Identification of the Brazilian academic roots through mining advisor-advisee relationships

R. J. P. Damaceno, L. Rossi, J. P. Mena-Chalco

Federal University of ABC, Brazil
{rafael.damaceno, luciano.rossi, jesus.mena}@ufabc.edu.br

Abstract. This study seek to carry out an identification and analysis of academic roots using academic genealogy graphs as data source. These graphs are used to identify the academic roots of 85 areas of knowledge and analyze the influences prevailing between them. The results show that science in Brazil is young, with most of the PhD and master's graduates having obtained an academic degree between the years 1980 and 2000. We detected some key areas of knowledge, such as Education and Medicine that exert a considerable influence on the mentoring of academics in several areas of knowledge. The significance of this study is that it employs a method to use mentoring relationships for the identification of the academic roots of areas of knowledge, that could be applied to any academic genealogical graph.

Categories and Subject Descriptors: H.2.8 [**Database Management**]: Database Applications; G.2.2 [**Discrete Mathematics**]: Graph Theory

Keywords: graph mining, academic roots, advisor-advisee relationships.

1. INTRODUCTION

Science involves classifying academic disciplines or areas of knowledge that are arranged in accordance with the similar features they possess. The way science evolves is partly due to the interactions between the different areas. These interactions entail the sharing of the scientific knowledge that is peculiar to each area. However, one area may give rise to other related areas because of the depth of expertise that it includes. An example of this latter phenomenon can be found between the areas of Electrical Engineering and Computer Science. These areas are mutually influential and there is a point of intersection between the specialist knowledge of each area.

This paper investigates the influences that exist between the areas that form contemporary science in Brazil. This influence is determined by the identification and quantification of the different areas that assist in the formation of a specific area. This assistance is obtained through the Academic Genealogy (AG), which can be defined as the study of the intellectual inheritance that is perpetuated through formal relationships of academic mentoring [Sugimoto 2014]. Thus, the advisor-advisee relationships and the areas of expertise of the academics form a hierarchical structure represented by an AG graph, in which the nodes and edges represent the areas of expertise of the academics and the graduate mentoring, respectively.

The graph mining was carried out by examining the largest repository of academic curricula in Latin America - the Lattes platform [Damaceno et al. 2017]. This data source represents the history of science in Brazil by recording the academic activities of more than seven hundred thousand academics with a Master's or Ph.D.'s degree. In addition to graph mining, we used label propagation techniques to complete the missing attributes, as well as to create an "origins identification algorithm" that is

The authors would like to thank the Federal University of ABC and CAPES for supporting this work.

tailor-made for this study. The result is a social network with a wealth of opportunities for exploration. The following factors area highlighted: temporal ordering, an influence map and the distribution of the areas of knowledge in two-dimensional space. These describe the similarities between the areas of knowledge on the basis of the influences experienced.

This study can be regarded as original because it includes mentoring relationships as an interconnecting feature between the areas of expertise of the different scholars. The significance of this study is that it establishes a framework resulting from academic genealogical data, where there is a panorama of Brazilian science that relates the areas of knowledge, their influential relationships and the respective emergence of a temporal order.

2. RELATED WORK

The increasing availability of genealogical data requires the development of models and methods that can be applied to represent knowledge in terms of complex networks and analysis of these structures [Arruda et al. 2017]. In addition to the academic subject, there have been measures taken to devise methods for analyzing the relationship between knowledge in industrial areas [Dezfoulian et al. 2017]. In this regard, networks can describe elements as "patents" to provide of the transfer of theoretical knowledge to practical technological applications [Ding et al. 2017]. Networks representing strategic alliances and their respective flow of knowledge are useful for investigating the evolving pattern of this type of structure, especially in the context of knowledge management [Jianyu et al. 2017].

Currently, a large number of the publications devoted to the study of the transfer of scientific knowledge, seek to make a correlation between the advance of science and the socioeconomic development provided by this advance. These studies investigate the flow of knowledge in both the internal and external environment, and among companies whose field of expertise is interrelated [Gao et al. 2015]. In contrast, [Sorenson et al. 2006] analyzed the possible advantages that actors closer to the sources of knowledge may have compared with those that are more distant. This study is based on patent data and uses a network of citations to study the impact of knowledge in various fields.

