# A Study on the Process of Migration to Training in Brazil: Analysis Based on Academic Education Data

Higor Alexandre Duarte Mascarenhas[1(✉)], Thiago Magela Rodrigues Dias[1], and Patrícia Mascarenhas Dias[1,2]

[1] CEFET-MG, Divinópolis, Minas Gerais, Brazil
thiagomagela@cefetmg.br
[2] UEMG, Divinópolis, Minas Gerais, Brazil

**Abstract.** In recent years, a fact that stands out at the national level is the movement of individuals to other locations at some point in their lives. There are several causes that motivate these types of displacement, among them, one of the main reasons for training, especially at the level of academic training. Given this scenario, this work aims to carry out an analysis of how Brazilian academic mobility occurred, through data extracted from currencies and institutions registered in the Lattes Platform. Thus, extracted from the curricula of Brazilians, resulting in 308,317 records. From the extraction, the data was filtered, obtaining the following relevant research items, performed, treatment of the analysis items, removing irrelevant and incomplete terms and, subsequently, improving the data with information on the geographical location of each institution. At each level of education the entire group analyzed, from the place of birth to the individual's current professional performance. Soon after, with the set of detected data, it was possible to carry out the monitoring in the measurements of the social networks, in which the networks were characterized considering the displacements between the places in the academic formation process used. As a result, an image of how the scientific formation process took place after a long process of capacitation was used, making it possible to measure the migratory flow of individuals and trends in the formation processes.

**Keywords:** Platform Lattes · Brazilian scientific exodus · Migratory flow · Data analysis

## 1 Introduction

The emigration of Brazilians to other countries and to other states has increased significantly, so that in Brazil, studies show that some cities have rates of 10 to 30% of migrants who do not live in their home state [2]. In many cases, Brazilians go out for a job or study, always looking for quality of life.


E. Bisset Álvarez (Ed.): DIONE 2021, LNICST 378, pp. 26–40, 2021.
https://doi.org/10.1007/978-3-030-77417-2_2

Among the main reasons for applying is the need for training at a high level of training. One of the main causes for the mobility option in the Brazilian territory, refers to the quality of higher education in other states, the search for new opportunities and more experiences in their areas [9]. Another refuge for these students is directed to other countries, thus seeking cultural exchange and better investment in research grants. The student's departure to other countries is not only interesting for the student, but also for the institutions of origin, as it the same return in most cases more productive, with the most extensive contact network, greater experience and available in the future with the sharing their experiences with other students from the home institution.

According to [6], every day it has become more difficult to produce scientific research in Brazil, due to investment cuts using education. One of the main reasons for the migration of Brazilian researchers to other countries can be pointed out by the lack of government support. Therefore, with this scenario of Brazilian researchers go out in the country, thus making it difficult to return due to the lack of opportunities. Most of the Brazilian scientists who return to Brazil do not work in their area of training, so they do not progress in their careers.

A program that facilitated and helped a lot or the entry of students in institutions with institutions in other countries is Ciência sem Fronteiras, for being a program that supports students, that offers scholarships. In 2015, the government planed to reach 101,000 scholarships for researchers, graduates, PhD students, post-doctoral students, encouraging students to capacitate in institutions of recognized relevance [4]. Today, with only 5,000 scholarships available, the Program has lost a lot of influence in the entry of students to other countries, due to investment cuts.

As a motivation for the study to have an understanding of the Brazilian scientific performance aiming to obtain an opportunity to understand the current scenario, and to adopt measures to promote possible openings of new undergraduate or postgraduate courses in areas where a deficit in specific areas of the knowledge. Another issue related to the economic issue, is a better exploration of the specific area of education in a region of Brazil.

Given this scenario, this work presents a study on the exod of individuals that left your state/city of birth to other states/cities and/or those that went to other countries in search of training. To use the data of Brazilian students analyzed in this study, the framework LattesDataXplorer [7] was used, a tool responsible for extracting and processing CVs of users registered on the Lattes Platform. Currently, the Lattes Platform curricula repository that records academic/scientific and professional information, has approximately 6,750,000 registered curricula. Therefore, a set of components created for the purposes of this study and incorporated into the *framework*, thus enabling a broad and unprecedented view of the Brazilian scientific exod.

## 2 Related Works

According to [8] the shifts in the researcher's training demonstrate a correlation as characteristics of the individual, one of the main characteristics is the degree of international cooperation or scientific production.

[1] in his study carried out an analysis of the migratory flow of people born in 196 countries on all continents, his research carried out an analysis in the mid-1990s until the year 2010, with a study to understand of patterns and trends in flow of immigration from countries and continents selected by the authors. The authors were able to identify the migration flows of individuals registered in the study according to the level of development of the countries.

