Scientometrics (2019) 119:303–333
https://doi.org/10.1007/s11192-019-03023-0

The Brazilian academic genealogy: evidence of advisor–advisee relationships through quantitative analysis

Rafael J. P. Damaceno[1] · Luciano Rossi[1] · Rogério Mugnaini[2] · Jesús P. Mena-Chalco[1]

Received: 21 August 2018 / Published online: 18 February 2019

Abstract
Science can be examined from several standpoints, such as through a bibliometric analysis of the scientific output of researchers, research groups or institutions. However, there is little information about the advisor–advisee relationships or the academic supervision of researchers or between teachers and students. In this paper, we examine the results of the academic genealogy of PhD and Master's students working in Brazil, which was obtained from 737,919 curriculum vitae extracted from the Lattes Platform. Our findings bring to light three main sources of evidence related to the Brazilian academic genealogy: (1) the degree of interdisciplinarity between main areas of knowledge, (2) the structural features and evolving patterns with regard to both areas of knowledge and researchers, and (3) the patterns in the levels of training that affect the topological metrics. We conclude that academic genealogy offers a great opportunity to assess researchers and their areas of research from the perspective of human resource training.

Keywords Academic genealogy · Mentoring · Network analysis · Graphs

Introduction

Scientific publications are one of the principal means of disseminating significant knowledge to the academic community. Thus, it is usual to seek to characterize researchers through the impact of their output. Several approaches have been adopted for science in assessing the performance of researchers throughout their academic careers. Among these endeavors are those based on the number of publications produced, the citations each receives and the publications resulting from co-authorship.

Another way to improve science is through the training of new academics to enable them to make a contribution to scientific development. Moreover, studies have been carried

✉ Rafael J. P. Damaceno
rafael.damaceno@ufabc.edu.br

[1] Center for Mathematics, Computation and Cognition, Federal University of ABC, Av. dos Estados 5001, Santo Andre, SP 09210-580, Brazil

[2] School of Communication and Arts, University of São Paulo, Av. Prof. Lúcio Martins Rodrigues 443, São Paulo, SP 05508-020, Brazil

out to find alternative ways of evaluating researchers, which can be complemented by the existing bibliometric indicators. Evaluation measures have been devised that are related to the training of human resources at Master's and PhD levels and used to assess professors, research groups, and institutions. Several methods can be employed for studies involving mentoring relationships between academics, and these include those that are classified as academic genealogy.

Academic genealogy is the study of the intellectual heritage that is perpetuated between researchers, through formal mentoring relationships (Sugimoto 2014). Postdoctoral guidance and supervision can be used as input to evaluate academics, and groups of researchers brought together by institutions, regions and even a nation. These methods have attracted the particular interest of researchers because they make it possible to investigate the influences that a mentoring activity can exert both on an academic's career and consolidating the scientific community. In addition, academic genealogy studies may reveal factors that have been neglected by bibliometric studies since they focus on the importance of collaboration, and recognizing the different kinds of contributions made by mentors that are not always evident in co-authorships or citations (Sugimoto 2014). By being more than a sum of collective activities, scientific collaboration makes it possible to share the meaning of the tasks that are being carried out, which in the advisor–advisee relationship have a special place, by enabling training to occur (Sonnenwald 2007).

Regardless of the level, the kind of collaboration that usually takes place between teachers and students is a reality, although not considered by many authors to be scientific collaboration, owing to the asymmetry of their roles (Vanz and Stumpf 2010). On the other hand, this might be another reason to analyze influences in the training process, rather than concentrate on co-publications. This underlines the importance of academic genealogy as an alternative form of linkage, by enlarging the branch of relational bibliometrics.

Advisor–advisee relationships have been represented through networks of academic genealogy, in which the key players are denoted by nodes and the relationships between actors (mentoring or academic supervision) by directed edges. These structures, also called academic genealogy graphs, are an essential object of study to obtain a better understanding of the configuration and evolution of academics and group of researchers. They can be used to assist funding agencies, or any other academic institution, in the evaluation of how human resources are formed and complement the assessment of research output.

Different measures are taking for structuring the academic genealogy of specific areas of knowledge, such as the Mathematics Genealogy Project (Jackson 2007), which has around 231,000 registered PhD mathematicians gathered on a Web platform. NeuroTree (David and Hayden 2012) is another Web academic database, specifically designed for the area of Neuroscience (with 120,000 registered academics) and which later originated the Academic Family Tree (with 704,000 registered academics). This concentrates on the academic genealogy of 61 areas of Science, such as Computer Science, Chemistry and Theology.

Technically, these types of networks can be analyzed as a social network. The most common networks in the academic world are those designed for analyzing co-authorships, and investigating the possible correlations between the authors' connections and their productivity. However, in the last few years, owing to the increase in the availability of academic genealogy databases, studies have focused on the analysis of advisor–advisee relationships, and been closely linked to scholarly output.

In Brazil, some studies discuss the evaluation of academics by taking into account the nature of co-authorship. An example was carried out by Mena-Chalco et al. (2014) in which there is evidence of scientific collaboration; and by Rossi and Mena-Chalco (2014)

in which scientific mentoring is used as a substrate to develop evaluation metrics for academic networks. Also, Tuesta et al. (2015) investigated the existing correlation between the duration of the advisor–advisee relationship and the advisee's productivity (as measured in published journals) in the field of Earth sciences.

In this work, we explore the evolving pattern of genealogy graphs of academics registered in the Lattes Platform (available at http://lattes.cnpq.br), a mandatory curriculum vitae database for scholars who wish to take part in postgraduate programmes and are requesting grants or financial support. We used an algorithm created in a previous study (Damaceno et al. 2017), which was concerned with establishing a framework to draw information from the curricula data in academic genealogy graphs.

Our study examines the academic genealogy of PhD and Master's degree students, by analyzing graphs automatically and using the data gathered from the Lattes Platform. The graphs are characterized in terms of topological metrics that are especially designed to characterize academic genealogical graphs. The metrics were calculated for academics (nodes) both individually and collectively, according to their area of knowledge. To the best of our knowledge, this is the first study to address the entire network of researchers working in Brazil at Master's and Doctoral levels, and draws on data from more than 737,000 researchers. The central questions that this study seeks to address are as follows:

Q1 What degree of interdisciplinarity is there between the actors (e.g., those involved in areas of knowledge in Brazil, such as researchers), concerning academic genealogy?
Q2 What are the structural features of Brazilian academic genealogical graphs and how did these features evolve?
Q3 Are these features similar in different groups (e.g., graphs, areas of knowledge) and levels of training?

As an additional contribution to research, the graphs examined in this study and their respective attributes have been made available through a web platform called Acacia (available at http://plataforma-acacia.org). This was especially designed to enable genealogical information to be shared with members of the academic community and other stakeholders.

Related work

This section examines studies that have employed academic genealogy in several areas. Some of them make a correlation between the advisor's characteristics and advisee's performance. Others analyze the academic genealogy of areas of knowledge or individual scientists. The establishment of genealogical metrics to evaluate the formation of human resources has also been investigated. Finally, some works try to form an academic genealogy by gathering data from different information sources.

Malmgren et al. (2010) have studied the role of mentoring in a student's performance, by investigating the training of human resources. With the aid of the database of the Mathematics Genealogy Project, the authors found there was a correlation between the number of mentoring relationships of the advisors and the number of mentoring relationships of the advisees. In the other hand, Montoye and Washburn (1980) focused on the main contributors of a journal between the 1930s and 1976 (the study encompassed 135 people). They traced their academic ancestry and verified which advisors mentored students that contributed to that journal.

Chariker et al. (2017) investigated the mentoring patterns of academics who were winners of a Nobel prize. The authors used a subset of the Academic Family Tree consisting of 57,381 nodes, 402 of which were Nobel laureates. They found that the winners of a Nobel prize have a higher number of ancestors that had also won a Nobel prize when compared to those who did not win this prize. Using the method outlined by Wang et al. (2017), Liu et al. (2018) investigate the correlation between the advisors' academic characteristics and advisees' academic performance in Computer Science.

The authors found that the academic seniority (i.e., academic age) of the advisors plays an important role in the performance of the advisees (regarding the number of publications, citations and h-index). When there is an increase in the seniority of an advisor, there is initially an improvement in the performance of the advisees followed by a stage of a sustained standard and concluding with a decline.

Other studies have investigated specialist areas of knowledge, such as the work of Elias et al. (2016), which described the academic genealogy of Protozoology in Brazil, and the Kelley and Sussman (2007), on the academic genealogy of Primatology in the United States. Works like these are able to identify who are the "ancestors" in these fields as well, and ascertain in what way, human resource training is being undertaken. A further study by Sugimoto et al. (2011) used the academic genealogy to check the degree of interdisciplinarity in the area of Librarianship and Information sciences.

Other genealogical works involve the investigation of the academic lineage of a researcher (honorary genealogy). This is the case of the work carried out by Bennett and Lowe (2005), which examined the offspring of the American biologist George A. Bartholomew, who mentored one master's student and 39 doctorates and supervised five postdoctoral fellows. These students also formed mentoring relationships, making a total of 1200 individuals descending from Bartholomew. This kind of academic network can provide a literature review of an intellectual heritage, in a way that would not be possible through a conventional search in the literature.

The adaptation of bibliometric indicators was also suggested by Rossi et al. (2017) who designed a kind of h-index for genealogy graphs - the genealogical index. The definition of this is that a researcher has a genealogical index g if at least g of his descendants have at least g descendants.

David and Hayden (2012) developed the Neurotree, which is a collaborative genealogical database specifically designed for Neuroscience researchers. It currently contains about 115,000 registered researchers, a task undertaken by volunteers. The researchers are shown in graphic form, in which nodes represent the researchers and directed edges the advisor–advisee relationships. There is also a description of the biographical data of the researchers.

In a similar way to our study, Dores et al. (2017) formed graphs of the academic genealogy of researchers working in Brazil, using the Lattes Platform as their data source, but only taking into account the curriculum vitae of PhD researchers. The method used enabled 70,000 trees to be discovery, consisting of around 903,000 nodes and 1,444,000 edges. The most massive tree contains five thousand nodes, but 80% of them have less than 20 nodes.

Unlike these last two works, we automatically identify academic genealogy graphs from curricular data by including the curriculum vitae of both PhD and Master's researchers and employ name matching to merge duplicated researchers and relationships. Additionally, we analyze the data of the academic genealogy in Brazil on the basis of topological metrics, by examining the nodes both individually and grouped in main areas of knowledge.

A more recent approach has involved the use of bibliographic information sources to discover the advisor–advisee relationships. As argued by Wang et al. (2010, 2017) and Li

et al. (2017), the mentoring relationship is hidden in the co-authoring field of the information sources, since an advisor often writes papers with his/her advisees. The argument is that more traditional genealogical databases rely on manual tasks where people fill in electronic forms and thus possibly add errors or omit some critical information.