Studies of the flow of scientific knowledge in the academic world usually converge to an analysis of publications, citations, and collaborative research for structuring the knowledge network. [Mohammadi and Thelwall 2014], by counting readers with the aid of the *Mendeley* platform, compared this with the analysis of citations. In addition, the work of [Boschma et al. 2014], sought to trace the main cognitive trends in this specific area through an analysis of expressions in the titles of scientific publications in the area biotechnology.

An exploratory study, described by [Rinia et al. 2002], (the objective of which is to measure the transfer of knowledge between the disciplines and subfields of science), makes interesting observations about how advances in a given area of knowledge can affect other areas. They discussed the metrics that can be used to quantify the rate of knowledge transfer between different areas of knowledge. The use of metrics to quantify this impact is also mentioned in the study by [Rafols and Meyer 2010].

3. METHODS

The graph employed in this study is the result of a mining process developed by [Damaceno et al. 2017] in which each vertex represents a researcher and each edge a relationship between two researchers. The graph is directed from the advisor (source) to the advisee (destiny). The curricula data provided by the Lattes platform were gathered and structured in accordance with the academic mentoring relationships stated in each curriculum vitae. The graph mining process involves a preprocessing of the data to improve the accuracy of the information by resolving factors such as (i) the lack of standards in the registers, (ii) incomplete data, and (iii) errors in the identification of the advisor/advisee. When studying the academic roots of areas of knowledge, some alterations were made to the graphs; these followed two phases (i) the propagation of labels and (ii) the identification of academic roots.

3.1 Propagation of labels

The graph used in this article does not have all its vertices labeled with an area of knowledge because the data source for generating this information is incomplete. The reason for this lack of information is that researchers have failed to refer to their area of knowledge in their curricula. An algorithm has been created to supply this information, by propagating the area of knowledge to the vertices that do not have it labeled. However, the propagation of labels may introduce some degree of noise in the graphs.

Initially, the vertices with a degree equal to 1 (i.e., in-degree 1 and out-degree 0 and vice-versa) were labeled with their parent/son's area of knowledge. The remaining vertices were labeled with the area of knowledge of their neighborhood, i.e., the area of a given vertex was the mode of knowledge of all its advisors and advisees. If there were two or more areas of knowledge with the same mode, account was taken of the mode of the ancestors or the mode of the descendants (whichever was greater). If the mode was still the same, we used the mode of the ancestors to label the vertex. The same approach was adopted to the main areas of knowledge, where similar areas are formed into bigger groups.

3.2 Identification of academic roots

After applying the propagation of labels, we carried out the "identification of academic roots" stage, which consists of finding out who are the ancestors of each vertex (grouped by area of knowledge). In the case of each vertex (source) of a given area, the algorithm makes it possible to rise in the graph until there are no more ancestors with the same area of knowledge as the source vertex. The last ancestor identified, which has a different area of knowledge from the source vertex, is the academic root of that vertex which possesses that area of knowledge. The same approach was applied to the main areas of knowledge.

For example, consider a vertex whose area of knowledge is Computer Science. If the area of knowledge of the ancestors of that vertex is also Computer Science, we continue to obtain more ancestors (now the ancestors of the ancestors). We proceed in this way until the area of knowledge of all the ancestors is not equal to the area of knowledge of the source vertex. This approach enabled us to estimate the roots (academics) of a given area of knowledge and determine the influence between them over a period of time. In addition, it allows us to estimate the age of the areas of knowledge and of the main areas of knowledge. Section 3.3 shows the algorithm we created and applied to discover the academic roots of different areas of knowledge.

3.3 Algorithm to identify the roots

Consider the graph-structured data from the Lattes platform as $\vec{G}(V, E)$, the set of areas of knowledge as $Area$ and a square matrix M with order proportional to the number of areas, initially populated with zeros. $\vec{G}(V, E)$ is formed of a set of vertices $\vec{G}(V)$ representing the academics and a set of edges $\vec{G}(E)$ representing the mentoring relationships between the academics. In the case of each academic $v \in \vec{G}(V)$, there is a label giving information about its area of knowledge, represented by $v.area$. The algorithm for the identification of academic roots of areas of knowledge, is outlined below.

```
AREAS-SELECTION(G⃗, Area, M)
1  for each area ∈ Area
2    for each v ∈ G⃗
3      if v.area = area
4        ROOT-IDENTIFICATION(G⃗, v, M)
```

```
ROOT-IDENTIFICATION(G⃗, v, M)
1  if G⃗.ascendancy[v] = ∅
2    return
3  else
4    for each u ∈ G⃗.ascendancy[v]
5      if u.area ≠ v.area
6        M[v.area][u.area] ← M[v.area][u.area] + 1
7      else
8        ROOT-IDENTIFICATION(G⃗, u, M)
```