[5] conducted a study with data from the Lattes Platform of postgraduate researchers, collecting data on the researcher's trajectory from birth to his last formation degree. Having analyzed the group of PhD, reaching the conclusion of 95% are from the South, Southeast and Northeast states. It was mentioned that 40% of the first PhD training courses were carried out in the cities of origin and 87% of those individuals displaced to other cities not exceeded or within the limit of 1,000 km. It was also highlighted that the city with the highest number of PhD in São Paulo.

Already [10] analyzes the mobility of Brazilian researchers and students throughout their academic training. It is worth mentioning that 20% of the researchers work, however, more than 500 km from the institution where they entered the academic trajectory, in contrast, the majority work at about 100 km. This mobility made the interviewed researchers involved in several lines of research, making their work more known, in the places of their trajectory. The study indicates that the states in the southeastern region, mainly São Paulo, are those that most researchers are natural from there; the other states have a temporary migratory pattern.

[3] carried out a study to analyze the circulation of people throughout the academic journey of individuals, as well as their workday. In your study the author, mentioned which researchers from different nationalities choose to engage in migration or obtain more experience in your area and expand a network of contacts at other universities. Subsequently, it is mentioned that in Brazil there is no form of incentive facilitated for Brazilian students emigrated to other countries, and this fact can harm even the same Brazil, as this way hinders the network of contacts between Brazilian researchers and researchers from other nationalities. However, Brazil invests in foreign researchers to study to Brazil, to make contacts networks, however, it is often difficult to find an attraction for foreigners, since there was a reduction in scholarships in the country. Other countries, such as China, invest in obtaining a network of contacts between other countries, and concluded that the United States is one of the countries in which more Chinese choose to perform the exodus.

[11] proposed a study with analysis of intrastate use in Rio Grande do Norte, with occurrences between the Metropolitan Region of Natal and the interior of the state, and between the interior of the state and the Metropolitan Region of Natal and the data obtained for this research, based on two periods, from 1995/2000 and 2005/2010, with data provided by the Instituto Brasileiro de Geografia e Estatística (IBGE). From the data obtained, the authors carried out statistical analyzes to compare the results and point out the main figures on the flows that occurred, and found that individuals choose to exercise the flow from

interior to metropolis, but in both cases in the Metropolitan Region of Natal shows a decline in population gains. About the analysis of migrations from cities that they observed in Natal, there is a greater volume of migrations "from" and "to" the interior of Rio Grande do Sul, because the capital concentrate as the with activities related to the sectors of service, commerce, tourism and education.

Therefore, it is notorious that a large scale of individuals choose to obtain capacitation at a high level of formation left your home city to another, and with a smaller scale, part of your home country seeking capacitação abroad. It is worth mentioning that many works related to this project prefer to extract data from the curricula registered in the Lattes Platform, as it is a repository of great importance for the study of Brazilian scientific production.

## 3 Methodology

In the present work, the main source of data used was the curriculum repository available on the Lattes Platform. Initially, it was necessary to use the *Lattes-DataXplorer* [7] to extract the data, given the difficulty of obtaining them, since the interface to query the Lattes Platform curricula allows access to only one curriculum per time, so the analysis of large groups of individuals becomes a limiting factor. Data extraction was carried out in May 2019, totaling 308,317 resumes from individuals with completed PhD degree, considering all PhD regardless of the date of completion of formation.

Soon after the data extraction was carried out, treatments were carried out with the objective of obtaining formatted data extracts in order to facilitate future analyzes. Thus, steps such as "Data Selection" and "Data processing" were performed according to the scheme shown in Fig. 1.

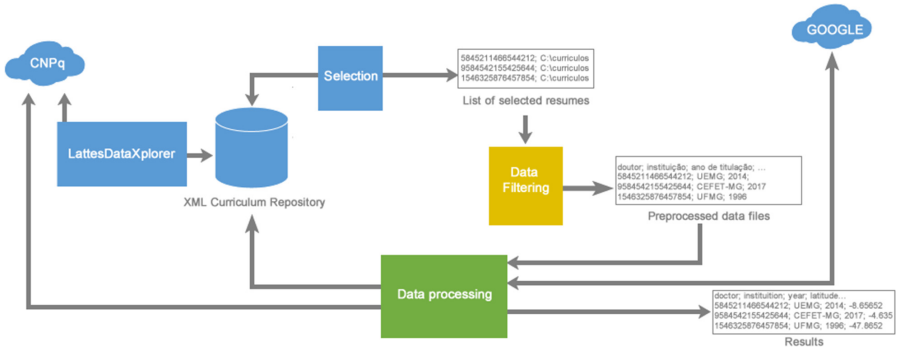


**Fig. 1.** General aspect of the set of components used. Source: Authors.

In the "Selection" Step, the XPath query language (XML Path Language) is used for research and subsequent generation of the subgroups to be analyzed. The XPath language allows the construction of expressions that will go through

an XML document in a similar way to the use of regular expressions. Therefore, it allows the grouping of a set of curricula with desired parameters, such as academic training or areas of expertise.

