GEOGRAPHIC AND DISCIPLINARY DISTRIBUTION OF THE BRAZILIAN'S PHD COMMUNITY: PATTERNS OF THE SCIENTIFIC COLLABORATION STRUCTURE

Luciano Antonio Digiampietri (1), Rogério Mugnaini (2), Caio Trucolo (3), Karina Valdivia Delgado (4), Jesus Pascual Mena-Chalco (5), André Fontan Köhler (6)

(1) Universidade de São Paulo, Programa de Pós-Graduação em Sistemas de Informação, Av. Arlindo Bettio, 1000, CEP 03828-000, São Paulo, SP, Brasil, digiampietri@usp.br, 0000-0003-4890-1548. (2) Universidade de São Paulo, Programa de Pós-Graduação em Ciência da Informação, Av. Prof. Lúcio M. Rodrigues, 443, CEP 05508-020, São Paulo, SP, Brasil, mugnaini@usp.br, 0000-0001-9334-3448. (3) Universidade de São Paulo, Programa de Pós-Graduação em Sistemas de Informação, trucolo@gmail.com, 0000-0003-4297-720X. (4) Universidade de São Paulo, Programa de Pós-Graduação em Sistemas de Informação, kvd@usp.br, 0000-0003-3835-3859. (5) Universidade Federal do ABC, Programa de Pós-Graduação em Ciência da Computação, Rua Santa Adélia, 166, CEP 09210-170, Santo André, SP, Brasil, jesus.mena@ufabc.edu.br, 0000-0001-7509-5532. (6) Universidade de São Paulo, Programa de Pós-Graduação em Estudos Culturais, Av. Arlindo Bettio, 1000, CEP 03828-000, São Paulo, SP, Brasil, afontan@usp.br, 0000-0002-8291-1654.

Abstract

The study of national academic characteristics is an imperative task for the understanding of national scientific production and the creation of effective science policy. Using a dataset of more than 3.2 million Brazilian curricula, we explore the academic community of PhDs working in Brazil in order to identify characteristics of the whole national network and in the knowledge area level. We used metrics from social network analysis and text mining techniques, as well as the patterns of collaboration between areas and the regional distribution of PhDs. The results show different general characteristics of the PhDs working in each Brazilian state and knowledge area, according to the social and economic characteristics of each of the five Brazilian regions. Different interaction profiles were described, like a less connected network in Linguistics, Letter, and Arts, in which each researcher is related, on average, to less than three other PhDs; on the opposite side, Agricultural Sciences each researcher is related, on average, to more than nine other PhDs of the network. It is clear that besides the capital and one or other major city, the Northeast Region is devoid of PhDs, a situation that is particularly problematic for the most destitute region of Brazil.

Keywords: Academic social network, scientific assessment, social network analysis, Lattes Platform, Brazil

1 Introduction

The assessment of the academic social network from any country is a complex task that involves the treatment of a large volume of data and requires different kinds of studies. Brazil is the fifth largest country in the world and the fifth most populous. Additionally, the country is composed of five major regions - North, Northeast, Central West, Southeast, and South -, which present great differences, regarding economic and cultural issues, level of literacy and entry into the higher education, history of occupation and colonization, etc.

For example, the Northern Region is mostly occupied by the Amazon Forest, it has a large portion of its population composed of indigenous people and their descendants, and has social and

economic indicators below the national average for all its seven federal units. On the other hand, the Southern Region, with a population predominantly descended from Europeans (mainly Italians and Germans), has a set of rich and affluent cities, with universities and colleges that offer master and doctorate courses, and has social and economic indicators above the Brazilian average.

The combination of these two points makes the description and analysis of the Brazilian scientific community particularly complex. In addition, the presentation of a general analysis for the country, which fails to portray the marked regional differences, risks presenting results that, on average, do not occur in any of its five major regions, not even in the different scientific areas (Leta, Glänzel and Thijs, 2006).

One real obstacle, considering assessing the full scientific production of any given country is to gather hundreds of information sources, considering the diversity expressed by documental typology, disciplinarity and scope (national or international). However, this challenge can be diminished given the existence of a national database of academic curricula, which is managed by a system called Lattes Platform. It congregates information about research groups, academic curricula, and academic institutions.

Scientific collaboration is influenced by several factors such as geographical proximity and relationships established among researchers. In addition, although studies on scientific collaboration often focus on co-publication, scientific collaboration should be considered beyond research results (Vanz and Stumpf, 2010). It permeates the whole cycle, from its conception (when, for example, a process of supervision begins), as well as its development in the execution of a project.

The information needed to establish the different kinds of relationships is typically dispersed. In this sense, the information provided in the academic curricula helps to overcome various obstacles: the problem of homonym, the dispersion of bibliographic output on different sources of information, the current affiliation of the researcher, the advisor-advisee relationship, the areas of interest, as well as collaboration with other researchers.

Lattes Platform is an online system maintained by the Brazilian National Council for Scientific and Technological Development (CNPq) to congregate academic and professional information from researchers that work in Brazil. The Lattes database records the past and current lives of Brazilian researchers, being used not only by CNPq but also by other federal and state institutions and development agencies.

Today, there are more than five million curricula registered in Lattes Platform, and many studies use information from this database as the primary source of research. However, there are some

challenges in the utilization of this dataset: (a) it is not possible to download the entire database, instead each curriculum must be individually downloaded in HTML or XML format; (b) there is a lack of standardization in many fields which are manually filled by the owner of the curriculum; (c) more than half of the curricula were last updated more than one year ago; and (d) there is an enormous volume of information (the total space occupied by the XML files is above 140 gigabytes, there are millions of curricula, dozens of millions of registers of publications, tens of millions of registers of academic degrees, millions of registers of interest in knowledge areas, and so forth).

This paper aims to describe, analyze and evaluate the Brazilian scientific/academic community, through its professors, researchers and other professionals, with at least the title of Ph.D., from two dimensions. We describe the distribution of PhD holders over areas of interest as well as over the five major Brazilian regions; the data were analyzed and can be visualized for the federal units as well as for the concentration of PhDs in the main cities. The assessments are based on well-known metrics from social network analysis (Wasserman and Faust 1994; Camacho, Kim and Trawinski, 2015).

