, and a bigger network included 732 researchers from 27 Brazilian graduate programs in Computer Science with their publications extracted from DBLP (August 2010); (7) in [Bazzan and Argenta 2011] the social network is formed by senior researchers acting as PC members in the Brazilian Computer Science conferences including Brazilian and non-Brazilian members; (8) in [Freire and Figueiredo 2011] the global network is formed by DBLP individuals affiliated with Brazilian institutions (about 1600 authors). The authors in [Lima et al. 2015; Lima et al. 2017] state that although the profile analysis used national data the approach may be used by other countries. In [Brandão and Moro 2017], the authors are from three real social networks built from publications collected at digital libraries on Computer Science using articles and proceedings in the DBLP, medicine with a US national library of the Medicine National Institute of Health that comprises biomedical publications - PubMed (publications from the top-20 journals classified by h-index), and physics with American Physical Society (sample dataset with its journal publications). In [Wagner et al. 2019] authors treat specifically international collaboration using data from Web of Science and Scopus in 2005.

This work used members of six Brazilian Computer Science graduate programs to present the scientific collaboration network analysis through Sci-Synergy involving publication data from the DBLP library. Data from the graduate program members are collected from the Sucupira platform. The authors from the programs are mostly from Brazil but co-authors worldwide.

4.3 Adopted Metrics, Scope of the Analysis and Implementation

Regarding the adopted metrics, all works considered authors' publication volume, but [Lima et al. 2015; Freire and Figueiredo 2011] did not use co-authoring. In this work, we calculate how many co-authors one author has worked with and how many times this collaboration occurred. In [Brandão and Moro 2017] authors apply the *tieness* metric to compute the strength of relationships between a pair of authors to identify when a tie is weak or strong. In [Lopes et al. 2012], authors consider the Gini coefficient (g_c) of the distribution of researchers who have co-authored publications, the average (ρ) of co-authored papers, and the β-*index* that combines average and Gini. In [Trucolo and Digiampietri 2017], authors use social network metrics of Ph.Ds. researchers in Computer Science computed based on the co-authorship network.

Considering the scope of analysis, in [Trucolo and Digiampietri 2017; Brandão and Moro 2017; Maia et al. 2013; Bazzan and Argenta 2011; Freire and Figueiredo 2011], the authors implemented network properties (average path length, graph diameter, clustering coefficient, etc...). Collaboration patterns were considered in [Wagner et al. 2019; Silva et al. 2017; Lima et al. 2017; Maia et al. 2013; Bazzan and Argenta 2011; Freire and Figueiredo 2011]. Specially small world and link density collaboration patterns were applied in [Lima et al. 2017; Bazzan and Argenta 2011]. Temporal evolution is taken into account in [Wagner et al. 2019; Silva et al. 2017; Lima et al. 2017; Trucolo and Digiampietri 2017; Lima et al. 2015; Maia et al. 2013; Lopes et al. 2012]. We used the absolute frequency of interaction, but we stand out by considering the elapsed time between the co-authorship. Using SCI-synergy, it is possible for year-round queries, where a researcher might query his network or the graduate program network within a specific period (e.g., the last three years). Note that only our work focuses on intra- and inter-program collaboration patterns.

Related to implementation, there are no works with implemented solutions and code available in the related work. The SCI-synergy online artifact is implemented with the code available in the GitLab as presented in the technologies summary section.

5. CONCLUSION

This work presents a scientific collaboration network view of six Brazilian Computer Science graduate programs using social network metrics and the SCI-synergy artifact. The case focuses on the researchers' collaboration involving publication data available on DBLP and Sucupira platform to identify scientific collaboration relationships allowing to understand patterns of intra- and inter-program relationships. Although the case study focuses on Computer Science graduate programs in Brazil the database can grow with data from other programs or even research fields with the corresponding enforcement of the infrastructure to accommodate.