The "AREAS-SELECTION" receives the AG graph and the set of areas of knowledge and selects the vertices that correspond to each of the areas, which are subjected to the "ROOT-IDENTIFICATION" procedure. In this last process, the linked areas between the vertex in question and its ancestors are compared and, if they do not match, the influence matrix M is incremented by one unit in the row and column that corresponds to the areas of the vertex in question and of its rise, respectively. If it is found that the areas match, the ascending vertex is taken as the parameter in this same process, recursively.

4. RESULTS

From the total number of vertices (n = 1 111 544), 472 637 (42.52%) do not have a defined main area of knowledge and 477 013 (42.91%) do not have a defined area of knowledge. After applying the propagation of labels, 20 845 (1.88%) vertices remain with an undefined main area of knowledge, and 28 219 (2.54%) with an undefined area of knowledge. Table I displays the number and percentage of vertices representing graduate (master's and doctorates) academics grouped by main area of knowledge after the propagation bas been applied.

Table I. Number and percentage of academics by academic degree and main area of knowledge.

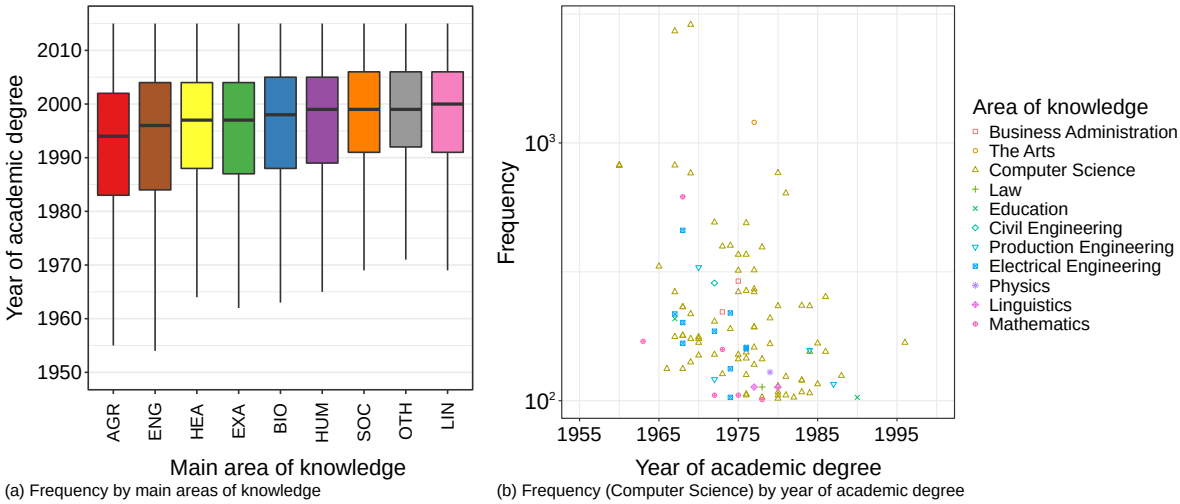
		Doctorates		Master's		All	
Main area of knowledge	Acronym	N	%	N	%	N	%
Humanities	HUM	59 938	16.26	143 202	19.28	203 140	18.28
Applied Social Sciences	SOC	38 445	10.43	135 526	18.24	173 971	15.65
Health Sciences	HEA	58 146	15.77	108 619	14.62	166 765	15.00
Exact and Earth Sciences	EXA	57 535	15.61	87 118	11.73	144 653	13.01
Engineering	ENG	36 198	9.82	83 248	11.21	119 446	10.75
Biological Sciences	BIO	50 239	13.63	58 985	7.94	109 224	9.83
Agricultural Sciences	AGR	34 933	9.48	53 522	7.20	88 455	7.96
Linguistics, Letters and Arts	LIN	21 788	5.91	49 588	6.67	71 376	6.42
Undefined	UND	7 637	2.07	13 208	1.78	20 845	1.88
Others	OTH	3 778	1.02	9 891	1.33	13 669	1.23
All		368 637	100.00	742 907	100.00	1 111 544	100.00

For a better understanding of the way the main areas of knowledge are ordered so that they can reveal a) the academic roots and b) the year of academic degree was awarded, Figure 1(a) shows a distribution of the roots grouped by main area of knowledge. Here it should be pointed out that there is a direct relationship between the academic age of researchers and the age of CAPES (established on July 11, 1951).