In this context, Wang et al. (2010) recommended a system based on a factor graph model which receives as input a set of papers from a bibliographic database and returns an academic genealogical graph with information about mentoring relationships. The experimental results showed an degree of accuracy of between 80 and 90%. Wang et al. (2017) designed Shifu, a deep-learning-based technique to extract the entire genealogical graph of the database used by the Digital Bibliography and Library Project, which includes more than 4.2 million papers of Computer Science. The method used can achieve a precision rate of 94%, which was validated by a smaller portion of that database (less than 3,300 pairs of advisor–advisees). Li et al. (2017) employed a technique based on the max.-confidence measure to infer the probability of a mentoring relationship exist, and this was more effective than the method proposed by Wang et al. (2010). Heinisch and Buenstorf (2018) employed machine-learning techniques to construct a dataset consisting of more than 20,000 German PhDs in Applied Physics and Electrical Engineering and analyzed the features of the advisors that were able to produce academic children who were responsive to advising. The authors used the Web of Science and the database of the German National Library for dissertation-based material to construct the graph.

Despite the pertinence of discovering the relationships of mentoring hidden in bibliographical sources, the relevance of using a curricular source is emphasized, where these formal relations are even certified by official academic institution. In addition, the national coverage of Lattes Platform, considering the continental dimension of Brazil, offers a significant amount of data. Finally, due to the fact that Tuesta et al. (2015) have analyzed, both the relationships of mentoring and bibliographical information for analysis of scientific productivity, we propose a broad, deep and dynamic look at the Brazilian academic genealogy, still involving its ancestors, whether foreign or not.

Background and hypotheses

An academic genealogy network can be represented by a directed graph, i.e., a mathematical framework that represents a set of elements that relate to each other and includes similar characteristics. Formally, a directed graph (**G**) consists of a finite set of nodes (V) and a set of edges (E) formed by ordered pairs (u, v), in which u and $v \in V$. The elements of the graphs may have different attributes, which differentiate them from each other. These attributes combined with the topological structure of the graph provide the necessary context for the study of the relationships between the represented elements.

For the purpose of this study, an academic genealogy network (or graph) is an interconnected group of researchers who spread knowledge through their advisor–advisee relationships. The actors who participate in it demonstrate their scientific activity through mentoring. Two actors form a relationship if one actor mentors another at one or more levels of training, i.e., master's and/or doctorates degrees. This type of network is usually represented by nodes and edges. Each node represents an actor, and directed edges represent the relationship between them. In our case, if there is more than one relationship between two actors, the oldest (in terms of years of completion) and highest degree is maintained.

By adopting a representation using nodes and edges, it becomes possible to conduct a data analysis based on graph theory. One can either calculate topological metrics individually, or for sets of nodes that have some common characteristic (e.g., area of knowledge, geographical region, or academic institution).

Hypotheses

In this study, by examining the topological metrics calculated for academic networks, we provide support for three hypotheses related to an evolving graph structure, i.e., the researchers considered individually and those grouped into areas of knowledge.

H1 From the interdisciplinarity that was noticed in previous studies concerning with co-authorship, we suspect that there is an intersection between the areas of knowledge and the formation process, since this has been observed in academic genealogical networks.

H2 Topological metrics are related to a researcher's post-graduate qualifications or, in other words, are based on the time an academic has had a degree.

H3 Different academic graphs have a similar pattern, with regard to topological metrics and interaction between areas of knowledge, including science in Brazil.

Regarding H1, the Coordination of Improvement of Higher Level Personnel, (CAPES) is an organization that coordinates post-graduate studies in Brazil, and divides subjects into nine main areas of knowledge. These are, as follows: Agricultural sciences (AGR), Biological sciences (BIO), Engineering (ENG), Exact and Earth sciences (EXA), Health sciences (HEA), Humanities (HUM), Linguistics, Letters and Arts (LIN), Applied Social sciences (SOC), and Others (OTH). These areas of knowledge establish some interrelationships with regard to human resource formation, i.e., academics of one academic field are able to mentor students from other areas. This characteristic was determined by Mena-Chalco et al. (2014) in the context of co-authorship networks, so it is also expected that it will occur in the domain of academic genealogy.

Concerning H2, recent studies have established a link between the advisee's performance and the features of the advisor (Malmgren et al. 2010; Chariker et al. 2017). In our study, an attempt is made to show that topological metrics, in particular those created for academic genealogical graphs, are related to the time when the academics obtained their degree, in the domain of Brazilian science.

Finally, H3 is based on the assumption that different graphs may behave in a similar way, with regard to topological metrics, and, in our case, show a related interaction between main areas of knowledge. Academic genealogical graphs have been formed by drawing on data from different sources, such as curricular and bibliometric databases. However, we believe, that even in the case where science is carried out in Brazil, these graphs might have similar characteristics.

Materials and methods

We investigated two academic genealogical graphs concerning research conducted in Brazil. The first only refers to doctorate degrees and the second to both master's and doctorate degrees. Both of them were formed from data taken from the Lattes Platform

and were characterized by three genealogical metrics, i.e., descendants, fecundity, and the genealogical index. In the graphs identified, each node (researcher) is formed of a researcher's identification code, full name, academic degree, year of academic degree, the main area of knowledge, current academic institution, and professional address. Each edge (advisor–advisee relationship) is formed of the node representing the identification code of origin (advisor), the target identification code (advisee), academic degree obtained and the concluding/graduation year.

Identification of academic genealogy graphs

Between July and August 2017, we collected and extracted data from the Lattes Platform, which contains the curriculum vitae of 737,919 researchers, of which 271,370 are PhDs and 466,549 Master's. Each academic in the Lattes Platform was able to update his/her curriculum at any time. As a means of ascertaining how updated the curricula sets are, we analyzed their dates at the time of the last update and found that about 80% of the academics had updated their curricula in the last 36 months, 50% being in the last eight months. Given the fact that master's and doctorate degrees last for two and four years in Brazil, respectively, most of the curricula have been updated. Figure 1 shows the frequency and cumulative percentage of curricula updated in the last 5 years.

A method devised in a previous study was employed to identify graphs from this curricular database (Damaceno et al. 2017). We started with curricular data providing information about personal life, scientific production and ancestors/descendants and ended with graphs in which each node represents a researcher, and each edge represents an advisor–advisee relationship between two researchers. The method is divided into six stages, as follows (see Fig. 2). The two initial stages involve collecting data and preprocessing the curricula. The third stage entails defining advisor–advisee relationships. Stages four and five involve name disambiguation through two forms of comparison, one with a global dictionary of names and node identifiers, and the other with the relative nodes of an given node. In stage six, the graphs are filtered to reduce the problem of noisy data (e.g., nodes with very short names were excluded).

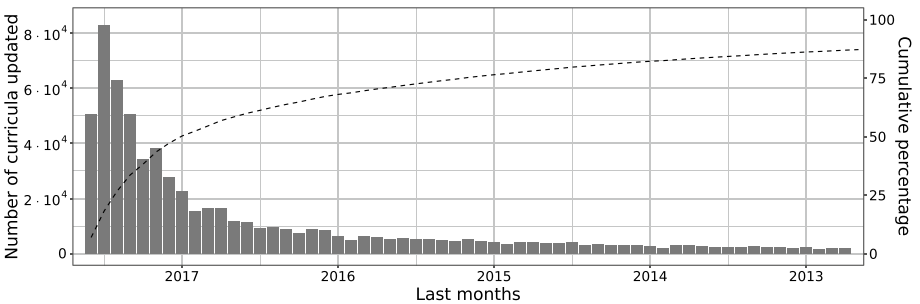

Fig. 1 Frequency and cumulative percentage of curriculum updates in the last 5 years. The first bar on the left represents August 2017

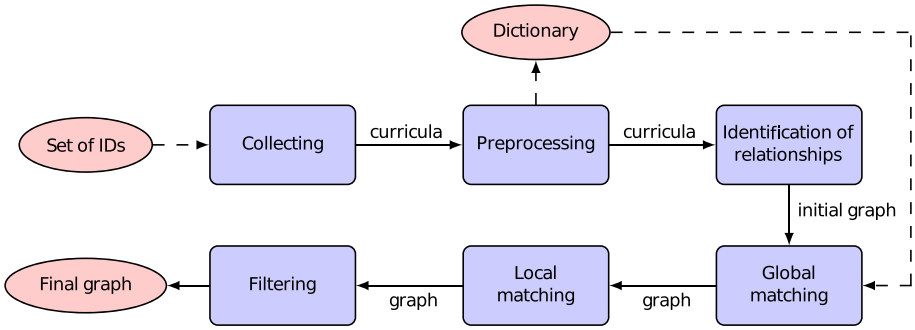

Fig. 2 A flowchart following the stages required to identify a genealogical graph from a curricular database and showing its set of identifiers for researchers

Collecting and preprocessing the curricula vitae

The algorithm uses as input the identification codes of the researchers of the Lattes Platform and produces as output a directed graph of academic genealogy. These identifiers, consisting of 16 digits, are unique to each researcher and were obtained through the platform itself, which provides a list of the identifiers of all the researchers registered, and shows the nature of their degree (e.g., graduate, master's or PhD). We collected the identifiers of all the masters and PhD researchers and obtained their curriculum vitae through the Lattes Extractor tool provided by the Lattes Platform. We preprocessed each curriculum vitae by removing accents and transforming all the characters to the lower case.

Identification of relationships

A curriculum is formed of three main parts, as follows. The first consists of personal information, i.e., the researcher's full name and curriculum vitae identification code (Lattes ID), main area of knowledge, name and address of institution. The second part includes the academic degree the research has, i.e., master's and doctorate degrees, year of graduation in that degrees, the advisor's name and Lattes ID. The last part includes the researcher's advisor–advisee relationships, i.e., academic degree resulting from the advising, year of the conclusion of the relationship and the advisee's name and Lattes ID. The algorithm generates a node in the graph for each curriculum vitae obtained. This node contains a unique code and the following information about the researcher: full name, Lattes ID, main area of knowledge and name and address of institution. Each academic qualification is converted to an edge with the advisor's node code, advisor's Lattes ID, advisee's node code, advisee's Lattes ID and year of the academic graduation. If the relationship of both graduation and academic mentoring does not contain the Lattes ID of the advisor or advisee's curriculum Lattes, respectively—which could result from a failure on the part of the academics to complete the information, we create an artificial node to represent this academic.

Global matching

At this stage, the algorithm matches the artificial nodes' full names with the full names and Lattes IDs of the academics. The names are matched to reduce noise, since the artificial nodes generated in the previous stage may represent an already existing academic in the Lattes Platform. This matching consists of comparing the complete name of each artificial node with a dictionary we compiled by extracting the full name and Lattes ID from all the curriculum vitae obtained during the preprocessing stage. Two similarity functions were used when comparing the names. The first is exact matching, in which all the characters of the compared names must match. If they did not match, we used the edit distance of one character (Levenshtein 1966), to represent the second function.