Three research questions were answered through the analysis of scientific co-authorship applying social network metrics – *average degree*, *connected components*, *betweenness centrality*, and *closeness centrality*. The first question – How is the scientific collaboration network of each graduate program? – focused on characterizing the collaboration profile of each program, we used the *average degree* to show that there is a regular pattern of distribution of degrees among programs. The second question – How cooperative is each researcher in the graduate program? – look for program members' particular collaboration profile, we used the *betweenness centrality* and *closeness centrality* to identify the centrality profile of collaboration. These metrics show that the collaboration flows in a distributed way. However, an old program like UFMG might present a great number of "bridges" according to *betweenness centrality*. The third question – What are the intra- and inter-program collaboration patterns? – explores the participation of program members in the biggest network partition using the *connected components* metric. This metric highlights the bound of collaboration in the entire network.

In future work, the SCI-synergy database might be extended to include other Computer Science graduate programs in Brazil. Affiliation attributes can be incuded in the database to allow institutional analysis and internationalization aspects of co-authorship. Also, different data sources can be explored like Web of Science and Scopus to include evaluation of other knowledge domains. With an extended database disambiguation techniques have to be improved to explore authors' correspondence. Finally, additional network metrics, such as degree centrality, Eigenvector centrality, Katz centrality, transitivity, reciprocity, and similarity, can be used to compute different collaboration aspects.

Acknowledgements

Professor C. G. Ralha thanks for the support received from the Brazilian National Council for Scientific and Technological Development (CNPq) for the research productivity grant in the Computer Science area (PQ-2) under grant number 309688/2021-3.

REFERENCES

Arruda, D., Bezerra, F., Neris, V. A., Toro, P. R. D., and Wainera, J. Brazilian computer science research: Gender and regional distributions. *Scientometrics* vol. 79, pp. 651–665, 2009.

Bazzan, A. L. C. and Argenta, V. F. Network of collaboration among PC members of brazilian computer science conferences. *J. Braz. Comp. Soc.* 17 (2): 133–139, 2011.

Bonacich, P. Power and centrality: A family of measures. *American Journal of Sociology* 92 (5): 1170–1182, 1987.

Brandão, M. A. and Moro, M. M. The strength of co-authorship ties through different topological properties. *J. Braz. Comp. Soc.* 23 (1): 5, 2017.

Brandes, U. and Pich, C. Centrality estimation in large networks. *International Journal of Bifurcation and Chaos* 17 (07): 2303–2318, 2007.

Chesbrough, H., Vanhaverbeke, W., and West, J. *Open innovation: Researching a new paradigm.* Oxford University Press, Oxford, UK, 2006.

Costa, A. R. and Ralha, C. G. Analysis of scientific collaboration network as a management tool for graduate programs. In *Proceedings of the Annual Conference on Brazilian Symposium on Information Systems: A Computer Socio-Technical Perspective - Volume 1*. SBSI 2015. Brazilian Computer Society, pp. 275–282, 2015.

Costa, A. R. and Ralha, C. G. *Análise de Rede de Colaboração Científica como Ferramenta na Gestão de Programas de Pós-graduação.* Vol. Princípios e Aplicações da Computação no Brasil 2. Atena Editora, Ponta Grossa, Paraná, Brazil, 2019. `https://www.atenaeditora.com.br/post-artigo/4924` (visited 2019-04-11).

Coxon, A. P. M. Perspectives on social network. In *Perspectives on Social Network Research*, P. W. Holland and S. Leinhardt (Eds.). Academic Press, Oxford, 23, pp. 489–500, 1979.

Cronin, B. *The hand of science: Academic writing and its rewards.* Scarecrow Press, Lanham, MD, USA, 2005.

Cugmas, M., Mali, F., and Žiberna, A. Scientific collaboration of researchers and organizations: a two-level blockmodeling approach. *Scientometrics* vol. 125, pp. 2471–2489, 2020.