A more in-depth analysis was conducted to illustrate the influence experienced by specific areas of knowledge, and Figure 1(b) shows which areas of knowledge exert an influence on Computer Science. In (b) the roots are marked with different types of points and colors and the roots that were pointed the fewest times were omitted (i.e. lower than 100 times).

Agricultural Sciences is the main area of knowledge that has the lowest median for years of academic degree. Linguistics, Letters and Arts have the highest median for years of academic degree. The median number can be found between the years 1994 and 2000, which shows that the science conducted in Brazil is still young. The areas of knowledge that exert the greatest influence on Computer Science are Electrical Engineering, Mathematics and Education. These areas of knowledge have higher frequencies, with most of the roots occurring in the early years, i.e., between 1960 and 1970. The root pointed at the year 1960, and those pointed most often are from Computer Science.

We also estimated the age of the areas of knowledge by using the root year of academic degree as a measure. Figure 2 shows the frequency of the roots year of academic degree in terms of areas of knowledge. Nuclear Engineering is the area of knowledge that has the lowest median of years of academic degree. Robotics, Mechatronics and Automation have the highest median for years of



(a) Frequency by main areas of knowledge

(b) Frequency (Computer Science) by year of academic degree

Fig. 1. Frequency of roots by year grouped into areas of knowledge.

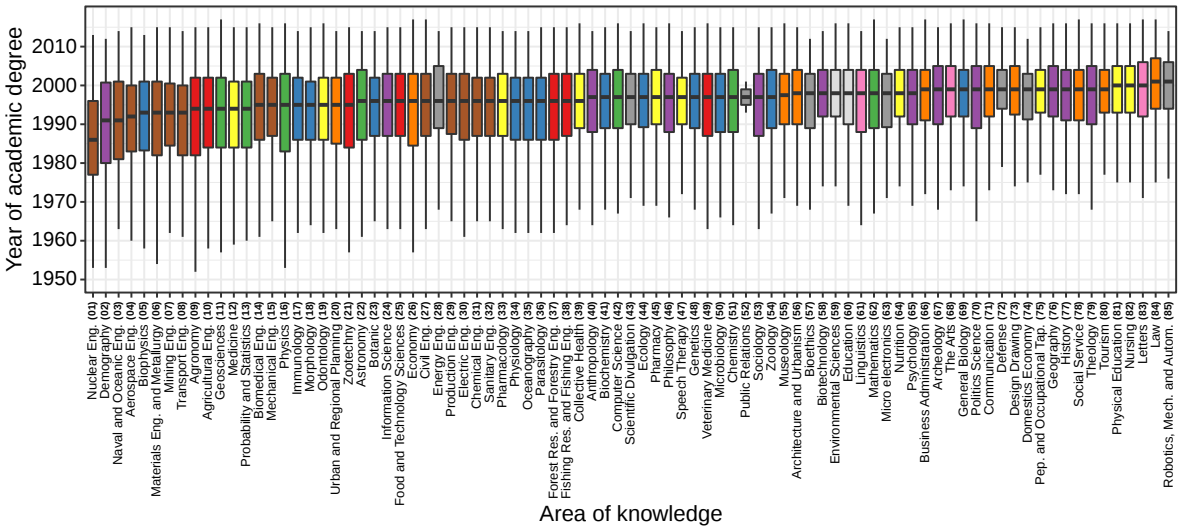


Fig. 2. Frequency of roots by year grouped by areas of knowledge. The bar color indicates the main area of knowledge for each subject.

academic degree. The median is found between the years 1986 and 1996. Most of the brown bars are on the left of the graph, and this corroborates the information displayed in Figure 1(a), that shows Engineering has one of the oldest roots. With regard to Heath Sciences, Medicine had a lower median for the years of academic degree than the other areas of the Health Sciences.