Local matching

We also carried out a local matching of names, that involved comparing the name of each artificial node with the names of its academic relatives, and using the same similarity functions as for the global matching. However, in this comparison, we only toke note of the first and last names of each academic (ignoring the middle name(s)). We defined two types of artificial nodes: "artificial node child" which is when an advisor with Lattes ID indicates as her/his advisee a researcher without Lattes ID, and "artificial node parent" when an advisee with Lattes ID indicates as her/his advisor a researcher without Lattes ID. The academic relatives of an artificial node child n_c are the nodes mentored by its ascending nodes except for n_c itself. The academic relatives of an artificial node parent n_p are the ascending nodes of all n_p's children except n_p itself (see Fig. 3).

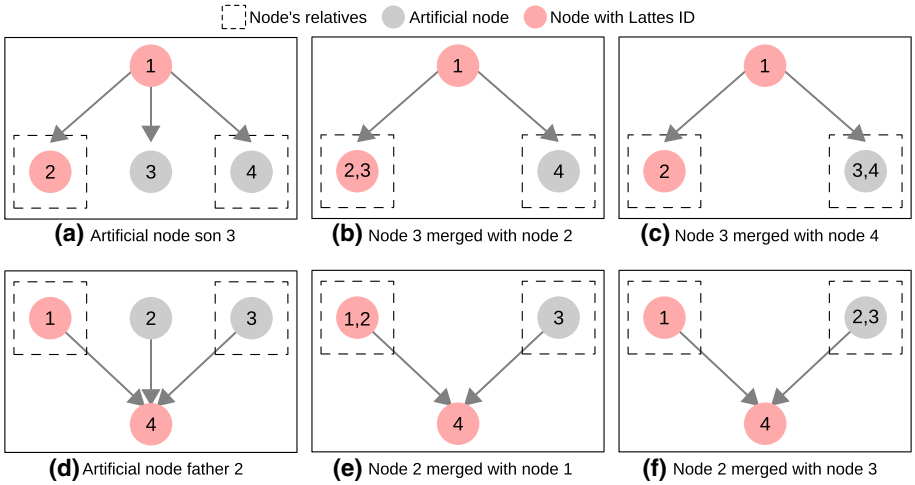

Fig. 3 Representation of the local matching of nodes. **a** An artificial node child (node 3), its parent (node 1) and its relatives (nodes 2 and 4). **b**, **c** The possible local matching of an artificial node child. **d** An artificial node parent (node 2), its child (node 4) and its relatives (nodes 1 and 3). **e**, **f** The possible local matching of an artificial node parent

Filtering

The last stage in the algorithm is to filter the graph, i.e., to remove nodes and artificial edges in accordance with two criteria. The first criterion removes nodes whose names contain less than six characters (e.g., "joao") or only one type of character (e.g., "aaaaaaa"), and their respective edges. The other criterion removes nodes whose names are in a list of invalid names (e.g., "to define", "waiting for definition" and "unaddressed") and their respective edges. We created this list of invalid names in advance in the preprocessing stage.

Indicators of academic genealogical graphs

We conducted an analysis of the academic training carried out in Brazil between 1927 and 2017, which encompassed both one level of formation (Doctorate—graph PhD) and two levels of formation (Master's and Doctorate degrees—graph MSc & PhD), which allowed us to test H3. Two analyses were conducted for each graph—one that groups the nodes according to their main areas of knowledge and another that analyzes nodes individually. In the case of the former, we were able to test H1, regarding the degree of interdisciplinarity of the areas from the standpoint of academic genealogy. In the case of the latter, we were able to identify the pioneering and/or distinguished scientific researchers in Brazil (whether foreign or not), through their genealogical metric measures for performance (H2).

Where the main areas of knowledge were grouped, it was possible to confirm how many advisor–advisee relationships have been established and experienced, and hence to check H1 (for example, how many advisor–advisee relationships exists between the areas Exact and Earth sciences and Engineering). The influence exerted by an area is defined as number of academic mentoring sessions carried out by a researcher from that area to another area. Moreover, the degree of influence experienced by an area can be defined as the number of academic mentoring sessions received by a researcher in that area from another area.

When analyzing the evolving pattern of academic genealogy graphs, account was taken to the time measurement called Academic Age (AA), which can be defined as the time (in years) since an academic has obtained his/her highest degree. Thus, given a researcher r with formation year f and the current year being c, his academic age is $\mathrm{AA}(r) = c - f$. For instance, defining $c = 2018$ (last year) and $f = 1995$ (for a given researcher R_1), the academic age of R_1 is 2018–1995 = 23. On the basis of the AA concepts, it was possible to determine H2 and answer the question of how science evolves.

In all the analyses, we calculated five genealogical metrics as follows: descendants, inverted descendants, fecundity, inverted fecundity and genealogical index. The following is a description of each of the measurements. Other genealogical metrics also can be calculated, such as those discussed by Rossi et al. (2018).

Descendants

Descendants (d^+) are the number of advisees that a researcher has mentored, either directly or indirectly, i.e., this metric takes into account mentoring at all levels of training, as an academic child, grandchild, great-grandchild and so on. With this metric, we were able to assess the impact of a researcher on the formation of the scientific community. "High descendants" values may indicate older researcher, i.e., a researcher with a higher academic

status/position, and low values may indicate young researchers. On the other hand, if the data is analyzed a opposite way, that is, by counting the number of ancestors, i.e., considering the parents, grandparents, great-grandparents and so on, there is and inverse descendant metric (d^-). In this case, the metric indicates the number of academics who exerted some influence on the training of an academic. High values can indicate a high degree of interdisciplinarity, and low values can represent academics predecessors the origins of the graph.

Fecundity

Fecundity (f^+) is the metric for measuring the number of direct descendants that a researcher has mentored. This metric represents the direct effect a researcher has had on the formation of scientific community. In Graph Theory terms, this is the out-degree of nodes. High fecundity values indicate that the researcher has had a significant impact on the community they belong to, since they represent the researchers that have advised most students. In contrast, low values indicate young researchers, i.e., researchers that have obtained their academic degree recently and still depend on an advisor to pursue their academic career. From an standpoint, it is worth examining inverse fecundity (f^-), that is, the number of "parents" that an academic owns. In terms of Graph Theory, this is the in-degree node. High values of inverse fecundity indicate a high degree of interdisciplinarity in the training of academics. Low or zero values represent the academic roots, that may be the predecessors of an area of knowledge in a graph, but could also represent information that is missing from the graphs.

Genealogical index

The genealogical index (gi) of a node v is defined as the largest number g of relationships between v (and adjacent to v) that have, at least, the same number g of relationships. It is a metric that takes into account the quantitative factor (number of descendants) and the qualitative factor (number of generations) (Rossi et al. 2017).

Results and discussion

This section discusses the results obtained from following topics: (1) the features of the graphs, (2) the influence exerted between the main areas of knowledge, (3) genealogical metrics for the main areas of knowledge, (4) genealogical metrics for nodes individually, (5) analysis of artificial nodes, and (6) the evolving pattern of genealogical metrics.

Features of the graphs

We identified two graphs, one representing only PhD researchers (abbreviated as PhD) obtained from the curriculum vitae of PhDs ($n = 271{,}330$) and the other representing Master's and PhDs (abbreviated as MSc & PhD), obtained from the curriculum vitae of Master's and PhDs ($n = 737{,}919$). Table 1 shows the number of ancestors, descendants, name matching, and nodes/edges filtered during the application of the algorithm described in this work.

Table 1 Number of curricula collected, ancestors and descendants identified, name matching and nodes and edges filtered

Number of	PhD	MSc & PhD
Curricula	271,330	737,919
Ancestors with Lattes ID	179,656	658,642
Ancestors without Lattes ID	100,051	329,420
Descendants with Lattes ID	177,322	678,320
Descendants without Lattes ID	67,598	427,359
Global matching of ancestors	39,649	151,622
Global matching of descendants	35,183	92,692
Local matching	16,412	69,805
Nodes filtered	600	1774
Edges filtered	2929	10,689

In the case of both graphs, there are a larger number of ancestors and descendants without Lattes ID, that were obtained in the initial phase of identifying relationships. The algorithm was capable of achieving more than 74,000 global name matching tasks (39,000 for ancestors and 35,000 for descendants) for the PhD graph, and over 243,000 for the MSc & PhD graph (151,000 for ancestors and 92,000 for descendants). The number of local name matching activities was lower, with more than 16,000 for PhD and over 69,000 for MSc & PhD.

The algorithm made it possible to reduce the number of nodes and edges representing the same entity in the graph. By only including the data needed to identify the MSc & PhD graph, before applying the matching, Table 1 shows that there were 2,093,741 relationships, 1,336,962 (64%) of them occurring solely between researchers with a Lattes ID (real relationships) and 756,779 (36%) between researchers with and without a Lattes ID (artificial relationships).

The global matching stage was responsible for matching 244,314 (32%) of the artificial relationships (756,779). This is the percentage of nodes without a Lattes ID that was

Table 2 Number of nodes and edges, density, mean and maximum values for out-degree and total degree, for both graphs and their respective giant components

Attribute	PhD		MSc & PhD	
	All nodes/edges	Giant component	All nodes/edges	Giant component
Nodes	381,306	233,666 (61.28%)	1,111,544	999,274 (89.90%)
Isolated nodes	4137 (1.08%)	–	5939 (0.53%)	–
Artificial nodes	97,458 (25.56%)	47,408 (20.29%)	351,386 (31.61%)	293,471 (29.37%)
Edges	348,315	238,761 (68.55%)	1,208,398	1,142,279 (94.53%)
Artificial edges	110,010 (31.58%)	54,684 (22.90%)	396,422 (32.81%)	334,305 (29.27%)
Standardized edges	4.79e−06	8.75e−06	1.95e−06	2.30e−06
Density	2.39e−06	4.37e−06	9.78e−07	1.15e−06
Mean degree	1.83	2.04	2.17	2.29
Maximum degree	131	131	422	422
Mean out-degree	0.91	1.02	1.09	1.14
Maximum out-degree	130	130	421	421
Connected components	38,296	–	48,199	-

converted to nodes with a Lattes ID. The local matching stage was also responsible for a large number of name matching occurrences, which made it easier to improve the graphs in terms of completeness, to a limited extent (e.g., by refining the number of direct descendants or ancestors of a node).

Table 2 shows the number of nodes and edges, mean and maximum values for in, out and total degree of nodes, for both the graphs and their giant component. The number of nodes (academics) is 381,306 for the PhD graph and 1,111,544 for the MSc & PhD graph.

With regard to the PhD graph, the highest out-degree is 130, i.e., the same academic has mentored 130 different students at the Doctoral level. In the case of the MSc & PhD graph, this value rises to 421, since in this case it also takes account of Master's degrees.

