

# Building the Brazilian Academic Genealogy Tree

Wellington Dores

Elias Soares

Fabrício Benevenuto

Alberto H. F. Laender

Department of Computer Science  
 Universidade Federal de Minas Gerais  
 Belo Horizonte, MG, Brazil  
 {wellingtond, eliassoares, fabricio, laender}@dcc.ufmg.br

## ABSTRACT

Along the history, many researchers provided remarkable contributions to science, not only advancing knowledge but also in terms of mentoring new scientists. Currently, identifying and studying the formation of researchers over the years is a challenging task as current repositories of theses and dissertations are cataloged in a decentralized way through many local digital libraries. Following our previous work in which we created and analyzed a large collection of genealogy trees extracted from NDLTd, in this paper we focus our attention on building such trees for the Brazilian research community. For this, we use data from the Lattes Platform, an internationally renowned initiative from CNPq, the Brazilian National Council for Scientific and Technological Development, for managing information about individual researchers and research groups in Brazil.

## CCS Concepts

•Information systems → *Digital libraries and archives*;

## Keywords

Academic genealogy trees; Academic mentorship; Lattes Platform.

## 1. INTRODUCTION

Science has evolved over the centuries as a system that not only promotes progress through the scientific method, but that is also centered on the processes of mentoring and teaching. The academic mentoring activity is a form of relationship that promotes the scientific development, as well as the formation and evolution of new researchers. Despite the complex system behind science, most of the existing efforts in the literature that aim at measuring individuals' research productivity within a scientific community usually account only for the publications produced [11], citations received [2] and collaborations established [1, 12], neglecting the formation of new researchers.

There has been only a limited number of initiatives, by specific academic communities, in the sense of documenting, analyzing and classifying advisor-advisee relationships. Sometimes this kind of study considers a representation usually called academic genealogy tree [3, 4, 9], in which nodes represent researchers and relations indicate that a researcher was the advisor of another one. However, these efforts have focused on specific fields, such as Mathematics [9] and Neuroscience [4], or have been restricted to a specific community

as in the cases of a career retrospect of prominent American physicists [3] and the tropical meteorology's academic community [8]. Although limited to specific locations and research areas, overall these efforts show that the analysis of such relationships in the form of a genealogy structure contributes to a greater understanding of a scientific community and of its individual values, allowing us to identify the impact generated by individuals in the formation of a community. For instance, Tuesta *et al.* [14] have analyzed the advisor-advisee relationship in the Brazilian exact and earth science field, correlating time and productivity throughout the advising relationship. Malmgreen *et al.* [13] have investigated mentorship fecundity using data from the Mathematics Genealogy Project.

Complementary to all these efforts, we have started an ambitious project towards building a large network that records the academic genealogy of researchers across fields and countries [5]. Our preliminary work used data from NDLTd, the Networked Digital Library of Theses and Dissertations<sup>1</sup> [7], and aimed to reconstruct advisor-advisee relationships from ETD records from many institutions around the world and from distinct disciplines.

In this paper, we move one step forward by constructing academic genealogy trees from a completely different data source, the Lattes Platform<sup>2</sup>. Maintained by CNPq, the Brazilian National Council for Scientific and Technological Development, this platform is an internationally renowned initiative [10] that provides a repository of researchers' curricula vitae and research groups, all integrated into a single system. All researchers in Brazil, from all levels (from junior to senior), are required to keep their curricula updated in this platform, which provides a great amount of information about the researchers' activities and their scientific production that can be used for many purposes. We then crawled the entire Lattes Platform and collected the curricula of all researchers holding a PhD degree. Next, we developed a basic framework to extract specific data from the collected curricula, identify and disambiguate the respective researchers, and establish their advisor-advisee relationships, from which we carried out a series of analyses that describe the main properties of the genealogy trees we were able to construct. Finally, we developed a first version of a system that allows users to browse and explore the academic genealogy trees. We believe that this is the first large-scale effort to generate a general academic genealogy tree involving as much distinct research fields as possible. We hope

<sup>1</sup><http://www.ndltd.org>

<sup>2</sup><http://lattes.cnpq.br>

our framework can evolve into a much larger crowdsourcing system that stores a comprehensive collection of academic genealogy trees.