The remainder of this paper is organized as follows: next subsection contains a brief review of the related work; Section 2 presents the methodology and Section 3 presents the findings and discussion; Section 4 brings the limitations of the research; and Section 5 provides the conclusions.

1.1 Related Work

In the last years, several studies analyzed the Brazilian academic social networks on a micro level. There are still relatively few studies that assessed the whole country academic social networks composed of researchers who work in different knowledge areas. Some of them (Melo, 2011; Digiampietri et al., 2014; Mena-Chalco, 2013; Mena-Chalco et al., 2014; Tuesta et al., 2015; Lima et al., 2015; Silva et al., 2017; Dias and Moita, 2018; Damaceno et al., 2019) use the information registered in the Lattes Platform.

Tuesta et al. (2015) examine the advisor and advisee relationship for the researchers who are involved in the area of Exact and Earth Sciences in Brazil and its eight subareas. The authors identified a positive correlation between the time of cooperation of the advisee and the advisor and the productivity of the advisees. Additionally, they analyzed the gradual decrease in intellectual dependence between the advisor and the advisee.

Damaceno et al. (2019) use the set of individuals with at least a master's degree who have a curriculum registered in the Lattes Platform in order to examine the Brazilian academic genealogy.

Based on the curricula registered in the Lattes Platform, the authors were able to draw a broad overview of the advisor-advisee relationship and to measure the level of interdisciplinarity between areas of knowledge, among other points.

Mena-Chalco et al. (2014) analyzed the Brazilian coauthorship network built using data from more than one million of curricula from Lattes Platform. The authors analyzed the network using different graph metrics and constructed subnetworks according to the knowledge areas declared in the curricula. They assessed the structure of these networks and the social behavior of the researchers in the different areas.

In a PhD Thesis, Melo (2011) characterized the Brazilian scientific community considering three aspects: productivity, internationalization, and visibility. The author examined the curricula published in the Lattes Platform of 51,080 PhDs that are participants of research groups (the groups are also registered in the Lattes Platform). The aspect of internationalization was further studied by Mugnaini, Leite and Leta (2012), comparing internationalization profile between two subgroups of the 51,080 PhDs in Lattes Platform: those who published at least one article in Web of Science journal, and those whose name were not present in that database.

According to Dias and Moita (2018), although individuals with PhDs constitute only 5.38% of the total curriculum registered in the Lattes Platform, they are responsible for 74.51% of journal papers and 64.67% of communications published in annals of technical-scientific events. Besides, PhD individuals tend to have more up-to-date curricula and to have at least one registered publication. Dias and Moita (2018) also draw attention to the fact that PhDs are responsible for the masters and doctorate advisoring in the stricto sensu graduate programs in Brazil.

Another information source used in academic social network analysis around the world is the DBLP, a platform that provides bibliographic information on major computer science journals and proceedings that contain information of more than 2.3 million papers.

Freire and Figueiredo (2011) used data from DBLP to analyze the Brazilian coauthorship network. They grouped the researchers according to the graduate program where they worked and compared the network metrics with a ranked attributed by the Brazilian Coordination for the Improvement of Higher Education (CAPES). The same comparison was developed by Digiampietri et al. (2014) using information about the professors that work in the Brazilian Computer Science graduate programs. This information was extracted from the Lattes Platform and was combined with information about citation of papers extracted from Microsoft Academic Search and Google Scholar.

Mena-Chalco (2013) studied the Brazilian social network composed of persons that have declared an interest in Humanities, Applied Social Sciences, and the Linguistics, Letters and Arts. They used information from more than 650,000 curricula from Lattes Platform. The authors identified some characteristics of the dynamic of these knowledge areas.

Bornmann and Moya-Anegón (2019) argue that the visualization of bibliometric data on a map, taking the municipalities as a unit of analysis, improves the perception of spatial results and the distribution of what is under analysis (number of articles published, number of PhDs with professional addresses in the city, etc.). When viewing the data by reference to the municipalities, they can be presented in maps without overlapping, which does not happen when the institutions are referenced.

Our work differs from related papers by analyzing all the PhDs working in Brazil (more than 150,000) identified from more than 3.2 million of curriculum vitae extracted from the Lattes Platform. A comparison of metrics involving the whole network and the sub-nets composed of PhDs working in one of the knowledge areas is presented, as well the identification of the number of researchers, professors and other professionals working on each area of knowledge in each one of the 27 Brazilian federal units.

2 Methodology

The present work is divided into five activities: (a) data gathering; (b) sample selection; (c) information extraction; (d) metrics' calculation; and (e) analysis of the results.

Data Gathering: All data used in this work were extracted from the Lattes Platform curricula. On July 2013, the search service of this platform was queried in order to retrieve the identifiers from all the curricula. About 3.2 million identifiers were retrieved, and all the XML files from the respective curricula were downloaded.

Sample Selection: From the universe of about 3.2 million curricula registered in the Lattes Platform, the sample selection was made through two criteria. First, individuals with a PhD degree were selected, regardless of the area of interest or the date of obtaining the degree (184,764 PhDs were identified). From this set of individuals (Brazilian and foreign ones), those without a professional address or who had a professional address outside Brazil were excluded. After applying these two filters, we reached the sample used in the present study, composed of 156,421 curricula. From these curricula, 150,859 belong to Brazilians and 5,562 to foreign ones.

Information Extraction: For each of the 156,421 curricula, the information extracted consists of areas of interest, professional address, and list of collaborators (coauthors, advisees, advisors, collaborators in scientific projects, etc.).

Metrics' Calculation: Two kinds of metrics were calculated. The first one corresponds to metrics only to characterize the sample considering the areas of interest and the Brazilian's federal unit where the PhDs work. The second one corresponds to social network analysis metrics, that were calculated considering nine networks: one corresponding to the network composed of all the selected sample, and eight networks each one composed only by the PhDs of the respective knowledge area (Agricultural Sciences, Biological Sciences, Health Sciences, Exact and Earth Sciences, Humanities, Applied Social Sciences, Engineering, and Linguistics, Letters and Arts). These metrics are useful to understand the structure and behavior of the academic and professional communities composed of the PhDs from each of the knowledge areas. All the networks produced correspond to undirected and unweighted graphs, in which each node corresponds to a PhD and each edge (link) corresponds to a co-authorship relation between two PhDs.