The number of artificial nodes for the PhD graph is more than 97,000 researchers, and for MSc & PhD, it is over 351,000. These are the academics who either have no curriculum vitae on the Lattes Platform or do, but were unable to be linked to an existing curriculum by the algorithm (possibly because the name was written in a different way from what is registered by the real researcher). The number of isolated nodes for the PhD graph is more than four thousand academics, and for MSc & PhD, it is almost six thousand researchers. These are academics who did not register any relatives in their curriculum.

Table 2 also displays data from the giant component of both graphs, whose highest value for out-degree remained the same as that for the total number of nodes. In the case of the PhD graph, the giant component comprised 61.28% of the nodes and 68.55% of the edges. In the case of the MSc & PhD graph, these values rise to 89.9% for the nodes and 94.53% for the edges. These elevated values can probably be explained by the relationships between Master's and PhDs academics in MSc & PhD. A possible reason for this behavior is that there are a large number of PhDs that do not have mentored doctorates yet.

We also count the number of academics by generation, i.e., there is a rise in each on the graph until there is no possibility of rising further (the highest distance from the source node to the most distant ancestor is the generation of the source node). The PhD graph has nine generations of academics with most of them (225,174—59.05%) being between generations three and four. In the case of the MSc & PhD graph, there are thirteen generations with most of the academics (768,050—69.1%) being between generations three and five.

Table 3 Number and percentage of researchers for each main area of knowledge

Acronym	Main area of knowledge	Researchers	
		PhD	MSc & PhD
AGR	Agricultural sciences	26,479 (6.94%)	51,675 (4.65%)
BIO	Biological sciences	35,417 (9.29%)	66,033 (5.94%)
ENG	Engineering	24,890 (6.53%)	58,103 (5.23%)
EXA	Exact and Earth sciences	38,682 (10.14%)	81,263 (7.31%)
HEA	Health sciences	44,952 (11.79%)	105,364 (9.48%)
HUM	Humanities	44,634 (11.71%)	119,837 (10.78%)
LIN	Linguistics, Letters and Arts	16,241 (4.26%)	43,935 (3.95%)
SOC	Applied Social sciences	29,368 (7.70%)	103,577 (9.32%)
OTH	Others	2672 (0.70%)	9120 (0.82%)
UND	Undefined	117,971 (30.94%)	472,637 (42.52%)
–	All	381,306 (100.00%)	1,111,544 (100.00%)

Influence between main areas of knowledge

This section examines the influence exerted and experienced in the main areas of knowledge. Table 3 provides an overview of the size of each of these areas, and shows the number of researchers grouped by main area of knowledge and the number of advisor–advisee relationships established by each area. We classify as Undefined (UND) artificial nodes and researchers that do not register any field of knowledge.

The main area of knowledge that has the largest number of researchers is Health sciences with 44,952 researchers (11.79%) for the PhD graph and Humanities with 119,837 researchers (10.78%) for the MSc & PhD graph (ignoring OTH and UND in both cases). The area of knowledge that has fewest researchers is Linguistics, Letters and Arts for both graphs (ignoring OTH and UND) with 16,241 (4.26%) researchers for PhD and 43,935 (3.95%) for MSc & PhD. A much larger number of researchers do not have an area of knowledge that is defined, 117,971 (30.94%) in PhD and 472,637 (42.52%) in MSc & PhD. We included researchers without a Lattes ID in this group. The high values obtained can be attributed to the number of artificial nodes created in both graphs (see Table 2), which was an essential feature of our algorithm for identifying foreign and pioneering researchers.

To illustrate the general influence exerted and experienced by each main area of knowledge, we created a visual MSc & PhD graph (Fig. 4), in which the color of the node represents its field of knowledge. We excluded from the graph any nodes with an in-degree equal to one and out-degree equal to zero, since this represents researchers that have only one advisor and no academic children, and nodes without a defined area of knowledge. This graph contains 197,034 nodes and 296,889 edges.

The graphic visualization in Fig. 4 shows a larger number of relationships between the nodes within the same area of knowledge (the same colors are close to each other). Moreover, a considerable number of nodes with different areas of knowledge have close

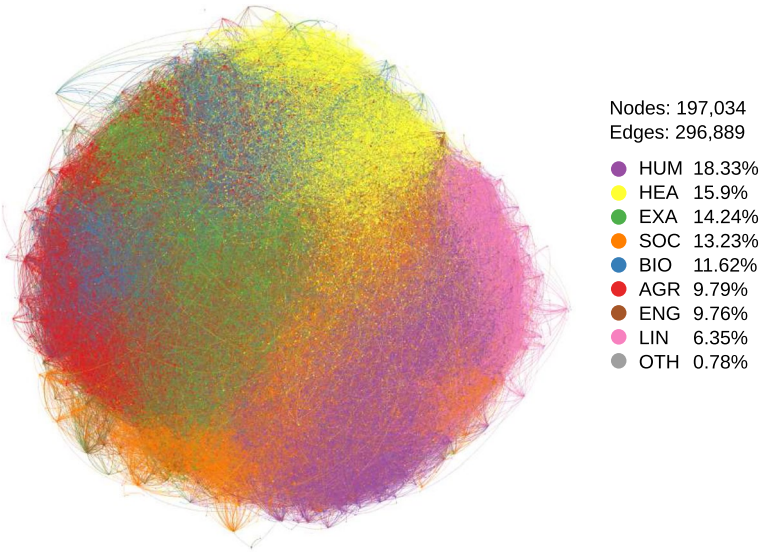

Fig. 4 The Brazilian academic genealogy. This illustration represents 197,034 researchers (nodes) and 296,889 advisor–advisee relationships (edges) registered on the Lattes Platform. The OpenOrd algorithm was used for the visual representation (Martin et al. 2011). (Color figure online)

relationships with each other. For example, HUM (dark purple) is near SOC (orange) and LIN (light purple), BIO (blue) is near AGR (red) and HEA (yellow). The large number of mentoring relationships reflected in the proximity between the nodes representing academics from different areas supports H1, i.e., there is a certain degree of interdisciplinarity that can be seen in the training schemes.

We counted the number of mentoring relationships that occurred between the areas of knowledge for both the PhD and MSc & PhD graphs (see Table 4). As expected, a main area of knowledge establishes relationships largely with the same area (main diagonal of the tables), but in each case there is a second area with also exerts and experiences a considerable influence. For example, researchers of SOC establishes relationships largely with researchers of SOC (71,720 for MSc & PhD), but also has a considerable number of relationships with researchers of HUM (16,378 for MSc & PhD). This tendency is found for both the PhD and MSc & PhD graphs. We ignored the main areas of knowledge related with "Others" and "Undefined" since the former has a small number of researchers and relationships (see Table 3) and the latter represents those researchers who either did not register any main area of knowledge or are artificial nodes. In addition, we created 16 radar charts, based on Table 4 and designed to visualize the similarities between the two graphs more clearly. These show the influence exerted on each area of knowledge for both the PhD and MSc & PhD graphs (see Appendix Fig. 9).

Table 4 Influence exerted on (a) the PhD graph (a) and (b) the MSc & PhD graph

	AGR	BIO	ENG	EXA	HEA	HUM	LIN	SOC	Total
(a) PhD									
AGR	18,567	2619	543	732	631	249	15	253	23,609
BIO	2868	23,728	252	773	4749	547	28	165	33,110
ENG	723	446	16,813	3998	621	381	41	1427	24,450
EXA	629	1025	1916	23,282	874	995	48	465	29,234
HEA	339	2635	106	365	30,364	1133	61	343	35,346
HUM	182	300	217	1194	1913	30,025	1464	3839	39,134
LIN	6	8	16	33	214	1126	10,999	745	13,147
SOC	135	65	453	307	266	1499	535	15,415	18,675
Total	23,449	30,826	20,316	30,684	39,632	35,955	13,191	22,652	216,705
(b) MSc & PhD									
AGR	47,705	7218	1531	2224	2300	1076	80	1175	63,309
BIO	6720	54,226	735	1963	13,237	2050	127	743	79,801
ENG	1771	1413	50,261	11,575	2252	1656	233	6230	75,391
EXA	1658	3054	6056	63,386	2691	4144	239	2384	83,612
HEA	872	5973	347	989	82,738	4273	292	1435	96,919
HUM	714	1428	823	5044	7976	100,939	6171	16,378	139,473
LIN	33	46	63	182	704	4869	35,782	3129	44,808
SOC	672	286	2145	1907	1394	7238	2276	71,720	87,638
Total	60,145	73,644	61,961	87,270	113,292	126,245	45,200	103,194	670,951

In the last row, the figures in the column represent the total influence experienced by the main area of knowledge. In the last column, the figures in the row represent the total amount of influence exerted by the main area of knowledge

The data shown in Table 4 supports H1, i.e., the main areas of knowledge have a certain degree of interdisciplinarity. Researchers working in a main area of knowledge have mentored students actively involved in other main area of knowledge. This result is similar to what was obtained for the same curricular platform, when it was analyzed from the standpoint of co-authorship networks, as confirmed by Mena-Chalco et al. (2014). While comparing the two graphs, it is also possible to see a similarity in the proportion of relationships between the same areas of knowledge. This suggests that H3 is plausible in the context of interdisciplinarity, i.e., different academic graphs display a similar pattern for this concept.

In seeking to determine how the mentoring grouped by the main areas of knowledge evolved in the period 1987–2017, we selected a subset of the PhD and MSc & PhD graphs, that only included the mentoring relationships for this period. We quantified the number of completed mentoring sessions (with regard to the advisees) related to the eight areas for each area of knowledge and each of the years (with regard to the advisors) (see Fig. 5). Figure (bar graph) 5a represents the relationships established between advisors of AGR and advisees of all the areas (except UND), Fig. (bar graph) 5b advisors of BIO with advisees of all the areas (except for UND), and so on.