The list stores the identifiers for each curriculum and the path in which it is stored locally, so it will be possible to analyze only the selected curricula. In view of the above, only curricula were collected from individuals with completed PhD degree, as this is the group with the highest level of academic education; since these are curricula that are frequently updated and most of the parameters required for the present work are registered in their curricula.

After selecting the set to be analyzed, the "Data filtering" module, which is responsible for analyzing the resumes in XML files in order to obtain relevant information for research, features an extract of formatted data (Preprocessed data files). The curriculum information registered in the file has: curriculum identifier; individual's state and city of birth; institution code, name and zip code of the individual's current employment relationship, in addition to the identification code and name of the institution for each level of education completed, considering since undergraduate to PhD.

Afterwards, the module "Data processing" (Fig. 2) is executed, in which four steps are performed: Obtaining the institution's CEP; Search by geographic location; Data cleaning and grouping and Data normalization. The first step carried out is the "Obtaining CEP of the institution" in which, from the institution's code retrieved from the curriculum, it is consulted in the institutions directory of the Lattes Platform, in order to obtain the institution's data and, thus, retrieve it from the address section, the institution's characteristics data, from then on, the website will return the institutions' information and thus obtain the institution's zip code.

The "search for geographic location" stage is a task performed with the purpose of geolocating an institution. Accessing the Google geolocation API (Application Programming Interface), the institution's address will be sent, to later have the institution's geographic location (latitude and longitude) returned.

In the "Data cleaning and grouping" stage, exclusion of possible terms that are irrelevant to the search occurs, in order to reduce the volume of data to be processed and analyzed. As an example: removing *stopWorlds* in city names; normalization to extract accented words, and replace them with their equivalent without accent.

The "Data normalization" stage, on the other hand, aims to reduce the redundancy of information, discarding attributes with the absence of data, such as CEP with no digits.

Subsequently, the "Results File" is generated, representing a summary of all the data obtained in the curricula of Brazilian PhD degree, without needing to consult the XML files of the extracted curricula, having all the specific data for carrying out the analyzes of this research.

Soon, after all the steps described above have been carried out, several metrics are applied to understand how the mobility of Brazilian PhD degree has occurs throughout their academic training process.

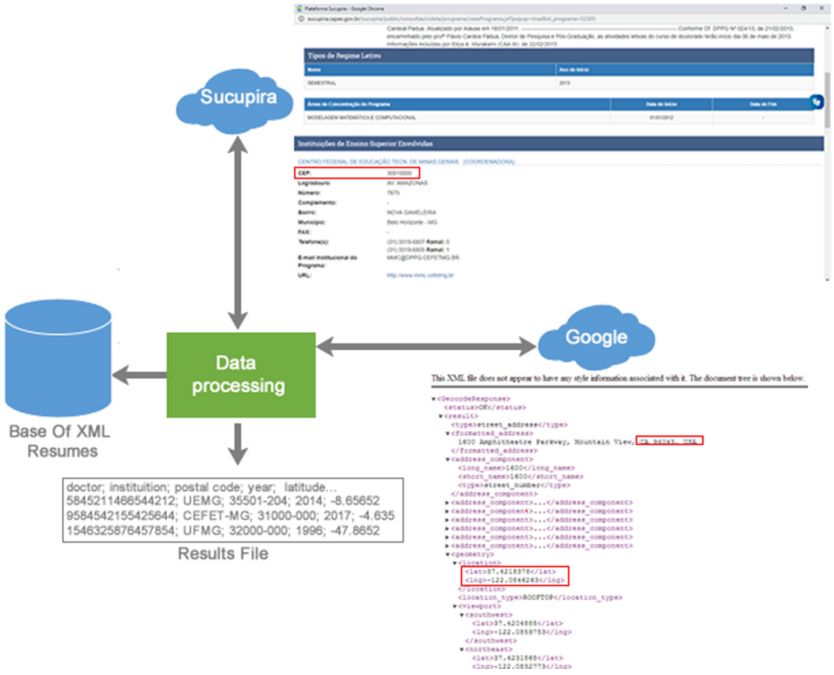


**Fig. 2.** Data processing. Source: Authors.

## 4 General Characterization

Initially, it was possible to characterize the analyzed set. As it is a group of individuals who have the highest level of education, to be considered, only those formations whose status in the curriculum are "completed" were included in the analysis, resulting in a total of 308,317 individuals. In order to evaluate the analysis potential of the extracted data, Fig. 3 shows the number of curricula that have attributes registered, such as: city of birth, institution of activity and formation institutions.