The vertices and edges of the graph that result from the process of root identification have the areas of knowledge and the influence exerted between them, respectively. Additionally, the edges are weighted with the number of roots identified. Figure 3 shows a partial representation of the graph, where only 85 areas are included, with the highest number of roots and only the maximum weight edge that is found in each area. The purpose of restricting the representation is to make it easier to visualize and interpret the structure that represents a map of influence between the areas of knowledge. As an example of interpretation, consider two areas (A and B) that are connected by a weight edge w that emerges from A and focuses on B. This means that area B has w roots that belong to area

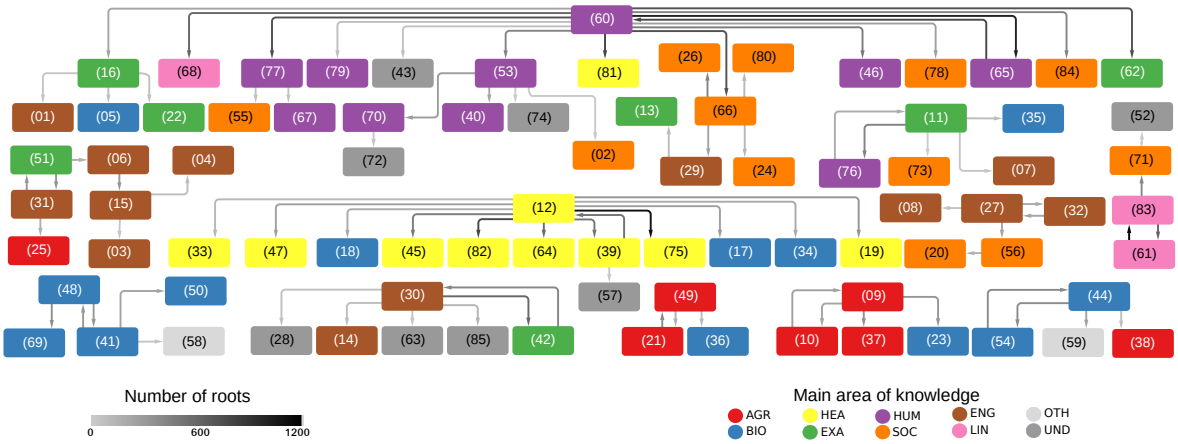


Fig. 3. Map of influence between areas of knowledge. The directed edges connect the areas that exerts an influence (origin) on the vertex that experiences influence (destination). The grayscale tones show the number of roots and the color of the bars (areas) refers to the main area of knowledge to which the subject belongs.

A, and w is the most significant weight observed among all the edges that focus on B. The areas of knowledge in the map of influences, represented in Figure 3, are colored in a way that corresponds to a clustering in the main area of knowledge, which is a formal classification used in Brazil. The edges are represented in grayscale and define the number of roots found in the area that experiences influence (i.e. the vertex which intersects the edge) and that belong to the area that exerts an influence (i.e. the vertex from which the edge emerges).

The map of influence exerted/experienced between different areas, displays how groups are formed since they share the same influential area. The largest group of influence is formed from the academics who act in the area of Education. This area exerts a direct influence on thirteen other areas, such as Business Administration, Sociology, and Physics. There are ten other groups exerting influence on the map that are formed from the following influential areas (and number of influenced areas): Medicine (11), Electrical Engineering (5), Geosciences (4), Agronomy (3), Biochemistry (3), Civil Engineering (3), Ecology (3), Chemistry (2), Letters (2), and Veterinary Medicine (2). Another exciting feature revealed by the map is the presence of mutual influences in all areas where there is a more significant influence.

The term "mutual influence" is defined as meaning when one area is the most influential on another and this, in turn, is the most influential on the first, as is the case, for example, between the areas of Electrical Engineering and Computer Science. The mutual influence between areas may be indicative of the emergence of one area caused by the higher degree of specialization of another. Other areas with mutual influence on the map are (Education $\leftrightarrow$ Psychology), (Geosciences $\leftrightarrow$ Geography), (Letters $\leftrightarrow$ Linguistics), (Ecology $\leftrightarrow$ Zoology), (Civil Engineering $\leftrightarrow$ Sanitary Engineering), (Agronomy $\leftrightarrow$ Agricultural Engineering), (Medicine $\leftrightarrow$ Collective Health), (Veterinary Medicine $\leftrightarrow$ Zootechnics), (Biochemistry $\leftrightarrow$ Genetics), and (Chemistry $\leftrightarrow$ Chemical Engineering).