Analysis of the Results: The academic social networks were comparatively analyzed as well as the patterns of collaboration between areas and the regional distribution of PhDs in Brazil.

For the sake of better presentation of the figures, the abbreviations of the names of the 27 Brazilian federal units (26 states and one federal district) were used, according to the following list: AC - Acre; AL - Alagoas; AM - Amazonas; AP - Amapá; BA - Bahia; CE - Ceará; DF - Distrito Federal; ES - Espírito Santo; GO - Goiás; MA - Maranhão; MG - Minas Gerais; MS - Mato Grosso do Sul; MT - Mato Grosso; PA - Pará; PB - Paraíba; PE - Pernambuco; PI - Piauí; PR - Paraná; RJ - Rio de Janeiro; RN - Rio Grande do Norte; RO - Rondônia; RR - Roraima; RS - Rio Grande do Sul; SC - Santa Catarina; SE - Sergipe; SP - São Paulo; TO - Tocantins.

3 Findings and Discussion

In the Lattes Platform, it is possible to register from zero to six areas of interest (or areas of expertise). From the 156,421 PhDs studied in this paper, 6,511 did not register any area of interest; 103,378 had registered only one area of interest, and 46,532 registered more than one area of interest. Figure 1 presents the distribution of PhDs by areas of interest.

Brazil is organized into five major regions, which are subdivided into 26 states and a Federal District. In Figure 2, each colored node corresponds to a PhD positioned in the geographic region of him/her professional address. In Figure 3, the position of the PhDs are the same, but are colored

according to interest area, but including only an amount of 103,378 curricula, with only one area of interest.

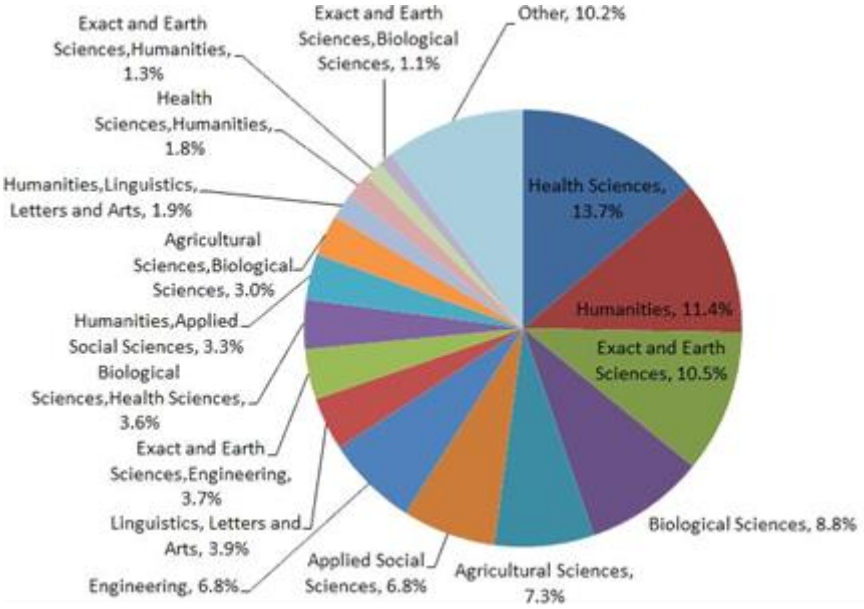
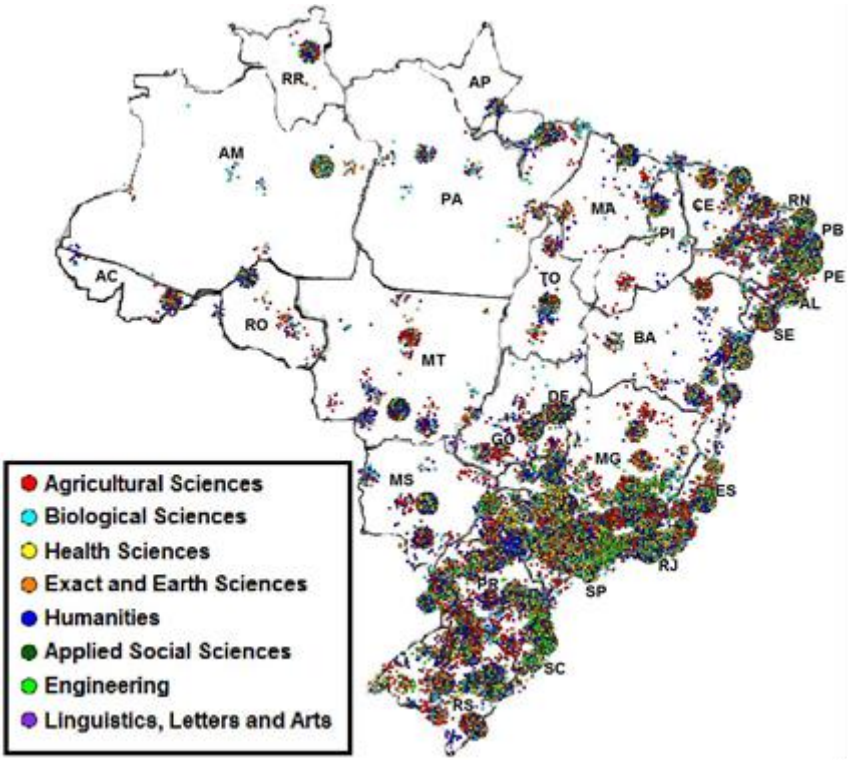