Digiampietri, L. A., Mena-Chalco, J. P., de Melo, P. O. S. V., Malheiro, A. P. R., Meira, D. N. O., Franco, L. F., and Oliveira, L. B. Brax-ray: An x-ray of the brazilian computer science graduate programs. *Plos One* 9 (4): 1–12, 2014.

Ferreira, A. A., Gonçalves, M. A., and Laender, A. H. F. *Automatic Disambiguation of Author Names in Bibliographic Repositories.* Morgan & Claypool, San Rafael, CA, USA, 2020.

Franceschet, M. Collaboration in computer science: A network science approach. *Journal of the Association for Information Science & Technology* 62 (10): 1992–2012, 2011.

Freire, V. P. and Figueiredo, D. R. Ranking in collaboration networks using a group based metric. *J. Braz. Comp. Soc.* 17 (4): 255–266, November, 2011.

Glänzel, W. National characteristics in international scientific co-authorship relations. *Scientometrics* 51 (1): 69–115, 2001.

Glänzel, W. Coauthorship patterns and trends in the sciences (1980-1998): A bibliometric study with implications for database indexing and search strategies. *Library trends* 50 (3): 461–475, 2002.

Gospodnetic, O., Hatcher, E., and McCandless, M. *Lucene in action.* Simon and Schuster, 2010.

Hagberg, A., Swart, P., and S. Chult, D. Exploring network structure, dynamics, and function using networkx. In *Proceedings of the 7th Python in Science Conferences (SciPy 2008)*. pp. 11–16, 2008.

Hansen, D. L., Shneiderman, B., Smith, M. A., and Himelboim, I. Chapter 3 - social network analysis: Measuring, mapping, and modeling collections of connections. In *Analyzing Social Media Networks with NodeXL (Second Edition)*, Second Edition ed., D. L. Hansen, B. Shneiderman, M. A. Smith, and I. Himelboim (Eds.). Morgan Kaufmann, pp. 31–51, 2020.

Hevner, A. and Chatterjee, S. *Design Research in Information Systems: Theory and Practice*. Springer Publishing Company, New York, NY, USA, 2010.

Holzschuher, F. and Peinl, R. Performance of graph query languages: comparison of cypher, gremlin and native access in neo4j. In *Proceedings of the Joint EDBT/ICDT 2013 Workshops*. ACM, pp. 195–204, 2013.

Hu, Z., Tian, W., Guo, J., and Wang, X. Mapping research collaborations in different countries and regions: 1980–2019. *Scientometrics* vol. 124, pp. 729–745, 2020.

Jiang, J., Shi, P., An, B., Yu, J., and Wang, C. Measuring the social influences of scientist groups based on multiple types of collaboration relations. *Information Processing & Management* 53 (1): 1–20, 2017.

Kang, I., Na, S., Lee, S., Jung, H., Kim, P., Sung, W., and Lee, J. On co-authorship for author disambiguation. *Inf. Process. Manag.* 45 (1): 84–97, 2009.

Katz, J. S. and Martin, B. R. What is research collaboration? *Research Policy* 26 (1): 1 – 18, 1997.

Kim, K.-W. Measuring international research collaboration of peripheral countries: Taking the context into consideration. *Scientometrics* 66 (2): 231–240, 2006.

Laender, A. H. F., de Lucena, C. J. P., Maldonado, J. C., de Souza e Silva, E., and Ziviani, N. Assessing the research and education quality of the top brazilian computer science graduate programs. *ACM SIGCSE Bulletin* 40 (2): 135–145, 2008. Invited Report.

Ley, M. The DBLP computer science bibliography: Evolution, research issues, perspectives. In *International Symposium on String Processing and Information Retrieval (SPIRE)*, A. H. F. Laender and A. L. Oliveira (Eds.). Lecture Notes in Computer Science, vol. 2476. Springer, Berlin, Heidelberg, Lisbon, Portugal, pp. 1–10, 2002.