Most individuals have their birth city registered with a total of 293,340 (95%) records, as this is a mandatory field when registering on the Lattes Platform. Those individuals who do not have a birth city registration are considered to be older curricula, in which city registration was not mandatory. Of the other institutions shown in the graphic, the one that identifies itself as superior in quantitative data is the *institution of completion of the PhD degree*, totaling 297,815 (96%) registrations, as it is the group selected to carry out the study. The *postdoctoral institution* is the one with the lowest number of registrations, with an amount of 70,405 (22%), this fact is justified, because the selection of the groups was directed to individuals with a completed PhD degree; for this reason, an individual who holds a PhD degree does not always have a postdoctoral degree.

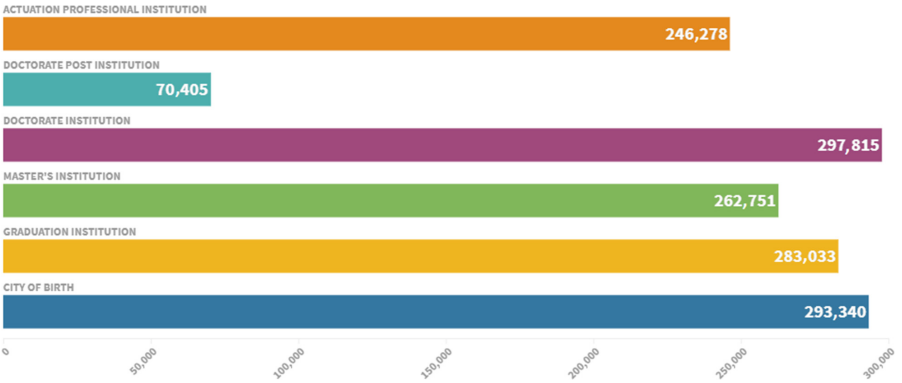


**Fig. 3.** Quantitative of curricula that have information that is subject to analysis. Source: Authors.

After obtaining the number of records registered in the Lattes Platform curricula, it was possible to present in Table 1, the calculation of the average distances in kilometers between one level of training to another.

**Table 1.** Average distance in kilometers between formation levels. Source: authors.

| Distance (km) | Graduation | Master's degree | PhD degree |
|---|---|---|---|
| Birth | 291.58 | 548.59 | 1,000.75 |
| Graduation | – | 432.38 | 901.26 |
| Master's degree | – | – | 619.00 |

It is possible to observe the result of the average distance of all stages of training of Brazilian PhD, during their academic training. It can be observed that the average distance between the stages has a variation considered. Initially, analyzing the average distance from the place of birth for graduation it is noticed that this is the shortest calculated distance. One of the factors that influence this phenomenon is that a large part of Brazilian cities have institutions that provide undergraduate courses to students, and those that do not, in most cases, are close to other cities that hold courses at this level of training of interest to students. The greatest distances, on the other hand, are between the place of birth and formation at the PhD level, followed by the undergraduate/PhD degree in which the displacement is greater than the other levels of training.

Figure 4 aims to demonstrate a characterization of the number of bonds obtained by PhD's at birth, at each level of education, as well as their professional performance. In order to explore better data visualization, the representation of the links has been separated by states, and the colors treated in the heat map vary according to the number of links the state has, with values ranging from 0

links to 40,000 links, aiming to address the discrepancy of the data in relation to the way it was distributed. In the image, there are five different graphics, dealing with the ties of PhD's at each level of education, they are: birth (a); graduation (b); master's degree (c); PhD (d) and professional performance (e).