The influences exerted and experienced can be analyzed by treating different levels as a chain of influential areas. It can be seen in these chains, how different types of knowledge, that are characteristic of a given area, are combined to form a different area. This unfolding of knowledge can be observed, for instance, in the chain that has Education as an influential area on Sociology, which in turn, led to the formation of the Political Sciences group, and the latter influenced the area of Civil Defense. This pattern of unfolding scientific chains is more evident when we examine societies that have a stronger academic tradition, which is not the case in Brazil.

When account is taken of only the largest area of influence in the composition of the map (Figure 3),

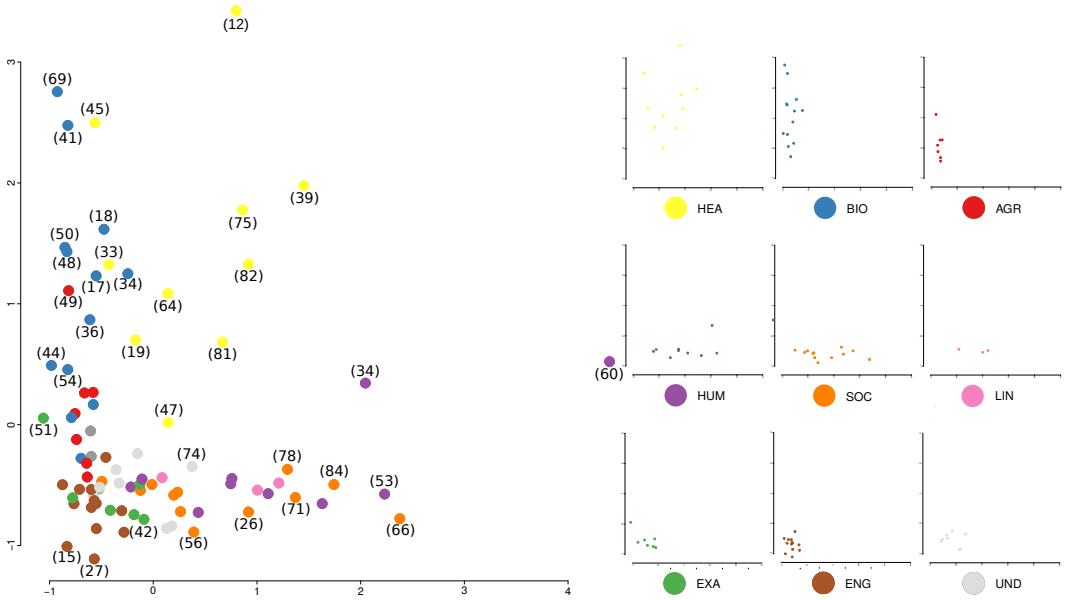


Fig. 4. The diagram, on the left, describes the distribution of the areas of knowledge in the two factors (axes) that best explain the variation; the colors are clustered according to the main area of knowledge to which they belong. On the right are the diagrams that corresponds to each grouping.

the structuring of knowledge makes evident the relationships between the areas and their natural hierarchical grouping. However, most of the information on influences was disregarded for the sake of interpretation. One way of examining all the information on influences in a representation that is able to illustrate the features of the areas according to their origins, is by conducting a Factor Analysis.

A Factor Analysis is a statistical method used to reduce the multidimensionality of data into a few representative factors. In the context of this study, the data is represented by a square matrix of order equal to 85 that represents the number of areas (see Section 3.3). Each row refers to an area, and the columns (variables) describe the number of roots identified by area. The fundamental concept underlying the Factor Analysis is that the multiple observable variables describe a pattern of similarity that is linked to a latent variable that has not been clearly measured. Each resulting factor is shown in order, according to the variation it can explain [Chatfield 2018]. Figure 4 shows the distribution of the areas as a function of two factors where the most significant variations are concentrated; however, describing the factors obtained according to the context of the original variables is not a trivial task.

The problem of assigning a semantic intuition for the factors is difficult because it is necessary to include all the original variables, which takes us back to the multidimensional context that we seek to reduce. However, it is possible to analyze the resulting distribution of the Factor Analysis and to trace interesting patterns in this configuration.

Figure 4 shows the distribution of the areas of knowledge; the colors represent groups that describe the main areas of knowledge to which the areas belong. The diagrams positioned to the right in Figure 4 represent the unique areas of each representative group. The groups labeled Health Sciences (HEA), Biological Sciences (BIO) and Agricultural Sciences (AGR) have their areas distributed according to the factor represented by the vertical axis. On the other hand, the factor represented by the horizontal axis influences the Human Sciences (HUM), Applied Social Sciences (SOC), and Linguistics, Letters and Arts (LIN). Exact and Earth Sciences (EXA), Engineering (ENG) and the undefined main area (UND), have their areas distributed according to two factors simultaneously. Thus, although there is no clear context for each factor, it can be stated that the areas show a pattern of formation reflecting their origins (roots), as well as the formal classification of the main areas of knowledge.