The rest of the paper is organized as follows. Next, we describe how we built our academic genealogy trees from the Lattes Platform. Then, we present a preliminary characterization of the academic genealogy trees we were able to build and discuss our findings. Finally, we conclude the paper and provide directions for future work.

## 2. BUILDING THE GENEALOGY TREES

In this section, we discuss how we built the researchers' individual academic genealogy trees (AGT's, for short) using data from the Lattes Platform. To build such AGT's, we first crawled the Lattes Platform and collected the curricula vitae (in XML format) of 222,674 researchers holding a PhD degree. Then, following the procedure described by Algorithm 1, we parsed each collected curricula extracting the data required to build the researchers' AGT's. Such data appears basically in two specific sections of each curriculum: the **Identification** section, which includes the researcher's name, institution and degrees held, and the **Mentorships** section, which includes the researcher's list of all Master's and PhD students she has advised in her career. Note that the output of this procedure is actually a directed acyclic graph, since in her academic life a researcher might have had more than one advisor (e.g., PhD and Master's) or acted as a co-advisor for one or more students.

Following Algorithm 1, in order to build the AGT's, we first sort the set of all collected curricula according to the researcher's PhD degree year (line 1). This aims to establish a chronological order to build the individual AGT's, thus avoiding unnecessary name matchings when processing the advisees' curricula. Then we set the graph  $G$  empty (line 2). Next, for each curriculum in the set  $C$  (lines 3 to 26), we execute the following three main steps: (i) search  $G$  for the respective researcher's node, creating a new node if it does not yet exist or updating it otherwise (lines 4 to 9); (ii) search  $G$  for the nodes of the researcher's PhD and Masters advisors, creating them if they do not yet exist or updating them otherwise, and then connect them to the researcher's node (lines 10 to 16); (iii) for each researcher's advisee, search  $G$  for her respective node, creating it if it does not yet exist or updating it otherwise, and then connect it to the researcher's node (lines 17 to 25).

A critical component of our algorithm is the search function present in lines 4, 10 and 17. Although the Lattes platform provides an internal identifier for each researcher with a registered curriculum, it is not always possible to use this mechanism to instantaneously identify another researcher whose name appears, for instance, in the list of mentorships of a specific curriculum. Thus, to overcome this problem, we have implemented a simple, but quite effective strategy to handle this typical name disambiguation problem [6], which considers the following parameters: the researchers' names, the names of their institutions, the titles of their theses or dissertations, and the respective years of defense. A detailed discussion of this name disambiguation strategy is out of the scope of this paper. However, it is worth noticing that, when connecting a researcher's node to the nodes of her advisors (lines 10 to 16), in most cases we use only her advisor's name and the name of the institution where she earned a degree to match the respective nodes.

---

### Algorithm 1: The AGT Building Procedure

---

**Input:** A set  $C$  of Lattes Curricula;  
**Output:** A graph  $G$  with all AGT's built;

```

1 Sort  $C$  by the researchers' PhD degree year;
2 Set  $G$  empty;
3 foreach Curriculum  $c$  in  $C$  do
4   Search  $G$  for the researcher's node  $n$ ;
5   if there is no such a node in  $G$  then
6     Create node  $n$ ;
7   else
8     Update the academic attributes of  $n$ ;
9   end
10  Search  $G$  for the nodes  $p$  and  $m$  of the researcher's
    PhD and Master's advisors;
11  if either  $p$  or  $m$  are not found then
12    Create them;
13  else
14    Update the academic attributes of  $p$  and  $m$ ;
15  end
16  Connect  $p$  and  $m$  to  $n$ ;
17  foreach advisee in  $c$  do
18    Search  $G$  for the advisee's node  $a$ ;
19    if there is no such a node in  $G$  then
20      Create node  $a$ ;
21    else
22      Update the academic attributes of  $a$ ;
23    end
24    Connect  $a$  to  $n$ ;
25  end
26 end

```

---

## 3. CHARACTERIZING THE AGT'S

In this section, we briefly characterize some aspects of the AGT's we have been able to build. Our main motivation is to identify aspects that highlight the legacy of a researcher, measured in terms of formation of other researchers, and not in terms of the traditional counts of publications, impact factor, and scientific discoveries.