Figure 1. *Distribution of the curricula according to the areas of interest*

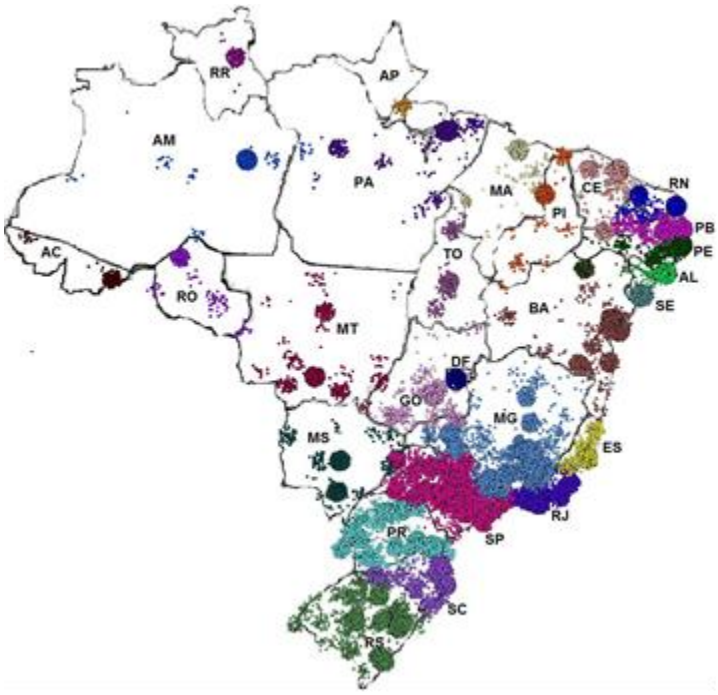
Figure 3 and Table 1 allow some basic analysis of the regional distribution of PhDs in Brazil. The Gross Domestic Product (GDP) and per capita income were extracted from the official government web site (Instituto Brasileiro de Geografia e Estatística, 2017a). First, there is a clear concentration of PhDs in the two most developed regions, the South (PR, SC, and RS) and the Southeast (SP, RJ, MG, and ES). In these two major regions, unlike the other regions, PhDs do not work essentially in state capitals. In addition to having important regional centers, which have national higher education institutions, such as Campinas (SP), Santa Maria (RS) and Juiz de Fora (MG), the states have dozens of cities where many PhDs work, which means that, when observing Figure 3, it appears that they are spread over all states.



Legend: nodes are colored according to the PhD's area of interest.

Figure 2. Distribution of the PhDs according to their professional address (federal unit)

The exception to the rule is the State of Rio de Janeiro (RJ); its capital was the Brazilian federal capital until 1960, and still concentrates a large number of federal government regulatory agencies, public services, and public higher education institutions.



Legend: nodes are colored according to the PhD's state of location.

Figure 3. Distribution of the PhDs according to their federal unit

Table 1 also shows the participation of each federal unit in the population and in the Brazil Gross Domestic Product (Instituto Brasileiro de Geografia e Estatística, 2017a, 2017b), as well as the number of PhDs and how much this represents of the total. A further two data complement the table, namely: (a) the percentage of PhDs in the state capital; and (b) the percentage of PhDs in the second city of the state with more PhDs.

The Central-West Region has a large part of its economy linked to agribusiness; the main Brazilian agricultural frontiers are in this great region. Since the 1970s, the Center-West has shown higher population and economic growth than the Brazilian average. In their three states, capitals account for just over half the number of PhDs. The second city with more PhDs was, until the 1970s, little more than a village, and is currently the main city of the most important state agricultural frontier (Jataí [GO], Sinop [MT] and Dourados [MS]).

Table 1. Distribution of population, GDP, per capita income, and PhDs among the federal units and considering the federal unities' capital and second main city

Federal Unit	Population	GDP	Per capita income*	PhDs	Federal Unit Capital	Percentage of PhDs in the State Capital	Second Main City	Percentage of PhDs in the Second Main City
AC	0.40%	0.17%	43.68%	0.22%	Rio Branco	87.70%	Cruzeiro do Sul	9.78%
AL	1.63%	0.61%	37.49%	0.72%	Maceió	80.90%	Arapiraca	6.85%
AM	1.94%	1.86%	95.62%	1.03%	Manaus	92.06%	Itacoatiara	1.64%
AP	0.38%	0.18%	47.98%	0.08%	Macapá	98.14%	Laranjal do Jari / Santana	0.62%
BA	7.41%	4.32%	58.28%	3.49%	Salvador	55.05%	Feira de Santana	8.69%
CE	4.35%	1.81%	41.64%	2.29%	Fortaleza	80.27%	Sobral	6.45%
DF	1.44%	2.80%	193.94%	3.23%	Brasília	100.00%	Not applicable	Not applicable
ES	1.93%	1.99%	103.03%	1.25%	Vitória	64.87%	Alegre	9.04%
GO	3.25%	2.89%	88.89%	2.00%	Goiânia	66.78%	Jataí	5.62%
MA	3.37%	1.12%	33.12%	0.75%	São Luís	83.56%	Imperatriz	6.75%
MG	10.19%	9.29%	91.16%	9.88%	Belo Horizonte	37.71%	Juiz de Fora	8.11%
MS	1.30%	1.34%	118.67%	1.23%	Campo Grande	50.49%	Dourados	28.06%
MT	1.60%	1.90%	102.87%	1.13%	Cuiabá	52.68%	Sinop	11.47%
PA	4.01%	1.96%	48.72%	1.64%	Belém	79.17%	Santarém	5.78%
PB	1.94%	0.71%	36.79%	2.29%	João Pessoa	51.15%	Campina Grande	27.75%
PE	4.57%	2.35%	51.47%	3.10%	Recife	80.72%	Petrolina	5.35%
PI	1.56%	0.50%	31.87%	0.70%	Teresina	75.15%	Parnaíba	9.45%
PR	5.46%	6.75%	123.72%	6.77%	Curitiba	39.64%	Londrina	15.83%
RJ	8.07%	10.95%	135.61%	12.65%	Rio de Janeiro	73.30%	Niterói	10.72%
RN	1.69%	0.87%	51.68%	1.67%	Natal	75.89%	Mossoró	15.17%
RO	0.87%	0.54%	62.40%	0.24%	Porto Velho	67.64%	Rolim de Moura	5.60%
RR	0.25%	0.13%	52.36%	0.17%	Boa Vista	95.70%	Rorainopolis e Amajari	1.79%

Table 1. Distribution of population, GDP, per capita income, and PhDs among the federal units and considering the federal unities' capital and second main city - continued