5. CONCLUSION

In this study, we developed and employed a method to identify the academic roots by using an academic genealogy graph as a data source. The "academic roots" were based on pioneering scientists that influence their successors through the formation of human resources, i.e., by establishing mentoring relationships. The influences between different areas of knowledge were measured by the identification and quantification of the academic areas of ancestors. The same approach was adopted for main areas of knowledge. Data from Brazil's academic genealogy were drawn on, as a case study, to illustrate the proposed method.

The results show that science in Brazil is still "young", with most of the academics having obtained a graduate degree between 1980 and 2000. With regard to the influence that one area exerts on other areas, we drew attention to some key areas of knowledge. Education and Medicine are the most important subjects since they exert an influence on several areas of knowledge. Education exerts an influence on 13 areas of knowledge in six different main areas of knowledge. Medicine exerts an influence on 11 areas of knowledge in Health Sciences and Biological Sciences. The proposed method was employed to study the academic roots of Computer Science to exemplify how an in-depth analysis could be conducted of specific areas of knowledge. We determined that Computer Science in Brazil had been greatly influenced by Electrical Engineering, Mathematics, and Education. By using the factor analysis technique, we were able to determine that there is a pattern of influence exerted/experienced between areas of knowledge which have similar vocational ends.

Future studies include the following: (i) the identification of scientific predecessors that exerted an influence on science conducted in Brazil, (ii) the analysis of the data generated in this work from an individual perspective, and (iii) the use of different databases to carry out this type of analysis.

REFERENCES

Arruda, H. F., Silva, F. N., Costa, L. d. F., and Amancio, D. R. Knowledge acquisition: A complex networks approach. *Information Sciences* vol. 421, pp. 154–166, 2017.

Boschma, R., Heimeriks, G., and Balland, P.-A. Scientific knowledge dynamics and relatedness in biotech cities. *Research Policy* 43 (1): 107–114, 2014.

Chatfield, C. *Introduction to multivariate analysis*. Routledge, 2018.

Damaceno, R. J. P., Rossi, L., and Mena-Chalco, J. P. Identificação do grafo de genealogia acadêmica de pesquisadores: Uma abordagem baseada na Plataforma Lattes. In *Proceedings of the 32th Brazilian Symposium on Databases*. SBC, 2017.

Dezfoulian, H., Afrazeh, A., and Karimi, B. A new model to optimize the knowledge exchange in industrial cluster: A case study of semnan plaster production industrial cluster. *Scientia Iranica. Transaction E, Industrial Engineering* 24 (2): 834, 2017.

Ding, C. G., Hung, W.-C., Lee, M.-C., and Wang, H.-J. Exploring paper characteristics that facilitate the knowledge flow from science to technology. *Journal of Informetrics* 11 (1): 244–256, 2017.

Gao, X., Chen, Y., Song, W., Peng, X., and Song, X. Regional university-industry knowledge flow: A study of chinese academic patent licensing data. *Open Journal of Social Sciences* 3 (02): 59, 2015.

Jianyu, Z., Baizhou, L., Xi, X., Guangdong, W., and Tienan, W. Research on the characteristics of evolution in knowledge flow networks of strategic alliance under different resource allocation. *Expert Systems with Applications*, 2017.

Mohammadi, E. and Thelwall, M. Mendeley readership altmetrics for the social sciences and humanities: Research evaluation and knowledge flows. *Journal of the Association for Information Science and Technology* 65 (8): 1627–1638, 2014.

Rafols, I. and Meyer, M. Diversity and network coherence as indicators of interdisciplinarity: case studies in bionanoscience. *Scientometrics* 82 (2): 263–287, 2010.

Rinia, E., van Leeuwen, T., Bruins, E., van Vuren, H., and van Raan, A. Measuring knowledge transfer between fields of science. *Scientometrics* 54 (3): 347–362, 2002.

Sorenson, O., Rivkin, J. W., and Fleming, L. Complexity, networks and knowledge flow. *Research policy* 35 (7): 994–1017, 2006.

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