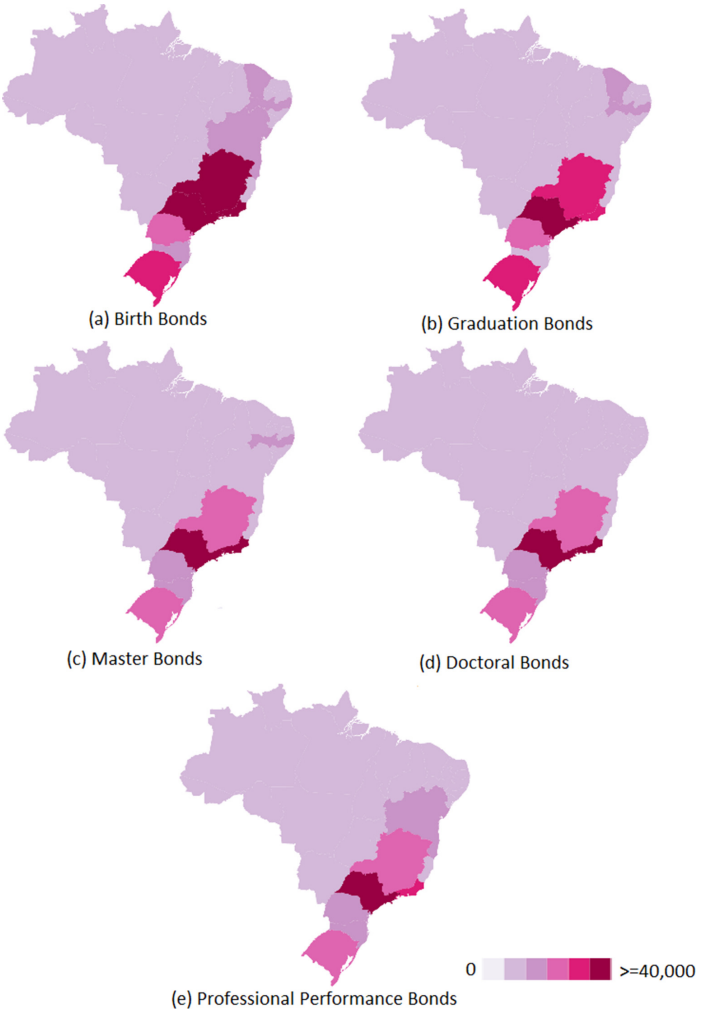


**Fig. 4.** Bonds of PhD's in Brazilian states. Source: authors.

In all the graphs obtained, the state of São Paulo has values on average well above the limit rated by the authors, because if the number of links obtained by the state of São Paulo was used as the highest value, data visualization would be unfeasible from heat maps.

It is observed that the region that stands out most at all levels present, is the Southeast region, which have greater influence in the states of São Paulo, Rio de Janeiro and Minas Gerais, a possible justification is the high concentration of institutions and universities federal governments in this region, in the state of Espírito Santo, do not have a high concentration of bonds in any of the graphs. With less influence, but highlighted in all charts, the state of Rio Grande do Sul also stands out at all levels of education, representing one of the states with a large concentration of Brazilian universities and institutions.

When it comes to the level of birth, in addition to the greater prominence in the southeastern region and the state of Rio Grande do Sul, it is possible to observe the highest concentration in the state of Paraná, and with the least amount of links, there are some states in the northeast region, such as: Bahia, Pernambuco and Ceará respectively. Unlike the graph with birth level, it is possible to observe that the states of Minas Gerais and Rio de Janeiro stand out less in the graph of undergraduate ties and that the state of Bahia does not stand out in this level of education, a possible justification for the lower assiduity of links from the three states, may be the search for capacitation in other states.

Analyzing the graphs that represent master's degrees, it is clear that in the Northeast region, the state of Pernambuco stands out with a greater number of degrees present at this formation level. When taking into account the number of PhD degrees, it confirms once again the hegemony of the state of São Paulo and Rio de Janeiro respectively, and right after Rio Grande do Sul and Minas Gerais, due to the amount of post-graduate courses offered is higher in the four states compared to the others.

When observing professional performance, the state of Bahia stands out compared to the other states, excluding the quartet of states (São Paulo, Rio de Janeiro, Minas Gerais and Rio Grande do Sul) characterizing one of the states with the highest number of professional performance bonds.

It is observed how less frequent are the states of the North and Center-West regions, due to the fact that these regions have fewer universities and institutions, and consequently a lower number of postgraduate courses on offer.

## 5 Results

This chapter presents a characterization of the formation based on the networks of PhD's throughout their academic capacitation and also their professional performance, through their links between Brazilian states and several countries in the world. The networks are directed, and the nodes are represented by the location where the individual was trained and the edge represents the interaction carried out between places where the individuals displaced. The diameter of each knot characterizes the number of degrees it has.

Figure 5 represents the interaction between nodes (Brazilian states/other countries), constituting five different networks: birth - graduation (a); graduation - master's degree (b); master's degree - PhD degree (c); PhD degree - professional performance (d) and birth - professional performance (e).