The caption for each bar graph shows the mean percentage (in this period) of advisees belonging to each of the main areas of knowledge. The bar graphs show that the main area of the advisees is in a significant part of the same area of knowledge as that

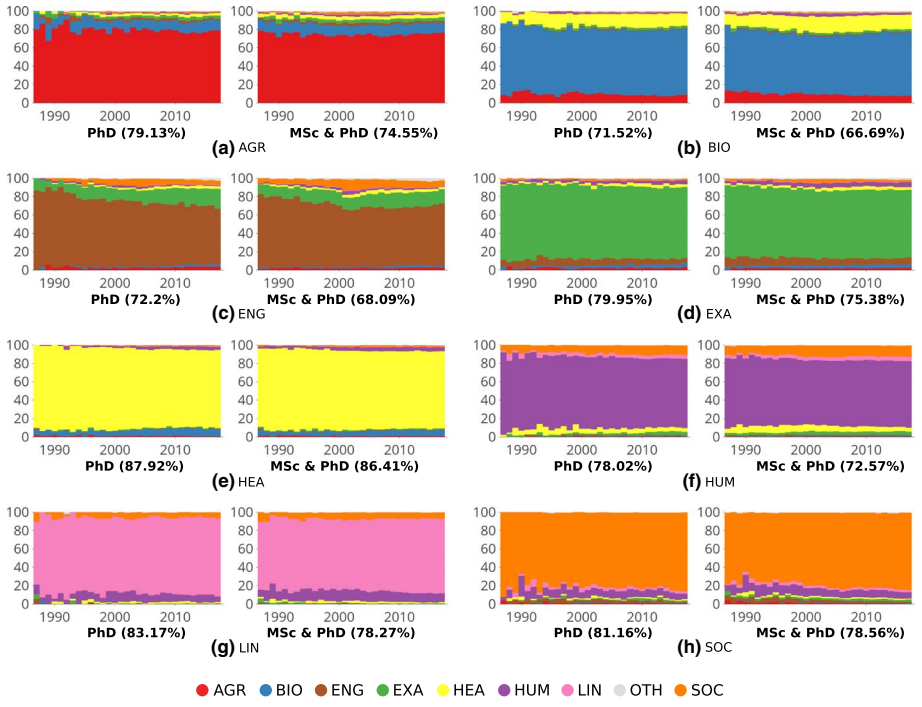

Fig. 5 Percentage of academics grouped by main areas of knowledge mentored in the period 1987–2017 for the PhD and MSc & PhD graphs. The *x*-axis represents the year an academic obtained a doctorate or master's degree. The *y*-axis represents the percentage of advisees' area of knowledge that was influenced by the advisors' area. (Color figure online)

of the advisors. In these cases, the mean percentage values range from 71.52% (BIO) to 87.92% (HEA) for the PhD graph and from 66.69% (BIO) to 86.41% (HEA) for the MSc & PhD graph.

In some cases, such as ENG, a third area (SOC) also has a larger number of relationships (after EXA). In contrast, Mena-Chalco et al. (2014) did not find many partners willing to collaborate with this third area. Another specific feature was observed in the LIN area, namely that it was more common to find co-authorship among HUM researchers, in contrast with our results, that showed a smaller percentage. These features of the Brazilian scientific community suggest that academic genealogical graphs could offer information that would not be available from co-authorship graphs.

When a comparison is made between both the PhD and MSc & PhD graphs, a similar trend is found, which supports H3 in the context of the last 30 years. The addition of the master's degree to the PhD graph led to a decline in the percentage of researchers from the same area of knowledge in the subgraphs. This results suggest that there is a higher degree of interdisciplinarity in the master's degree category than in the others.

With a few exceptions (e.g., area of knowledge of AGR in 1989, LIN in 1992, and others), the evolving pattern of the bar graphs shows a steady number of advisors from one area of knowledge, mentoring advisees in the same or other areas of knowledge,

Table 5 Metrics for descendants (d^+ and d^-), fecundity (f^+ and f^-) and genealogical index (gi) for the graphs: (a) PhD and (b) MSc & PhD

	d^+		d^-		f^+		f^-		gi	
	avg	max	avg	max	avg	max	avg	max	avg	max
(a) PhD										
AGR	1.95	450	2.73	21	1.16	63	1.03	4	0.07	9
BIO	2.74	1523	2.76	16	1.29	89	1.04	5	0.09	11
ENG	2.17	368	2.43	20	1.32	130	1.06	4	0.08	9
EXA	1.96	879	2.46	15	1.10	96	1.04	5	0.08	12
HEA	1.92	1164	2.86	22	1.01	76	1.03	5	0.08	11
HUM	2.08	1136	2.55	13	1.12	130	1.02	5	0.07	12
LIN	1.98	1050	2.70	13	1.06	124	1.02	4	0.07	11
SOC	1.42	913	2.51	18	0.81	97	1.02	5	0.05	8
OTH	0.27	39	2.00	10	0.22	35	0.97	3	0.01	2
UND	3.11	2022	1.76	17	0.50	53	0.64	5	0.14	12
(b) MSc & PhD										
AGR	4.91	2103	6.76	68	1.87	148	1.35	6	0.12	17
BIO	5.82	7304	5.56	71	1.78	164	1.30	7	0.13	24
ENG	6.50	4042	6.13	74	2.24	255	1.31	9	0.14	21
EXA	4.94	4092	6.02	72	1.67	161	1.32	7	0.13	16
HEA	3.59	3312	5.87	71	1.41	162	1.25	7	0.09	14
HUM	5.28	10,989	6.22	81	1.73	421	1.28	12	0.11	29
LIN	4.50	5670	6.00	69	1.50	211	1.27	14	0.10	21
SOC	3.79	5432	6.20	70	1.38	327	1.24	9	0.07	19
OTH	0.56	178	4.86	66	0.39	77	1.16	4	0.01	5
UND	6.19	18,679	4.39	77	0.34	152	0.82	14	0.09	23

over the past thirty years. The bar graphs indicate that the interdisciplinarity in the human resource training followed a regular pattern in that period, which strengthens H1.

Genealogical metrics for main areas of knowledge

This section examines the values of the genealogical metrics for the main areas of knowledge. Table 5 shows the mean and maximum values for the metrics of descendants, fecundity and the genealogical index. The main area of knowledge with the highest mean number of descendants is BIO (2.74) for the PhD graph, and ENG (6.50) for the MSc & PhD graph, in both cases ignoring UND. The maximum value is 1,523 descendants for the PhD graph (BIO) and 10,989 for the MSc & PhD graph (HUM), ignoring UND. The researchers in these areas have the largest numbers of academic successors, that is, children, grandchildren, great-grandchildren, and so on. This metric values provide evidence of the significant role played by human resource training in Brazil in those areas of knowledge.

Engineering has the highest mean fecundity values (f^+), 1.32 and 2.24, for the graphs PhD and MSc & PhD, respectively. ENG and HUM have the maximum value, 130 academic children each, for the PhD graph. Regarding MSc & PhD, HUM have the highest value, 421 academic children. These main areas of knowledge hold academics who have carried out a high number of mentoring at the master's and doctoral levels, being prominent influences in the formation of human resources. The age of the area of knowledge could influence these values, i.e., oldest areas of knowledge had more time to advising when compared with the newest areas of knowledge.

In the PhD graph, BIO has the highest average value (0.09) for the genealogical index (gi), and in the MSc & PhD graph, ENG has the highest average value (0.14), ignoring UND. The maximum value is 12 for both EXA and HUM in the PhD graph and 29 for HUM, in the MSc & PhD graph. These main areas of knowledge have researchers with productive academic children regarding human resource training. For example, the academic who has the highest genealogical index value (29) in HUM has 29 academic children whom each has a minimum of 29 children each.

Genealogical metrics at researcher's level

We also calculated the metrics for the nodes individually. As a subset of all of our results, Table 6 shows the values of the metrics descendants and fecundity, as well as the genealogical index for the 15 researchers with the highest values, by taking note of the PhD and the MSc & PhD graphs. In the case of the metric descendants with the 15 highest values (10 for the PhD graph, and 12 for the MSc & PhD graph), these represent researchers who do not have a curriculum vitae in the Lattes Platform. This metric makes it possible to investigate who are the predecessors or roots of the graphs. The researcher who obtained the highest fecundity value was Martins, J., with 2,022 (PhD) and 18,679 (MSc & PhD) descendants throughout his academic career.

One of the highest values for the genealogical index metric was obtained by an artificial node, (the researcher Bori, CM), and had a genealogical index equal to 12 (PhD graph). With regard to human resource formation, the highest number of doctoral degrees (Fialho, FAP, and Ebecken, NFF) is equal to 130. At master's and doctoral levels, this rate rises to 421 advisor–advisee relationships, (also from the researcher Fialho, FAP). In a similar way to this analysis, in Table 7 (in the Appendix), we list the top 15 for each metric, but only take into account artificial nodes. It should be noted that the artificial nodes have significant

Table 6 Top 15 values for the metrics descendants (d^+ and d^-), fecundity (f^+ and f^-) and genealogical index (gi): (a) PhD and (b) MSc & PhD. Researchers marked with * are artificial nodes

Researcher	d^+	Researcher	d^-	Researcher	f^+	Researcher	f^-	Researcher	gi
(a) PhD									
Martins, J*	2022	Silva, MF	22	Fialho, FAP	130	Silva, AC*	5	Bori, CM*	12
Brieger, FG*	1769	Reinehr, CO	21	Ebecken, NFF	130	Rassi, DM	5	Witter, GP	12
Dreyfus, A*	1524	Bortoluzzi, AC	20	Braga, MLS	124	Borges, JC*	5	Martins, J*	12
Pavan, C	1523	Oliveira Filho, JAJP	18	Cosenza, CAN	111	Pugliesi, M	5	Gottlieb, OR	12
Bori, CM*	1320	Novello, Z	18	Carvalho, PB	97	Oliveira, MA*	5	Pais, CT	11
Lima, JP*	1165	Marchesi, C*	17	Gottlieb, OR	96	Martins, MCB	5	Saviani, D	11
Böhm, GM	1164	Marson, LCG	17	Maciel Filho, R	91	Pereira, RGFA	5	Krieger, EM	11
Saviani, D	1136	Artur, AG*	16	Azevedo, JL	89	Almeida, RL	5	Cohn, G	11
Carlini, ELA	1115	Gomes, ASG	16	Souza, W	89	Carlini Sobrinho, T	5	Azevedo, JL	11
Pereira, L*	1107	Feihrmann, AC	16	Carvalho, EA	88	Stockmeier, TE	5	Chaui, MS	11
Pottier, B*	1100	Rassi, DM	16	Witter, GP	88	Gomes, AD*	4	Neves, EFA	10
Michel, L*	1051	Oliveira, JC*	16	Pinto, JCCS	85	Artur, AG*	4	Izquierdo, IA	10
Pais, CT	1050	Azevedo, LCP*	16	Silva, EL	81	Machado, A	4	Rodrigues, JA	10
Franca, ED*	1030	Tres, MV	16	Lopes, RT	81	Guarneri, AA	4	Travassos, LRRG	10
Roper, JA*	1025	Muniz, PR	16	Eckert, H	78	Guimarães, ACR	4	Braga, MLS	10
(b) MSc & PhD									
Martins, J*	18,679	Pimenta, TFF	81	Fialho, FAP	421	Silva, AC*	14	Saviani, D	29
Saviani, D	10,989	Barbosa, DNL*	77	Carvalho, PB	327	Silva, MA	14	Azevedo, JL	24
Pereira, L*	8138	Caruzzo, A	74	Cosenza, CAN	255	Oliveira, PC	12	Martins, J*	23
Bori, CM*	7578	Madeira Júnior, AG	74	Ebecken, NFF	219	Oliveira, PR*	12	Braga, MLS	21
Franca, ED*	7423	Pimentel Neto, AA	73	Braga, MLS	211	Pereira, ACS*	11	Barcia, RM	21
Dreyfus, A*	7305	Kramer, AHFR	72	Landau, L	203	Oliveira, CA*	10	Bori, CM*	20
Pavan, C	7304	Albuquerque, FCB	72	Diniz, MH	203	Oliveira, JC	9	Witter, GP	20
Fernandes, F*	6851	Dias, GKG	72	Beuren, IM	202	Santos, JC	9	Warde, MJ	20
Brieger, FG*	6369	Araújo, KF	72	Witter, GP	197	Oliveira, MA*	9	Catelli, A	19