Federal Unit	Population	GDP	Per capita income*	PhDs	Federal Unit Capital	Percentage of PhDs in the State Capital	Second Main City	Percentage of PhDs in the Second Main City
RS	5.48%	6.86%	125.33%	8.40%	Porto Alegre	44.97%	Santa Maria	11.35%
SC	3.35%	3.93%	117.17%	3.42%	Florianópolis	58.41%	Joinville	6.58%
SE	1.10%	0.58%	52.48%	0.76%	Aracaju	39.53%	São Cristóvão	51.09%
SP	21.71%	33.20%	152.91%	30.55%	São Paulo	43.49%	Campinas	11.17%
TO	0.74%	0.39%	52.56%	0.36%	Palmas	46.43%	Araguaína	21.43%
Brazil	100%	100%	100%	100%				

This importance of the agribusiness for the Central-West Region can be verified in Table 2, which shows the distribution, state to state, of the PhDs by the eight areas of interest. For GO, MT, and MS, Agricultural Sciences are over-represented. For Jataí, Sinop, and Dourados, this representation is even higher; together, Agricultural Sciences account for 26.27% of PhDs, when, for Brazil as a whole, they represent only 10.54%.

In the Northern Region, PhDs concentrate on state capitals, following what happens with their population and economy. They are the ones that concentrate the federal and state public service and the main institutions of higher education. The exception to the rule is the State of Tocantins (TO), created only in 1989. Created in the same year, Palmas concentrates only 46.43% of the PhDs, whereas Araguaína, one of the most important cities of the region that originally belonged to the state of Goiás, is the home of 21.43% of the PhDs.

Finally, the Northeast Region is the poorest in the country, even though it is the oldest one. In almost all states, PhDs concentrate on state capitals; there are few relevant regional centers with two exceptions. In the state of Paraíba (PB), Campina Grande has just over half the number of PhDs in the capital - and per 100,000 inhabitants, it has more PhDs than João Pessoa. This is due to the presence of dozens of software companies in the city and the Federal University of Campina Grande is a national reference in computer science. It is not by chance that 38.77% of the PhDs working in Campina Grande are from Exact and Earth Sciences and Engineering, while that number is only 24.96% for Brazil.

	Percentage of Total	Agricultural Sciences	Biological Sciences	Health Sciences	Exact and Earth Sciences	Humanities	Applied Social Sciences	Engineering	Linguistic, Letters and Arts	Population
AC	0.22%	0.56%	0.24%	0.12%	0.14%	0.32%	0.10%	0.04%	0.34%	0.40%
AL	0.72%	0.83%	0.53%	0.60%	0.95%	0.70%	0.78%	0.47%	1.03%	1.63%
AM	1.03%	1.17%	2.13%	0.80%	1.21%	0.84%	0.57%	0.73%	0.46%	1.94%
AP	0.08%	0.11%	0.09%	0.06%	0.11%	0.12%	0.09%	0.01%	0.05%	0.38%
BA	3.49%	3.63%	3.28%	3.33%	3.61%	4.09%	2.82%	2.12%	5.75%	7.41%
CE	2.29%	2.76%	1.43%	2.44%	2.45%	2.61%	1.88%	2.38%	2.01%	4.35%
DF	3.23%	2.87%	3.09%	2.21%	2.83%	4.03%	5.84%	2.61%	3.13%	1.44%
ES	1.25%	1.50%	1.15%	0.95%	1.27%	1.24%	1.34%	1.69%	1.19%	1.93%
GO	2.00%	3.50%	1.82%	1.55%	1.85%	2.70%	1.09%	1.35%	2.20%	3.25%
MA	0.75%	1.04%	0.57%	0.74%	0.80%	1.00%	0.56%	0.52%	0.46%	3.37%
MG	9.88%	13.74%	9.67%	8.30%	9.44%	8.67%	9.61%	11.56%	11.01%	10.19%
MS	1.23%	2.40%	1.31%	0.84%	1.08%	1.67%	0.77%	0.46%	1.58%	1.60%
MT	1.13%	2.74%	1.03%	0.70%	0.98%	1.39%	0.65%	0.46%	1.46%	1.30%
PA	1.64%	2.11%	2.26%	0.91%	1.80%	1.92%	1.14%	1.53%	1.68%	4.01%
PB	2.29%	2.92%	1.26%	1.64%	2.23%	2.84%	1.85%	3.28%	3.32%	1.94%
PE	3.10%	3.82%	3.34%	3.25%	3.11%	2.85%	2.95%	2.80%	2.13%	4.57%
PI	0.70%	1.49%	0.61%	0.63%	0.68%	0.85%	0.30%	0.27%	0.81%	1.56%
PR	6.77%	9.13%	6.09%	5.71%	6.19%	7.26%	7.72%	6.16%	7.09%	5.46%
RJ	12.65%	4.95%	16.20%	9.69%	16.09%	13.25%	12.53%	15.20%	14.19%	8.07%
RN	1.67%	1.28%	1.30%	1.27%	2.27%	1.89%	1.39%	2.19%	1.99%	1.69%
RO	0.24%	0.40%	0.30%	0.11%	0.21%	0.35%	0.22%	0.08%	0.36%	0.87%
RR	0.17%	0.45%	0.14%	0.01%	0.17%	0.21%	0.17%	0.09%	0.24%	0.25%
RS	8.40%	9.80%	6.52%	8.72%	7.31%	9.09%	10.25%	6.78%	9.40%	5.48%
SC	3.42%	3.29%	2.11%	2.85%	2.87%	3.83%	4.32%	5.69%	3.37%	3.35%
SE	0.76%	0.86%	0.52%	0.68%	0.92%	0.98%	0.66%	0.62%	0.64%	1.10%
SP	30.55%	21.69%	32.69%	41.79%	29.19%	24.83%	30.04%	30.73%	23.65%	21.71%
TO	0.36%	0.95%	0.31%	0.12%	0.23%	0.48%	0.36%	0.18%	0.46%	0.74%
# of PhDs	103,378	10,895	13,162	20,516	15,695	16,986	10,196	10,106	5,822	206,081,432

Table 2. Distribution of PhDs according to the area of interest and Brazilian state

Bahia (BA) has several cities with significant numbers of PhDs, the result of a state public policy of deconcentration of higher education in the state, through state public higher education institutions, without similar in the rest of the Northeast Region. Finally, in the case of the State of

Sergipe (SE), the municipality with the highest number of PhDs (São Cristóvão) is close to the state capital.