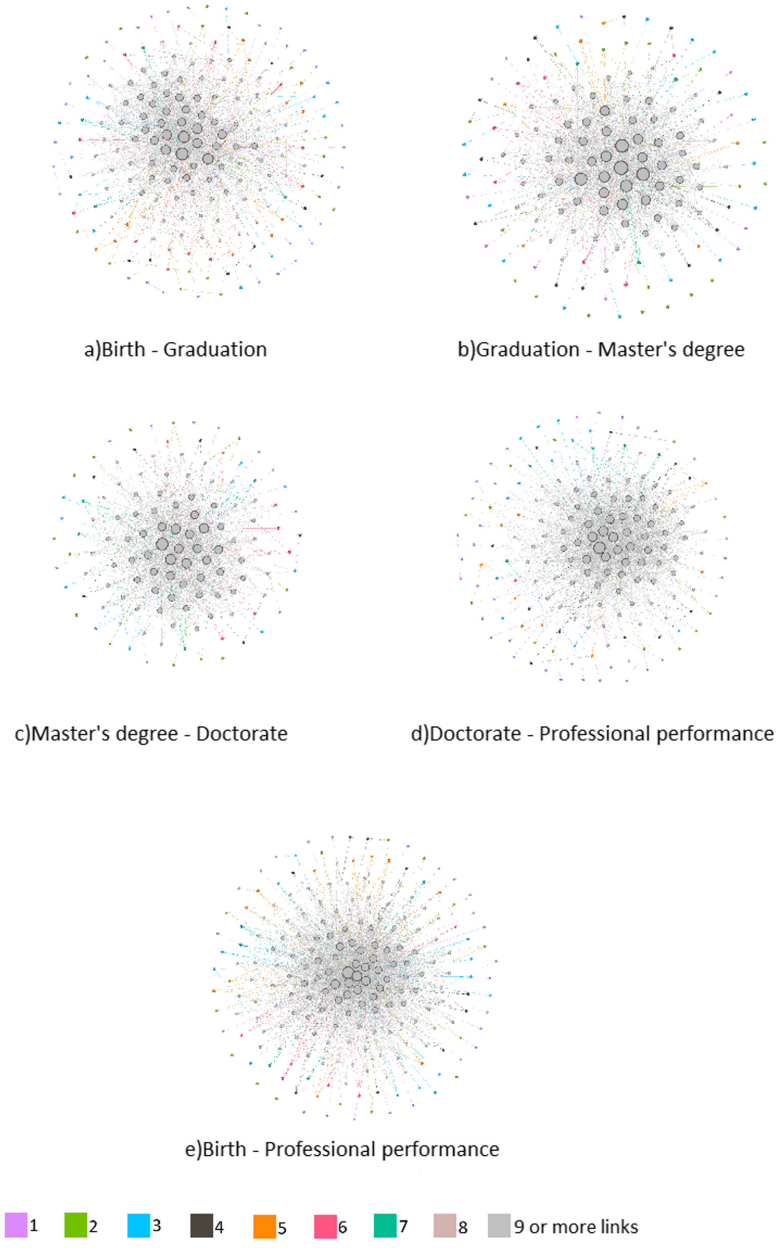


**Fig. 5.** International link networks. Source: authors.

It is possible to observe that some networks have a lower number of nodes in comparison to the others, for example the networks characterized in Figs. 5b and 5c, since generally individuals choose to remain in the same location in