Table 6 (continued)

Researcher	d^+	Researcher	d^-	Researcher	f^+	Researcher	f^-	Researcher	gi
Ianni, O*	6288	Domingues, LA	72	Lezana, ÁGR	189	Silva, PR	9	Pais, CT	19
Pottier, B*	6206	Costa, LCA	72	Rodriguez, AM	187	Silva, CR*	8	Cunha, CJCA	19
Moisés, M*	6005	Lima, MACB	72	Quelhas, OLG	186	Souza, MA	8	Velho, GCA	18
Michel, L*	5671	Cunha, NS	72	Fiorillo, CAP	183	Oliveira, MA*	8	Izquierdo, IA	18
Pais, CT	5670	Kramer, RHFR	72	Mukana, DMH	177	Silva, MA*	8	Lane, STM	18
Holanda, SB*	5639	Gonçalves, TJM	72	Kliemann Neto, FJ	170	Martins, MCB	8	Santanna, AR*	17

rates, and show high values even in Table 6 which includes all types of nodes. This analysis was possible since our approach does not ignore academics without a Lattes ID.

Analysis of artificial nodes

For a better understanding of who are the artificial nodes, or what types of researchers they represent, we conducted an analysis of in-degree and out-degree (fecundity and inverted fecundity, respectively) for these nodes. As a result, we found that most of them have an in- or out- degree equal to one (see Figure 6). Since they are researchers who do not have a curriculum vitae in the Lattes Platform, this subset of nodes probably belongs to foreign researchers, which could thus enable us to respond to Q2 in the context of foreign researchers. Note that as shown in Table 6, these nodes both exerts an influence on Science in Brazil and are also influenced by it.

By way of illustration, F. G. Brieger (academic age 74) is the most significant foreign artificial node with regard to the metric for descendants (d^+) with a total of 1769 and 6369 descendants for the PhD and MSc & PhD graphs, respectively. This is the German geneticist Friedrich Gustav Brieger who came to Brazil in the late 1930s. Concerning the MSc & PhD graph, about 32.2% of his descendants are from his main area of knowledge (BIO), and 27.3% are from AGR. With regard to the subareas of knowledge, about 18% of his descendants are from Agronomy, and 15.4% are from Genetics (the area of Brieger). If we only examine Brieger's direct descendants, all of them (two researchers) are in the area of Genetics. These data illustrate the influence exerted by Brieger on the field of Agronomy and Genetics in Brazil.

Another essential foreign node is the French scientist Bernard Pottier (academic age 48), a linguistic who is a specialist in semantics, (in the area of Humanities), who has 1,100 descendants in the PhD graph and 6,206 in the MSc & PhD graph. We also counted the number of descendants of Pottier and divided their areas of knowledge, for the MSc & PhD graph. About 43% of his descendants in this graph are from his main area of knowledge (LIN), and 9.9% are from SOC and HUM. With regard to the subareas of knowledge, 21% of his descendants are from Letters, and 20% from Linguistics. Keeping to the direct descendants (13 in total) of Pottier, ten of them are from Linguistics, two are from Letters, and one is from Theology. These direct descendants had mentored other researchers who, in turn, had mentored others, and this helped to consolidate the areas of Letters and Linguistics in Brazilian science.

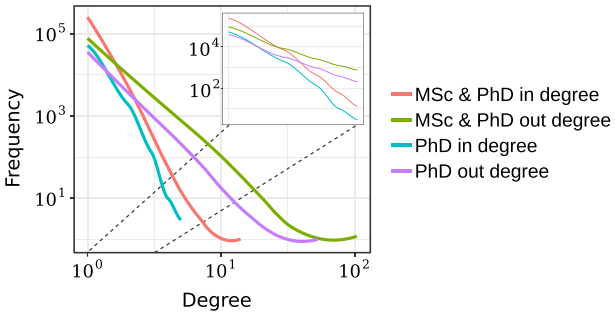

Fig. 6 Frequency of artificial nodes by in-degree and out-degree for the PhD and MSc & PhD graphs. The small square shows the intervals between degrees 0 and 5. (Color figure online)

Both artificial nodes were cited as mentors by Brazilian academics, who subsequently were very productive human resources trainers. These nodes are examples of foreigners who exerted/exert a direct and indirect influence on the scientific output of Brazil, with regard to human resource training. A high percentage of their descendants are from several subfields of knowledge, which shows that H1 occurs even in the early stages of formal mentoring in Brazil—F. G. Brieger is one of the oldest academics, in terms of academic age.

There was evidence that these foreigners have influenced Science in Brazil, that is, the number of academic children who have graduated in Brazil and how these have played a role in the genealogical metrics. However, we only conducted a brief analysis of two of the 351,386 artificial nodes that exist in the MSc & PhD graph. The analysis of the other artificial nodes involves determining exactly who they are and what country they come from. This analysis requires a detailed assessment and will be the focus of future work, which will allow us to respond to Q2 in depth.

Evolving patterns of the genealogical metrics

In this section, there is an analysis of the graphs of academic genealogy that are based on the academic ages of the researchers. Figure 7 shows the evolutionary pattern and the running average of the number of researchers by academic age. As expected, all the main areas of knowledge display the same trend, that is, there is a reduction in the number of

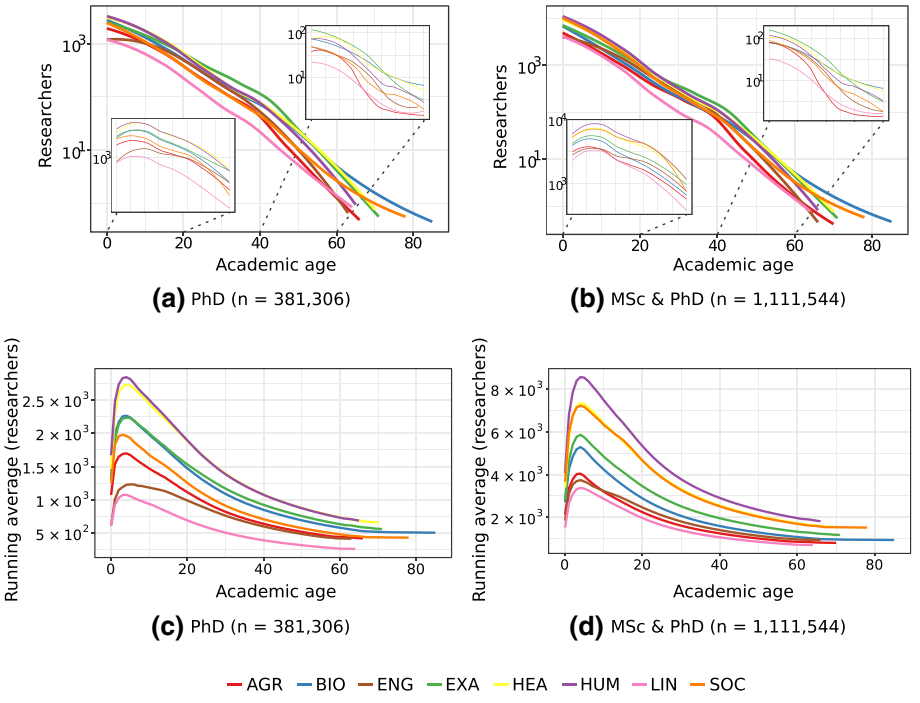

Fig. 7 Evolutionary pattern (**a**, **b**) and running average (**c**, **d**) of the number of researchers for the graphs based on academic ages. The small squares are spaced out for a better visualization. (Color figure online)

researchers as the academic age increases. This trend suggests that the number of researchers has increased in Brazil, for both levels of training (master's and doctorates) and in the eight areas shown in the chart.

In Fig. 7a, the small rectangle at the bottom left-hand side, shows a considerable increase in the number of researchers in Engineering, i.e., they suggest that in the last 5 years there has probably been a significant rise in the formation of that area of knowledge. Biological sciences is the area of knowledge with the largest number of older researchers for both graphs, and Humanities is the area of knowledge with the largest number of younger researchers. In Fig. 7c, d the curves have the same pattern concerning the number of academics. There is a peak between the academic years zero and five, which corroborates the view that science in Brazil is still young, i.e., there are more academics in the early stages of their career than senior academics, regardless of their main area of knowledge.

The large number of young researchers in Brazil can also be explained by the expansion of the number of postgraduate programmes, run by the Brazilian government in the last two decades. In 1996 there were 630 programmes and 2854 PhD researchers graduated,

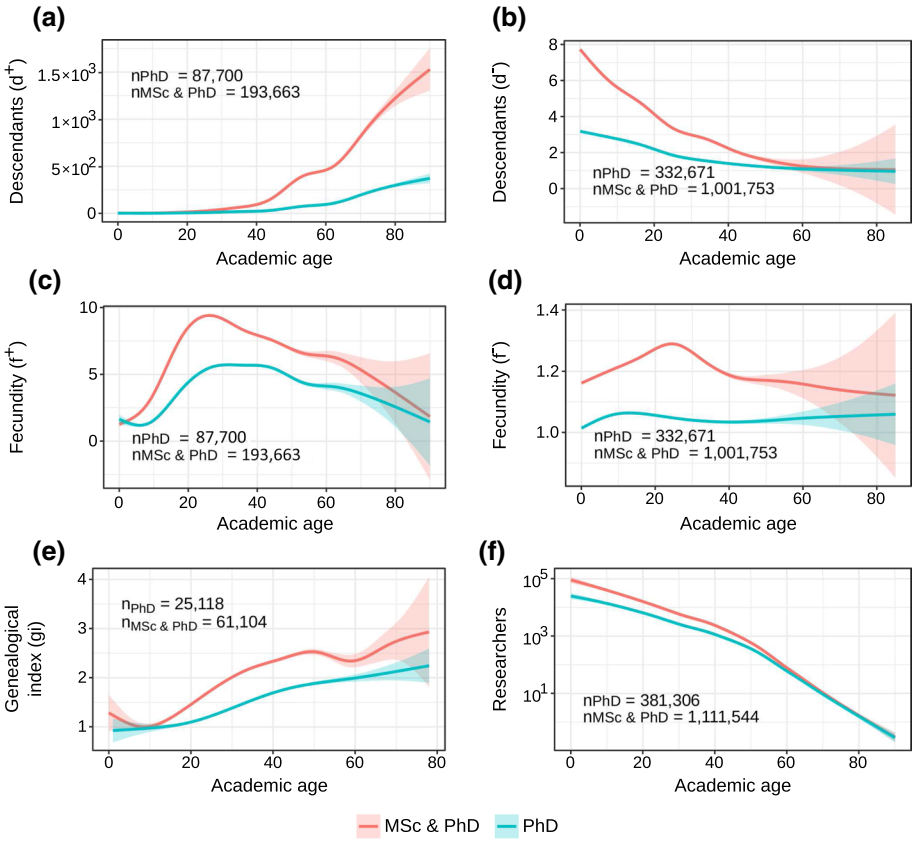

Fig. 8 Evolving pattern of the metrics related to descendants, fecundity and the genealogical index, and the number of researchers for the PhD and MSc & PhD graphs. (Color figure online)

while in 2016 this number was 1954 programs, involving 16,729 PhD researchers (Center for Strategic Studies and Management Science, Technology and Innovation 2016).