Lima, H., Silva, T. H. P., Moro, M. M., Santos, R. L. T., Meira, W., and Laender, A. H. F. Assessing the profile of top brazilian computer science researchers. *Scientometrics* 103 (3): 879–896, 2015.

Lima, L. H. C., Penha, G., de A. Rocha, L. M., Moro, M. M., da Silva, A. P. C., Laender, A. H. F., and de Oliveira, J. P. M. The collaboration network of the brazilian symposium on databases-30 editions of history. *J. Braz. Comp. Soc.* 23 (1): 10, 2017.

Linden, R., Barbosa, L. F., and Digiampietri, L. A. "brazilian style science"—an analysis of the difference between brazilian and international computer science departments and graduate programs using social networks analysis and bibliometrics. *Social Network Analysis and Mining* 44 (7), 2017.

Lopes, G. R., da Silva, R., Moro, M. M., and de Oliveira, J. P. M. Scientific collaboration in research networks: A quantification method by using gini coefficient. *Int. J. Comput. Sci. Appl.* 9 (2): 15–31, 2012.

Maia, G., de Melo, P. O. S. V., Guidoni, D. L., Souza, F. S. H., Silva, T. H. P., Almeida, J. M., and Loureiro, A. A. F. On the analysis of the collaboration network of the brazilian symposium on computer networks and distributed systems. *J. Braz. Comp. Soc.* 19 (3): 361, 2013.

Martin, C. and Niemeyer, P. On the impact of network size and average degree on the robustness of centrality measures. *Network Science* 9 (S1): S61–S82, 2021.

Milgram, S. The small world problem. *Psychology today* 2 (1): 60–67, 1967.

Newman, M. *Networks: An introduction.* Oxford University Press, Oxford, New York, 2010.

Newman, M. E. J. Coauthorship networks and patterns of scientific collaboration. *Proceedings of the national academy of sciences of the United States of America* 101 (suppl 1): 5200–5205, 2004.

Park, J.-H. and Nam, E. Intra- and inter-institutional collaborative pairing trends of lis field in south korea. *Libri* 66 (2): 125–136, 2016.

Silva, T. H. P., Laender, A. H. F., Jr., C. A. D., da Silva, A. P. C., and Moro, M. M. A profile analysis of the top brazilian computer science graduate programs. *Scientometrics* 113 (1): 237–255, 2017.

Simon, H. A. *The sciences of the artificial.* MIT press, Cambridge, MA, 1996.

Tang, L. and Walsh, J. Bibliometric fingerprints: name disambiguation based on approximate structure equivalence of cognitive maps. *Scientometrics* 84 (3): 763–784, 2010.

Tomizawa, H. and Shirabe, M. Likelihood of overseas access to international co-authorships. *Scientometrics* 53 (1): 113–121, 2002.

Trucolo, C. C. and Digiampietri, L. A. Improving trend analysis using social network features. *J. Braz. Comp. Soc.* 23 (1): 8, 2017.

Wagner, C. S., Whetsell, T. A., and Mukherjee, S. International research collaboration: Novelty, conventionality, and a typicality in knowledge recombination. *Research Policy* 48 (5): 1260–1270, 2019.

Weber, S. Design science research: Paradigm or approach? In *AMCIS 2010 Proceedings*, 2010.

Yoshikane, F. and Kageura, K. Comparative analysis of coauthorship networks of different domains: The growth and change of networks. *Scientometrics* 60 (3): 435–446, 2004.

Yu, S., Bedru, H. D., Lee, I., and Xia, F. Science of scientific team science: A survey. *Computer Science Review* vol. 31, pp. 72–83, 2019.

Zhang, C., Bu, Y., Ding, Y., and Xu, J. Understanding scientific collaboration: Homophily, transitivity, and preferential attachment. *J. of the Association for Information Science and Technology* 69 (1): 72–86, 2018.