Table 2 presents the relative distribution of PhDs according to the area of interest and the state where they work. The first column corresponds to the abbreviation of each state. The second column contains the percentage of the PhDs that works in the state. The other columns contain the percentage of PhDs per state for the corresponding areas. For each state, it is possible to identify which area of interest is more overrepresented, considering the total percentage of doctors in the state; the corresponding percentage appears in bold. For example, in Acre (AC) there are more than twice PhDs in Agricultural Sciences (0.56%) than the average percentage of PhDs in the state (0.22%). In Amazonas (AM), a state fully located in the Amazon Forest, there are twice PhDs working in Biological Sciences than the average number of PhDs per area in this state. The percentages of PhDs per area are deeply related to the main business and social activity of each state. The last line of this table includes the total number of PhDs in each area of interest. From this table, it is also important to note that more than half of the PhDs in Brazil work in only three states: SP, RJ and MG.

Nine social networks were constructed; one composed of all the 156,421 PhDs working in Brazil and eight networks composed each of them with PhDs that have only one knowledge area of interest. Table 3 (appendix) presents the networks' metrics. Each metric is described as follows, along with a discussion of the metric values for each of the analyzed networks.

The number of nodes corresponds to the number of PhDs in each network, and the number of edges corresponds to the number of relationships (coauthorship, advisee-advisor, and collaboration in scientific projects) between PhDs. In these networks, there are several nodes with degree zero, i.e., nodes without any relationship. The column "Nodes with degree greater than zero" presents the number of PhDs who have at least one relationship in the network, only these PhDs (nodes) were used in the calculation of the other metrics in Table 3.

The giant component is the largest connected component of each network. A connected component is a subset of nodes where all the nodes are linked (directed or undirected) to each other. The presence of the majority of the nodes in the giant component is considered a positive aspect of a social network because it means that a high amount of people is connected with the main information/knowledge of the network. Of the 156,421 PhDs, 75,467 have degrees greater than zero and, from these ones, 73,645 belong to the giant component. Thus, the Brazilian PhDs' network contains 98% of its nodes with degree greater than zero in the giant component. Only one network presents

less than 80% of its PhDs in the giant component: that corresponds to the area of Linguistics, Letters and Arts (75%).

The average degree indicates the average number of relationships (edges) for each node in the network. In the whole network, each PhD is connected, on average, to other eight PhDs. In the Agricultural Sciences' network this value is 9.73 (this network is the one where each node has the largest number of collaborators); on the other hand, this average is only 2.94 in Linguistics, Letters and Arts.

The density corresponds to the ratio between the number of edges of a graph and the maximum number of edges possible for it. The Brazilian network, as a whole, is not dense (the density is equal to 0.0001); the highest value of density occurs in the Linguistics, Letters and Arts' network, which is the smallest of the Brazilian Knowledge Areas' networks.

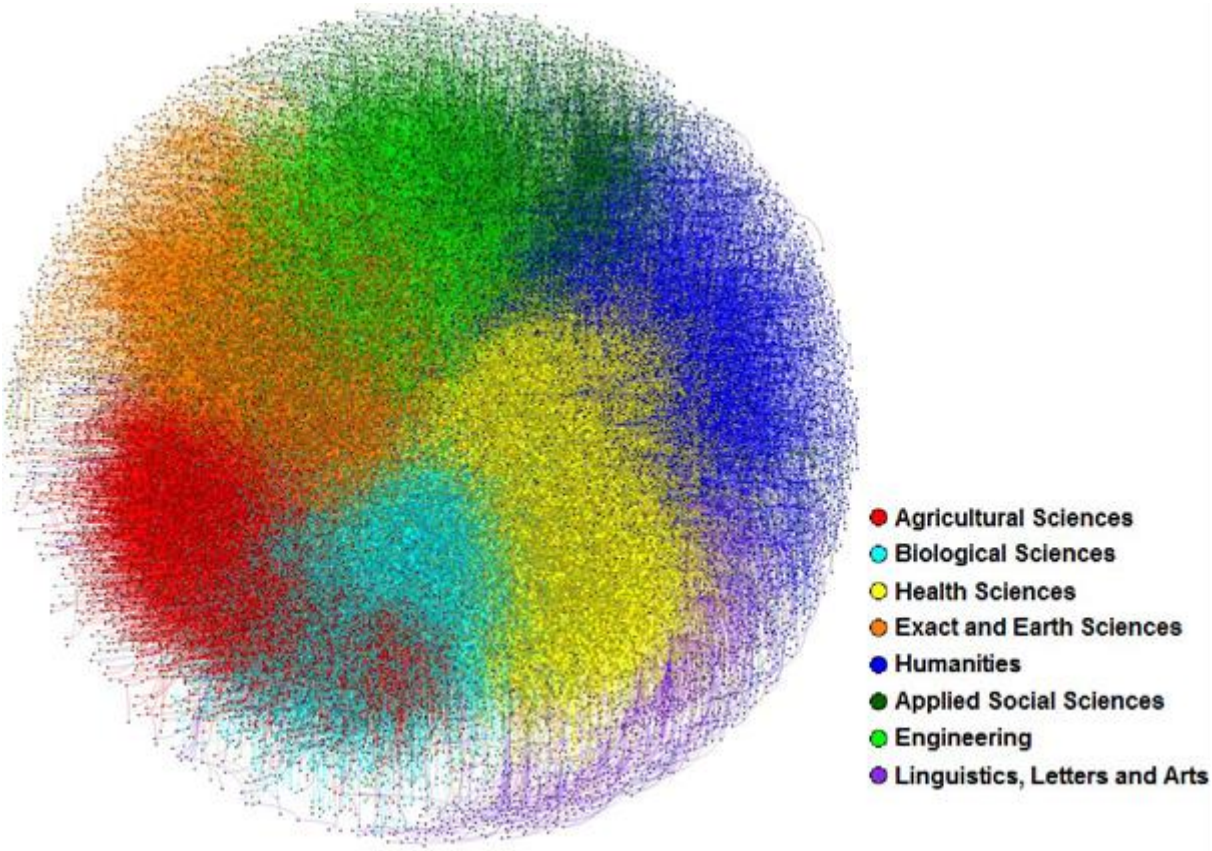
The assortativity measures the tendency of nodes with a common characteristic to connect with each other. For example, the degree assortativity measures if there is a tendency of nodes with the same degree to be connected. The value 1 for this metric means that all the connections (edges) occur between nodes with the same degree. On the other extreme, the value -1 means that all the connections occur between nodes with different degrees. In the nine networks, there are no significant tendencies considering the degree assortativity. The biggest positive tendency was found in the whole network (0.154), and the biggest negative tendency was found in the Linguistics, Letters and Arts' network. Another important kind of assortativity is the knowledge area assortativity that measures the tendency of PhDs to be interested in the same knowledge area to connect. This metric is applied only to the whole network and its value is 0.733, indicating a great tendency of the connections to occur between PhDs with the same area of interest.