With regard to the evolving pattern of genealogical metrics, Fig. 8 shows metrics descendants, fecundity, and the genealogical index, and the number of researchers for the graphs in question. In Fig. 8a–e we omitted academics with a value equal to zero for each analyzed metric. The shadows represent a confidence interval of 95% and the lines follow a conditional smooth mean that describes those curves. These results strengthen support for H3, i.e., the topological metrics are similar to those in the graphs witch only include academics at the doctorate level and with master's and PhD degrees. It also supports H2, i.e., topological metrics which vary in accordance with academic ages.

Figure 8a shows that younger academics (e.g., those with an academic age of less than 20 years), have fewer descendants. From the age of 40, the number of descendants grows steadily and peaks at around 80 years. This behavior is noticeable in both lines but in on a smaller scale in the case of PhD. The curve for both graphs begins to grow from the academic age of 20 when researchers begin to complete their first advisor–advisee relationships at master's and doctoral levels. When another twenty years is added, the grandchildren begin to appear, and this causes the curve to proliferate. This mean that younger academics begin to have direct and indirect descendants at master's and doctoral levels in the period of 20–40 years of age.

An inverted trend is seen in Fig. 8b for metric inverse descendants (d^-), when younger researchers have more direct and indirect ancestors, and, as the ages go by the number of ancestors declines. This trend may indicate an increase in interdisciplinarity in the area of human resource training since having more ancestors increases the chance of being mentored by academics from different field of knowledge. Advisees are able to gain experience and knowledge from different standpoints and in the future integrate areas or even create new disciplines.

Fecundity (f^+) and inverse fecundity (f^-), shown in Fig. 8c, d, both reach a peak at around 25 years of academic age (the only exception is that of the f^- PhD curve which is smoother, and peaks earlier, at around 10 years). Unlike the other metrics, both training levels show the same trend, since this is not a cumulative indicator.

In the genealogical index (gi), shown in Fig. 8e, there is a growth for both graphs from the academic age 20, which suggested that it is at this stage of academic maturity that academics begin complete their first mentorships. Moreover, it is between the academic ages 40 and 70 that the peak is reached for the indirect training of human resources. Figure 8e shows the number of researchers by academic age. There is a relatively young body of scientist in Brazil since most of the academics have an academic age of less than 20 years.

In the analysis by academic ages, we found that each metric behaves in a different way. It was possible to show rising curves for the descendant (d^+) and genealogical index (gi) and declining curves for the inverse of the descendant (d^-). In the case of the fecundity (f^+) metrics and inverse fecundity (f^-), there was evidence of an initial rise until the academic age was the equivalent of 25 and a declining value for greater academic ages. This trend supports H2, i.e., that topological metrics are related to the researcher's academic ages.

Conclusion

This study seeks to examine science in Brazil from the standpoint of academic genealogy. We analyzed data extracted from a public curricular database of researchers in Brazil (namely Lattes Platform), and focused on researchers who hold a master's or doctorate degrees. Three aspects of Brazilian science were examined, wich are as follows: (1) the degree of interdisciplinarity between areas of knowledge, (2) the evolving structural features regarding both areas of knowledge and researchers, and (3) the similarity between two different graphs, one representing researchers holding a doctorate and another which includes researchers with a master's or doctorate degrees.

The first factor addresses the degree of interdisciplinarity between main areas of knowledge in Brazil (Hypothesis 1). The argument was that, if there is a certain degree of interdisciplinarity from the perspective of co-authorship, as shown by Mena-Chalco et al. (2014), it is also possible that this feature occurs in mentoring relationships.

Our results revealed that researchers working in one of the leading areas of knowledge, had mentored students actively involved in other main areas of knowledge. This was inferred from what was shown in the graphs as a whole and also when confined to a period of mentoring (the last 30 years). From an individual standpoint, we found there was a high percentage of descendants from two foreign researchers involved in different subareas of knowledge, and this reveals that the interdisciplinarity had occurred earlier (i.e. at the beginning of the formal mentoring in Brazil). These results strengthen Hypothesis 1, that science in Brazil possesses a certain degree of interdisciplinary, from the standpoint of academic genealogy.

The structural features and their evolving patterns for both areas of knowledge and researchers individually, concern Hypothesis 2, in which we argue that these features are related to the time after an academic has obtained a degree. In addition, there is evidence of characteristics that appear to be hereditary, since some features were observed reflected in following generations, like the number of mentoring relationships and the possibility of winning a Nobel Prize, as observed in the studies of Malmgren et al. (2010) and Chariker et al. (2017), respectively.

On the one hand, by analyzing researchers when they were divided into different areas of knowledge, it could be seen how the genealogical metrics operate, with Biological sciences and Engineering being the most prominent areas in the metric descendants. On the other hand, it was also possible to analyze the most outstanding scientist that could be found in each of the five metrics. Additionally, by examining the evolving pattern of genealogical metrics, we saw that all the five metrics vary in accordance with the researchers' academic ages, which supports Hypothesis 2. It is worth taking note of the fecundity indicator, owing to its non-accumulative nature, which means it could be an important tool for evaluative purposes.

The last factor concerns the similarity between the structural features in light of the different graphs or levels of training. By comparing both the PhD and MSc & PhD graphs, it could be seen that Hypothesis 3 is supported in the context of interdisciplinarity, i.e., on this conceptual basis, the different academic graphs showed a similar pattern. Furthermore, the evolving of the genealogical metrics, grouped by graphs, displayed similar curves for

all the metrics, which also supports Hypothesis 3. The addition of a master's degree to the PhD graph, does not affect the behavior of the metrics or degree of the interdisciplinarity.

This approach opens up the field of bibliometric studies in Brazil, as it makes it possible to determine the influences of the pioneer researchers who carried out the first research projects. It suggests that the branch of relational bibliometrics can be enhanced by this kind of linkage. The same can be said about studies of interdisciplinarity in the training of researchers.

Finally, it can be concluded that the use of topological metrics applied to academic genealogy, can be a useful approach when adopted at a micro-level. It offers alternative and evolving information about a researcher's profile, which can be useful for evaluative purposes, for example, in funding agencies.

In future work, the study of artificial nodes, especially foreign ones, will be explored in greater depth to provide more detailed information about the foreign antecedents of Science in Brazil. Moreover, we will provide an analysis of the correlation between not only the advisors and their direct descendants but also between their grandchildren and cousins. Finally, we think it would be of great value to undertake studies of Brazilian scientific activity, by examining mentoring and publications jointly, in order to understand the scientific output in both dimensions.

Acknowledgements The authors would like to thank the Federal University of ABC for its financial support.

Appendix

We investigated the influence between main areas of knowledge in both PhD and MSc & PhD graphs. Figure 9 shows 16 radar charts (two by main area of knowledge). The axis of the radar charts uses the logarithmic function applied to the data shown in Table 4. As observed in Table 4, main areas of knowledge have a certain degree of interdisciplinarity. Moreover, while comparing the two graphs, there is a similar pattern in the influence exerted and experienced for both graphs PhD and MSc & PhD.

We also analyzed who are the fifteen artificial nodes with the highest values for each of the five metrics. Table 7 presents the values for the metrics descendants and fecundity, as well as the genealogical index for the 15 researchers without Lattes ID with the highest values, for both PhD and MSc & PhD graphs.

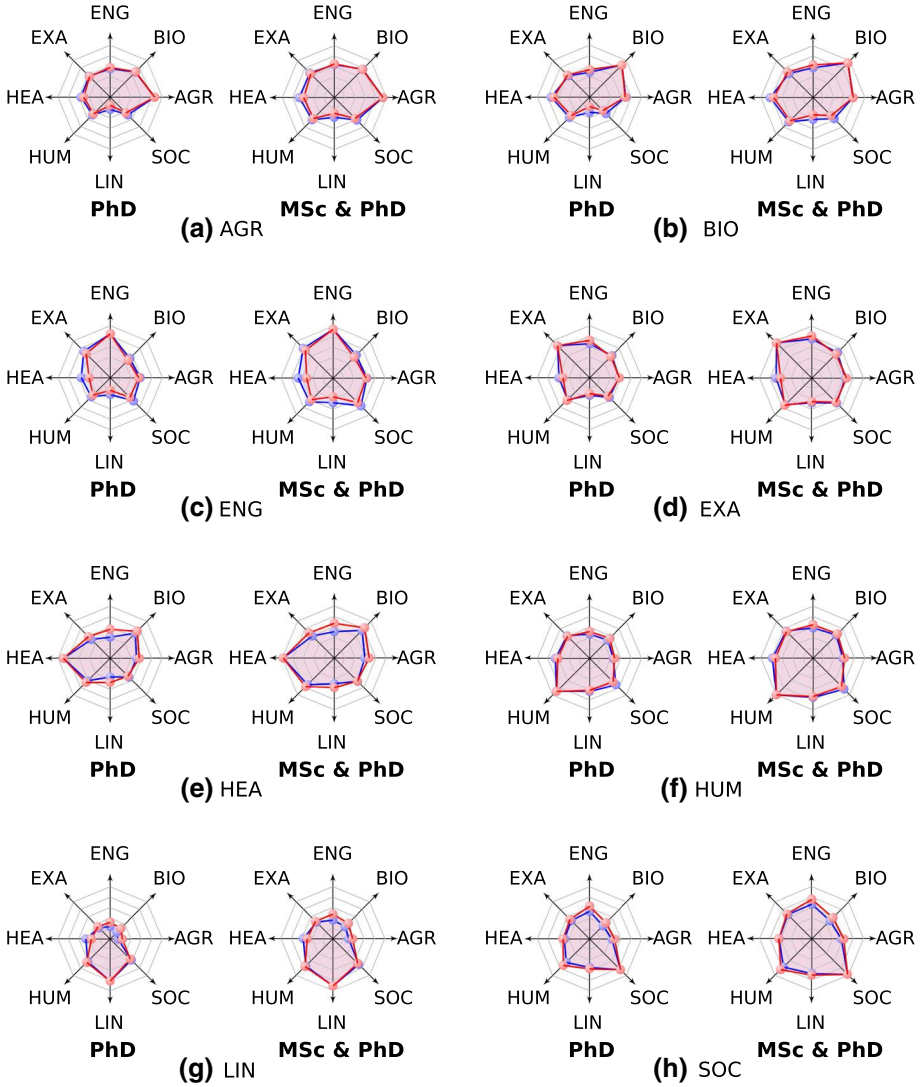

Fig. 9 Influence exerted (blue) and experienced (red) for each main area of knowledge. The radar charts are in logarithmic scale. (Color figure online)