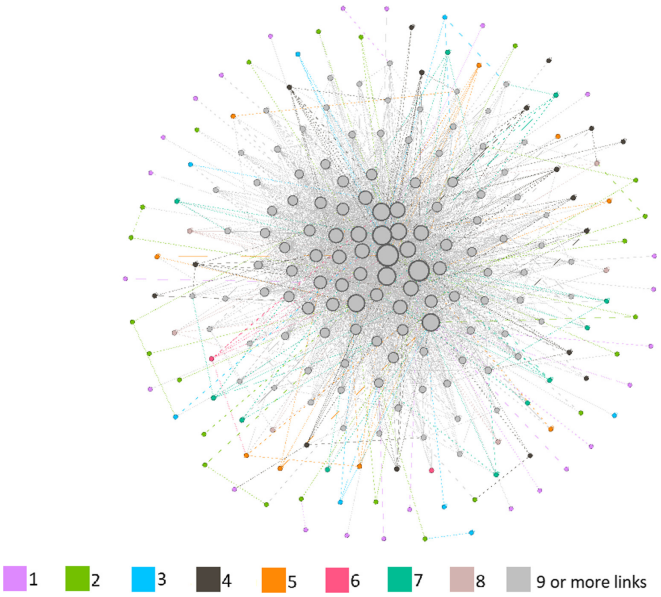


**Fig. 6.** Network of links at all levels of formation at the international level. Source: authors.

**Table 2.** Metrics extracted from the characterized international networks. Source: Authors.

| Metrics | Birth-Grad | Grad-Mest | Mest-Doct | Doct-Actuation | All | Birth-Actuation |
|---|---|---|---|---|---|---|
| Number of nodes | 201 | 144 | 124 | 174 | 217 | 210 |
| Number of Edges | 2,246 | 1,814 | 1,840 | 2,297 | 4,039 | 3,389 |
| Average degree of nodes | 11.174 | 12.597 | 14.839 | 13.201 | 18.613 | 16.138 |
| Nodes of G.C. | 189 | 135 | 119 | 171 | 217 | 206 |
| % of nodes in the G.C. | 94.03 | 93.75 | 95.97 | 98.28 | 100 | 98.1 |
| Edges of G.C. | 2,237 | 1,805 | 1,835 | 2,295 | 4,039 | 3,385 |
| % of edges in the G.C. | 99.99 | 100 | 99.72 | 99.91 | 100 | 99.88 |
| Network Density | 0.056 | 0.088 | 0.121 | 0.076 | 0.086 | 0.077 |
| Network Diameter | 5 | 5 | 5 | 5 | 6 | 5 |
| Average path length | 2.321 | 2.106 | 2.009 | 2.182 | 2.261 | 2.179 |

*Note*: *G.C.*: Giant Component

the formation levels, when it comes to undergraduate - master's degrees, and master's degrees - PhD degree, as a consequence of this, the networks have fewer links.

It is also observed that in the networks characterized by Figs. 5a and 5e, they present greater amounts of links, due to the presence of many countries that do not appear in other networks, mainly due to the fact that some PhD's registered in the Lattes Platform were born in other countries, not being Brazil, and/or in some cases, choose to work professionally in other countries.

Some dyads stand out in the link networks, in which when analyzing them, it is noticed that these dyads are represented by links between us with spelling errors, such as: the individual inserted the origin of the "Butsuana" link and the destination of the link "Botswana"; in another case, representing incomplete data, the individual inserted Bosnia and Hezergovina-Bosnia, also causing a dyad. However, really representing a dyad we have a fact that a PhD inserted the bond in Timor-Leste and the bond in Kosovo.

The networks do not have isolated components, since all links have a node characterizing the entry (being represented as the emigration location) and another node as the destination (representing the immigration location).

A network was created with all the links obtained at birth, using all levels of education, and professional performance (Fig. 6), just not taking into account the birth level for professional performance.

In this network, all nodes belong to a giant component, having no isolated component, and the greatest number of nodes with degrees whose degree value is greater than 9 stands out, it was observed that all Brazilian states are present in this network of links, with a degree higher than 9. It was noted that nodes with degree values greater than 60 have links with most nodes with amount of nodes greater than itself.

The Table 2 is intended to represent a summary of the characterized networks and calculations of metrics adopted.

When comparing the number of nodes and edge rate in the Giant Component, with the global network, it is noticed that the network that stands out the most is the network that has all the links, since it has the largest connected Giant Component, corresponding in 100% of the edges and 100% of the nodes, different from the other networks, in which the values of the rates of the nodes are also high, but with a percentage lower than the network of all links. As far as density is concerned, the network with all links has the third lowest value, different from the master's and PhD network that has the highest density, since the nodes are quite connected, and a possible justification for this is the networks of collaborations belonging to individuals who are in the transition between these levels of education.

It is observed that the largest diameter is represented by the network of all links, with a value equal to 6, where there is a distance of 5 nodes between the two most distant nodes. The other networks have the same diameter value totaling the value of 5, in which the two most distant nodes need to travel 4 nodes to meet.

The network from birth to graduation has the lowest average degree of nodes, being justified because some nodes have a low degree value because they do not have large amounts of bonds, thus decreasing the average degree value. Unlike the network with all links that has the highest average degree value, where the nodes with the highest grades increase the average. With regard to the average minimum path, it is possible to observe that the networks have very close values (approximately 2), on average it only takes two edges to reach a given node from any other node that makes up the network.

To expand the analysis and obtain a better view of the states and countries that have the highest degrees of entry, Fig. 7 was characterized.