The centralization metrics measure the relative importance of the most central node in the network about the entire network. High values of centralization are considered dangerous in a social network because it means that there is a node in the network that concentrates information/knowledge and the exclusion of this node can harm the entire flow of information in the respective network. The degree centralization is based on the degree centrality of the individual nodes (i.e., the importance of the nodes based on their degrees) and the closeness centralization is based on the closeness centrality (based on the proximity of the nodes). The centralizations in the whole network have low values. The highest value of degree centralization is associated with the Linguistics, Letters and Arts' network, and of closeness centralization occurs in the Agricultural Sciences' network. The lowest values of degree centralization are associated with Exact and Earth Sciences and Health Sciences' networks and the lowest value of the closeness centralization is associated with Engineering.

The diameter is the size of the maximum shortest path of a network, i.e., the shortest distance between the two nodes that are farthest from each other in a connected component of the network. In social network analysis, high values of this metric is usually problematic because the diameter is related to the time spent on the information to flow in the network. The highest value of diameter is associated with Humanities' network.

The maximum clique size is the number of elements of the largest subgraph of the graph where all nodes are connected with each other. High values for this metric typically indicate the presence of a strong and cohesive group of researchers working together. The largest clique in the Brazilian network is composed of 24 researchers. The maximum clique size from Agricultural Sciences and Biological Sciences networks is 13 and from Linguistics, Letters and Arts' network is only 5.

Figure 4. *Giant component considering only the PhDs that registered one area of interest*



The average path length corresponds to the mean value of the shortest paths between all pairs of nodes from the same connected component. The smaller the value of this metric the faster is, in average, the propagation of the information between an arbitrary pair of PhDs. The lowest average

path length was found in the Agricultural Sciences' network (4.77); on the other hand, the highest was found in the Humanities' network (7.55).

Figure 4 presents the graph of the giant component of the whole network (considering just the PhDs that registered only one area of interest). It is possible to note concentrations of nodes for the majority of the knowledge areas. It is also possible to observe the mixing of nodes from different areas (different colors) in some regions of the graph and, specifically with more intensity, in the borders between areas. For example, between Agricultural Sciences (red nodes) and Biological Sciences (cyan nodes), and between Engineering (light green nodes) and Exact and Earth Sciences (orange nodes).

Table 4. *Edges' distribution among knowledge areas*

	Agricultural Sciences	Biological Sciences	Health Sciences	Exact and Earth Sciences	Humanities	Applied Social Sciences	Engineering	Linguistics, Letters and Arts	Total
Agricultural Sciences	82.57%	9.82%	1.85%	3.54%	0.20%	0.80%	1.20%	0.01%	57,446
Biological Sciences	8.00%	67.38%	16.33%	6.59%	0.59%	0.16%	0.93%	0.02%	70,522
Health Sciences	1.25%	13.47%	79.70%	2.82%	1.53%	0.22%	0.96%	0.04%	85,497
Exact and Earth Sciences	3.22%	7.35%	3.82%	77.45%	1.05%	0.54%	6.46%	0.10%	63,223
Humanities	0.80%	2.89%	9.12%	4.65%	73.32%	6.01%	0.90%	2.30%	14,340
Applied Social Sciences	3.89%	0.95%	1.61%	2.90%	7.30%	77.67%	4.88%	0.80%	11,805
Engineering	2.41%	2.27%	2.86%	14.22%	0.45%	2.01%	75.71%	0.08%	28,714
Linguistics, Letters and Arts	0.25%	0.54%	1.50%	2.66%	13.72%	3.95%	0.91%	76.48%	2,406

Table 4 contains a detailed view of the edges in Figure 4. Each line contains the percentage of edges that connect the PhDs from the knowledge area of the respective line with PhDs from the area of the respective column. Thus, the sum of each line is 100%. In the last column, there is the total number of edges that involve the PhDs of the respective line. Most of the edges connect PhDs of the same area of interest, as can be seen in the main diagonal of Table 4. This can be easily seen in Figure 4. For example, 82.57% of the edges from the first line (Agricultural Sciences) connect two PhDs whose area of interest is Agricultural Sciences. In this table, it is highlighted in bold font the second largest percentage per line (i.e., the biggest excluding the principal diagonal value). It is possible to note that, proportionally, the Agricultural Sciences PhD's are more connect with Biological Sciences' PhDs than with the PhDs from other areas. With the exception of Exact and Earth Sciences and Humanities, the highlighted values occur between areas that are considered related.

For Agricultural Sciences, Health Sciences, Engineering and Linguistics, Letters and Arts, about 90% of the edges occur between PhDs of the same area of interest and the second one most connected. No other area of interest has a significant number of collaborations.

As for Biological Sciences, Exact and Earth Sciences, Humanities and Applied Social Sciences, there is a significant number of collaborations with PhDs from other two areas of interest. Once again, there is a predominance of collaborations with areas of interest correlated. For example, for Biological Sciences, 8% of the edges are linked to the Agricultural Sciences.