Table 7 Top 15 values for metrics descendants (d^+ and d^-), fecundity (f^+ and f^-) and genealogical index (g) for graphs: (a) PhD, (b) MSc & PhD. Only artificial nodes

Researcher	d^+	Researcher	d^-	Researcher	f^+	Researcher	f^-	Researcher	gi
(a) PhD									
Martins, J	2022	Marchesi, C	17	Salazar, LB	53	Silva, AC	5	Bori, CM	12
Brieger, FG	1769	Artur, AG	16	Maffesoli, M	40	Borges, JC	5	Martins, J	12
Dreyfus, A	1524	Oliveira, JC	16	Martins, J	34	Oliveira, MA	5	Portella, E	9
Bori, CM	1320	Azevedo, LCP	16	Villalba, OA	28	Gomes, AD	4	Santanna, AR	8
Lima, JP	1165	Silva, JA	15	Bori, CM	27	Artur, AG	4	Souza, ACME	8
Pereira, L	1107	Golunski, S	15	Sawaya, P	23	Hirschmann, ACO	4	Franca, ED	8
Pottier, B	1100	Pinheiro, AA	14	Segovia, AMC	21	Hernández, CELR	4	Mauro, F	8
Michel, L	1051	Borges, JC	14	Sachs, I	21	Chaves, CR	4	Cilento, G	8
Franca, ED	1030	Santos, LO	14	Santos, BS	20	Giunchetti, DSL	4	Maffesoli, M	8
Roper, JA	1025	Hirschmann, ACO	13	Pastor, FA	20	Oliveira, JC	4	Vanzolini, PE	8
Gurgel, JTdA	1024	Rossi, EB	13	Schijman, JH	19	Rosmaninho, MG	4	Giulietti, AM	7
Moisés, M	1021	Brunetto, G	13	Rochefort, M	19	Silva, MIN	4	Matson, E	7
Izquierdo, JA	970	Oliveira, MA	13	Ianni, O	19	Pinto, MCX	4	Sachs, I	7
Souza, GRME	929	Barata, MTA	13	Giulietti, AM	17	Araujo, AS	3	Almeida, PAM	7
Covian, MR	926	Moreira, RA	13	Mauro, F	17	Ferreira, AM	3	Cardoso, RCL	7
(b) MSc & PhD									
Martins, J	18,679	Barbosa, DNL	77	Villauba, OA	104	Silva, AC	14	Martins, J	23
Pereira, L	8138	Ligabó, AJP	70	Salazar, LB	93	Oliveira, PR	12	Bori, CM	20
Bori, CM	7578	Tarichi, AP	70	Martins, J	61	Pereira, ACS	11	Santanna, AR	17
Franca, ED	7423	Pacheco, BCS	70	Canese, ME	61	Oliveira, CA	10	Ianni, O	17
Dreyfus, A	7305	Rodrigues, LBS	70	Ferrari, D	50	Oliveira, MA	9	Maffesoli, M	15
Fernandes, F	6851	Assis, MVB	70	Ianni, O	50	Silva, CR	8	Rochefort, M	14
Brieger, FG	6369	Mauro, RA	70	Maffesoli, M	49	Oliveira, MA	8	Penna, AG	13
Ianni, O	6288	Senna, TF	70	Sartori, V	42	Silva, MA	8	Franca, ED	13
Pottier, B	6206	Silva, WF	69	Segovia, AMC	40	Silva, JC	7	Mauro, F	13
Moisés, M	6005	Tiradentes, MS	68	Penna, AG	35	Silva, JM	7	Souza, ACME	12

Table 7 (continued)

Researcher	d^+	Researcher	d^-	Researcher	f^+	Researcher	f^-	Researcher	gi
Michel, L	5671	Cunha, CR	67	Gemael, C	35	Santos, AC	6	Rodrigues, A	12
Holanda, SB	5639	Carvalho, GR	67	Morino, CIR	35	Guimarães, AP	6	Sheiham, A	12
Stevens, WK	5439	Toscano, RN	67	Rochefort, M	34	Silva, AP	6	Castro, CLM	12
Stilman, M	5433	Menezes, WWD	67	Bori, CM	33	Ferreira, JC	6	Portella, E	12
Roper, JA	5274	Morgan, AS	66	Otero, PMCC	33	Silva, JC	6	Haag, HP	12

References

Bennett, A. F., & Lowe, C. (2005). The academic genealogy of George A. Bartholomew. *Integrative and Comparative Biology*, *45*(2), 231–233.

Center for Strategic Studies and Management Science, Technology and Innovation. (2016). *Doutores 2015: Estudos da demografia da base técnico-científica brasileira*. Brasília, DF: Centro de Gestão e Estudos Estratégicos.

Chariker, J. H., Zhang, Y., Pani, J. R., & Rouchka, E. C. (2017). Identification of successful mentoring communities using network-based analysis of mentor–mentee relationships across nobel laureates. *Scientometrics*, *111*(3), 1733–1749.

Damaceno, R. J. P., Rossi, L., & Mena-Chalco, J. P. (2017). Identificação do grafo de genealogia acadêmica de pesquisadores: Uma abordagem baseada na plataforma Lattes. In *Proceedings of the 32nd Brazilian symposium on databases* (pp. 76–87).

David, S. V., & Hayden, B. Y. (2012). Neurotree: A collaborative, graphical database of the academic genealogy of neuroscience. *PLoS ONE*, *7*(10), e46608.

Dores, W., Soares, E., Benevenuto, F., & Laender, A. H. F. (2017). Building the Brazilian academic genealogy tree. In J. Kamps, G. Tsakonas, Y. Manolopoulos, L. Iliadis, & I. Karydis (Eds.), *Research and advanced technology for digital libraries* (pp. 537–543). Berlin: Springer.

Elias, M., Floeter-Winter, L. M., & Mena-Chalco, J. P. (2016). The dynamics of Brazilian protozoology over the past century. *Memórias do Instituto Oswaldo Cruz*, *111*(1), 67–74.

Heinisch, D. P., & Buenstorf, G. (2018). The next generation (plus one): An analysis of doctoral students' academic fecundity based on a novel approach to advisor identification. *Scientometrics*, *177*, 351–380.

Jackson, A. (2007). A labor of love: The mathematics genealogy project. *Notices of the American Mathematical Society*, *54*(8), 1002–1003.

Kelley, E. A., & Sussman, R. W. (2007). An academic genealogy on the history of American field primatologists. *American Journal of Physical Anthropology*, *132*(3), 406–425.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, *10*, 707.

Li, Y., Fang, N., Liu, Z., & Yu, H. (2017). Inferring advisor–student relationships from publication networks based on approximate maxconfidence measure. *Mathematical Problems in Engineering*. https://doi.org/10.1155/2017/8135464.

Liu, J., Tang, T., Kong, X., Tolba, A., AL-Makhadmeh, Z., & Xia, F. (2018). Understanding the advisor–advisee relationship via scholarly data analysis. *Scientometrics*, *116*(1), 161–180.

Malmgren, R. D., Ottino, J. M., & Amaral, L. A. N. (2010). The role of mentorship in protégé performance. *Nature*, *465*(7298), 622–626.

Martin, S., Brown, W. M., Klavans, R., & Boyack, K. W. (2011). Openord: An open-source toolbox for large graph layout. In *SPIE proceedings—Visualization and data analysis 2011* (Vol. 7868, pp. 786–806). International Society for Optics and Photonics.

Mena-Chalco, J. P., Digiampietri, L. A., Lopes, F. M., & Cesar Jr., R. M. (2014). Brazilian bibliometric coauthorship networks. *Journal of the Association for Information Science and Technology*, *65*(7), 1424–1445.

Montoye, H. J., & Washburn, R. (1980). Research quarterly contributors: An academic genealogy. *Research Quarterly for Exercise and Sport*, *51*(1), 261–266.

Rossi, L., Damaceno, R. J. P., Freire, I. L., Bechara, E. J. H., & Mena-Chalco, J. P. (2018). Topological metrics in academic genealogy graphs. *Journal of Informetrics*, *12*(4), 1042–1058.

Rossi, L., Freire, I. L., & Mena-Chalco, J. P. (2017). Genealogical index: A metric to analyze advisor–advisee relationships. *Journal of Informetrics*, *11*(2), 564–582.

Rossi, L., & Mena-Chalco, J. P. (2014). Caracterização de árvores de genealogia acadêmica por meio de métricas em grafos. In *Brazilian workshop on social network analysis and mining* (pp. 1–12).

Sonnenwald, D. H. (2007). Scientific collaboration. *Annual Review of Information Science and Technology*, *41*(1), 643–681.

Sugimoto, C. R. (2014). Academic genealogy. In B. Cronin & C. R. Sugimoto (Eds.), *Beyond bibliometrics: Harnessing multidimensional indicators of scholarly impact* (first ed., pp. 365–382). Cambridge: MIT Press.

Sugimoto, C. R., Ni, C., Russell, T. G., & Bychowski, B. (2011). Academic genealogy as an indicator of interdisciplinarity: An examination of dissertation networks in library and information science. *Journal of the American Society for Information Science and Technology*, *62*(9), 1808–1828.

Tuesta, E. F., Delgado, K. V., Mugnaini, R., Digiampietri, L. A., Mena-Chalco, J. P., & Pérez-Alcázar, J. J. (2015). Analysis of an advisor–advisee relationship: An exploratory study of the area of exact and earth sciences in Brazil. *PLoS ONE*, *10*(5), e0129065.

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

Wang, C., Han, J., Jia, Y., Tang, J., Zhang, D., Yu, Y., & Guo, J. (2010). Mining advisor–advisee relationships from research publication networks. In *Proceedings of the 16th international conference on knowledge discovery and data mining* (pp. 203–212). ACM.

Wang, W., Liu, J., Xia, F., King, I., & Tong, H. (2017). Shifu: Deep learning based advisor–advisee relationship mining in scholarly big data. In *Proceedings of the 26th international conference on World Wide Web companion* (pp. 303–310). International World Wide Web Conferences Steering Committee.