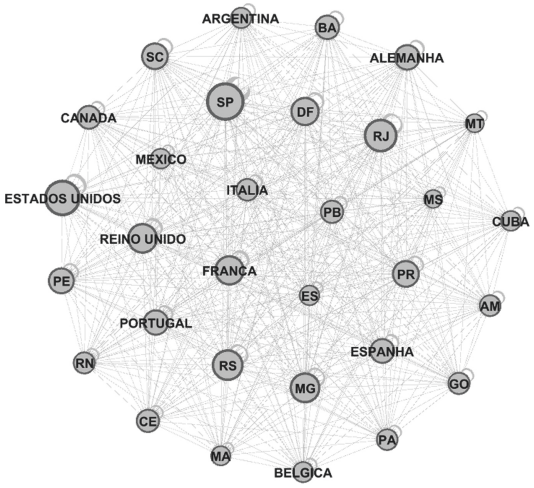


**Fig. 7.** Link networks with higher degrees of entry. Source: authors.

It is possible to highlight that the network is dense with a value of 1, since all nodes are connected to each other. It was also observed that all the states of the South, Midwest and Southeast regions are present, with a hypothesis for this the number of universities belonging to these regions. The North and Northeast regions, on the other hand, have lower numbers of states present, due to the inferior job offers and inferior qualities of graduate courses compared to the other regions.

When observing the countries present in the network, it is clear that Argentina is the only South American country. Still in analyzes aimed only at American countries, Cuba, Mexico, Canada and the United States stand out as the largest nodes, which are those with the greatest number of ties. The United States country stands out with the highest degree value, being inferior only to the state of São Paulo. Some of the possible reasons why the United States is the country that has the highest degree among nodes are: it is the country that has the largest world economy; have a universal language; have a large number of scholarship offers, among others.

When it comes to countries in Europe, seven countries are present, the United Kingdom, France, Portugal, Germany, Spain, Belgium and Italy with the highest degrees of entry of individuals respectively.

## 6 Conclusion

From the results obtained, it was possible to verify the feasibility of adopting the curricula registered in the Lattes Platform as a source of data for analysis on how the Brazilian Scientific Exodus occurs.

The choice of the group of PhD's is characterized as a significant portion of the entire set of data registered in the Lattes Platform, considering that they are the individuals with the highest level of academic education completed. It was also noticed that in general their resumes are recently updated and most of them have a registered professional address.

In addition, it also became clear how the southeast region and the state of Rio Grande do Sul concentrate the vast majority of Brazilian PhD's throughout the academic education process, a fact directly influenced by the concentration of the main public universities in the country.

After carrying out the general characterization of the set of individuals to be analyzed, the next stage of the work was to obtain the geolocation data from all the institutions in which the PhD's were trained at some level of academic training.

After it was possible to find the geolocation data of the institutions, it was possible to identify the average distance traveled by individuals throughout their academic training.

Also noted that the country United States stands out from the rest of the countries, since it corresponds to the country with the largest number of Brazilian fellows enrolled in the country. There is also a greater interaction between European countries compared to South American countries, because European countries have a better quality of education.

## References

1. Abel, G.J., Sander, N.: Quantifying global international migration flows. Science **343**(6178), 1520–1522 (2014)
2. Almeida, G.Z.R.: Fluxos migratórios: a distribuição da população de cada estado pelo país (2017). https://www.nexojornal.com.br/grafico/2017/12/01/Fluxos-migratorios-a-distribuicao-da-populacao-de-cada-estado-pelo-pais. Accessed 1 Jul 2020
3. Andrade, R.O.: O impacto da circulação de cérebros. Revista Fapesp, pp. 18–25 (2019)
4. Aveiro, T.M.M.: O programa ciência sem fronteiras como ferramenta de acesso à mobilidade internacional. Tear: Revista de Educação Ciência e Tecnologia **3**(2) (2014)
5. Chaves, L.C.R., et al.: Analisando a mobilidade de pesquisadores através de registros curriculares na Plataforma Lattes. Dissertation – Universidade Federal da Paraíba (2016)
6. Demartini, M.: Falta de oportunidades mantém cientistas brasileiros no exterior (2017). https://exame.abril.com.br/ciencia/falta-de-oportunidades-mantem-cientistas-brasileiros-no-exterior/. Accessed 1 Jul 2020

7. Dias, T.M.R.: Um estudo da produção científica brasileira a partir de dados da plataforma Lattes. Thesis (Ph.D. in Mathematical and Computational Modeling)-CEFET-MG (2016)
8. Jonkers, K., Tijssen, R.: Chinese researchers returning home: Impacts of internationalmobility on research collaboration and scientific productivity. Scientometrics **77**(2), 309–333 (2008)
9. Lombas, M.L.D.S.: A mobilidade internacional acadêmica: características dos percursos de pesquisadores brasileiros. Scielo **19**(44), 308–333 (2017)
10. de Pierro, B.: Circulação limitada. Pesquisa Fapesp, pp. 36–39 (2016)
11. de Silva, P.S., de Queiroz, S.N.: Migração intraestadual no rio grande do norte. Idéias, **11**, e020008 (2020)