Of the eight areas of interest, Linguistics, Letters and Arts is the most singular. The combination of the lowest percentage of nodes in the giant component with the lowest average degree shows an area in which the PhDs still usually work alone or with few collaborators. This notion of atomization of research is reinforced by the fact that the maximum click size is just 5, the lowest of all areas of interest. In Figure 4, Linguistics, Letters and Arts appears on the periphery of the scientific community; this is reinforced by the small size of this area of interest in Brazil.

4 Limitations

The research presents three main limitations. First, it focuses in a single country – Brazil; this limits its applications and conclusions. Second, as presented, the data extraction was made in 2013; it happened because the Lattes Platform does not permit, nowadays, the same kind of extraction that was made. Moreover, this research is part of a broad study that took years to complete. Third, the data collected in the Lattes Platform is inserted by professors, researchers and professionals themselves; there is no guarantee that the data is accurate and correct. Yet past studies point that the data is in general reliable.

5 Conclusions

This paper analyzed the Brazilian Academic Social Network composed of the PhDs whose professional address is in Brazil.

Nine networks were constructed and analyzed. One corresponding to all the PhDs working in Brazil and one for each of the eight main knowledge areas. It was possible to observe a great heterogeneity in the distribution of PhDs according to their area of interest. This distribution follows the social and economic characteristics of each of the five Brazilian regions.

As expected, there is a greater concentration of PhDs in the capitals of the federal units, as well as in their major cities, with rare exceptions.

It was also observed the interaction between researchers of different knowledge areas. There is a strong relationship between researchers from the same area and a smaller interaction, but also important, among researchers from related areas (such as Applied Social Sciences and Humanities, or Engineering and Exact and Earth Sciences). Different interaction profiles were assessed in the networks. Less connected networks were observed in areas such as Linguistics, Letter, and Arts in which publications with few authors are common and each researcher is related, on average, to less than three other PhDs. On the other hand, much more connected networks could be found in areas such as Agricultural Sciences in which each researcher is related, on average, to more than nine other PhDs of the network.

Finally, there are significant differences between the regional distribution and the predominance of areas of interest between the regions and states of Brazil. For example, it is clear that besides the capital and one or other major city, the Northeast Region is devoid of PhDs, a situation that is particularly problematic for the most destitute region of Brazil.

As future work, we intend to develop a study about the research subjects extracted from the papers published by PhDs working in Brazil, including the analysis of the temporal evolution of collaborations as well as the main research topics. We also intend to investigate the PhDs that declared to work in more than one knowledge area as well to investigate in detail interdisciplinarity collaborations.

References

Bornmann, L.; Moya-Anegón, F. (2019). Spatial bibliometrics on the city level. // Journal of Information Science. 45:3 (2019) 416-425.

Camacho, D.; Kim, S. W.; Trawinski, B. (eds.) (2015). New trends in computational collective intelligence. Berlin: Springer International Publishing, 2015.

Damaceno, R. J. P. et al. (2019). The Brazilian academic genealogy: evidence of advisor-advisee relationships through quantitative analysis. // Scientometrics. 119:1 (2019) 303-333.

Dias, T. M. R.; Moita, G. F. (2018). Um retrato da produção científica brasileira baseado em dados da Plataforma Lattes. // Brazilian Journal of Information Studies: Research Trends. 12:4 (2018) 62-74.

Digiampietri, L. A. et al. (2014). Brax-ray: an x-ray of the brazilian computer science graduate programs. // Plos One. 9:4 (2014) 1-12

Freire, V. P.; Figueiredo, D. R. (2011). Ranking in collaboration networks using a group based metric. // Journal of the Brazilian Computer Society. 17:4 (2011) 255-266.

Instituto Brasileiro de Geografia e Estatística. (2017a). Contas regionais 2015: queda no PIB atinge todas as unidades da federação pela primeira vez na série. 2019a. https://agenciadenoticias.ibge.gov.br/agencia-sala-de-imprensa/2013-agencia-enoticias/releases/17999-contas-regionais-2015-queda-no-pib-atinge-todas-asunidades-da-federacao-pela-primeira-vez-na-serie, accessed on July 20, 2019.

Instituto Brasileiro de Geografia e Estatística. (2017b). Estimativas de população para 1º de julho de 2015. https://ww2.ibge.gov.br/home/estatistica/populacao/estimativa2015/estimativa_tcu.shtm, accessed on July 20, 2019.

Leta, J.; Glänzel, W.; Thijs, B. (2006). Science in Brazil. Part 2: Sectoral and institutional research profiles. // Scientometrics. 67:1 (2006) 87-105.

Lima, H. et al. (2015). Assessing the profile of top Brazilian computer science researchers. // Scientometrics. 103:3 (2015) 879-896.

Medeiros, C. B.; Mena-Chalco, J. P. (2013). The dynamics of multidisciplinary research networks-mining a public repository of scientists CVs. In World Social Science Forum (pp. 1-17). Montreal, Canada.

Melo, P. L. C. (2011). Produtividade, Internacionalização e Visibilidade da Comunidade Científica Brasileira na Virada do Milênio. Rio de Janeiro. (Universidade Federal do Rio de Janeiro Ph.D. dissertation). 2011.
Mena-Chalco, J. P. et al. (2014). Brazilian bibliometric coauthorship networks. // Journal of the Association for Information Science and Technology. 65:7 (2014) 1424-1445.

Mugnaini, R.; Leite, P.; Leta, J. (2012). Fontes de informação para análise de internacionalização da produção científica brasileira. // PontodeAcesso. 5:3 (2012) 87-102.

Silva, T. H. et al. (2017). A profile analysis of the top Brazilian Computer Science graduate programs. // Scientometrics. 113:1 (2017) 237-255.

Tuesta, E. F. et al. (2015). Analysis of an advisor-advisee relationship: An exploratory study of the area of exact and earth sciences in Brazil. // PloS One. 10:5 (2015) 1-18

Vanz, S. A. S.; Stumpf, I. R. C. (2010). Colaboração científica: revisão teórico-conceitual. // Perspectivas em Ciência da Informação. 15:2 (2010) 42-55.

Wasserman, S.; Faust, K. (1994). Social network analysis: Methods and applications (Vol. 8). Cambridge: Cambridge University Press, 1994.

Recebido: 09/09/2019
Aceito: 29/11/2019