**Table 1: Graph Characterization**

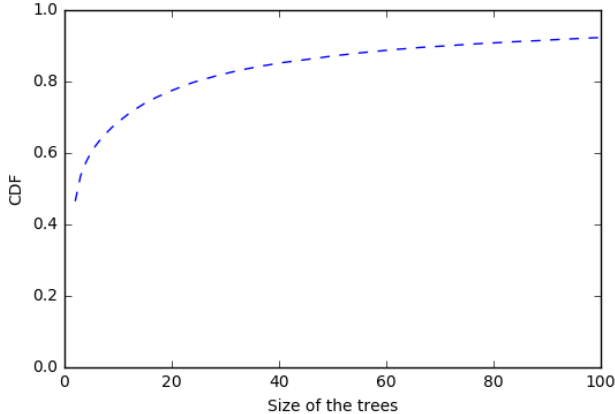
# of Nodes	903,183
# of Edges	1,144,051
# of Trees	70,610
# of Components	22,061
Avg. Tree Size	40.19
Avg. Tree Width	3.81

Table 1 shows some figures about the AGT's. Besides basic figures such as number of nodes, edges and trees, the later defined by the number of "roots" found in the graph (i.e., nodes without a known advisor), the table also shows the number of components (i.e., connected trees) and the values of two important metrics: the average tree size and the average tree width. The values of these two last metrics are calculated by dividing, respectively, the number of descendants by the number of subtrees (average size) and the number of out-links of all nodes by the number of nodes (width).

We have found in total 70,610 AGT's with 40.19 nodes on average. The average width of such trees is 3.81, i.e.,

each advisor in our dataset advised on average 3.81 PhD or Master’s students. Despite the average size of the trees being 40.19, the 10 largest trees have more than 5,000 nodes, although 80% of them have less than 20 nodes, as shown by the graph in Figure 1. On the other hand, almost half of the trees have depth 1, as shown in Figure 2. If we consider the width and the depth of a tree as its largest width and depth, respectively, we noted that trees are about 6.77 times wider than deeper in the Brazilian AGT’s. This number is much higher in comparison with the same ratio for trees built from NDLTD data [5], which is 2.48. We conjecture that this difference might be related to the quality of the trees we have obtained from both sources. NDLTD contains theses and dissertations from many institutions and countries, but it is unclear which scientific community it represents. On the other hand, Lattes represents an entire and complete scientific community, as basically all Brazilian researchers are forced to regularly update their academic records on the platform. We hope to incorporate many different data sources in our system and also allow users to fix and add their specific data, thus allowing one to better understand the idiosyncrasies from particular countries, research areas, or scientific groups, and their impact on scientific formation.

**Figure 1: Cumulative Distribution Function of the Tree Sizes**

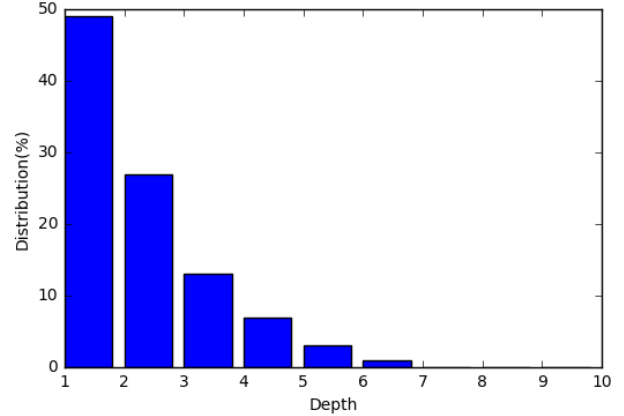


We now comment on Table 2 that lists the six most important foreign countries where Brazilian researchers obtained a Master’s or PhD degree. These six countries accounts for over 90% of the Brazilian researchers who chose to study abroad. We note that Portugal appears in first place, which might be explained by the same language spoken in both countries. These results highlight how rich the data from Lattes is and the kind of findings we can exploit by deepening our analysis of the AGT’s built from them.

#### 4. CONCLUSIONS AND FUTURE WORK

In this work, we used data crawled from the Lattes Platform to construct academic genealogy trees. Although still preliminary, our effort identified a number of interesting findings related to the structure of academic formation in Brazil, which highlight the importance of cataloging academic genealogy trees. Our effort, together with our previous work using data from the NDLTD [5], allowed us to identify many challenges that we need to tackle towards develop-

**Figure 2: Tree Depth Distribution**



**Table 2: The six most popular foreign countries from where Brazilian researchers earned a PhD or Master’s degree**

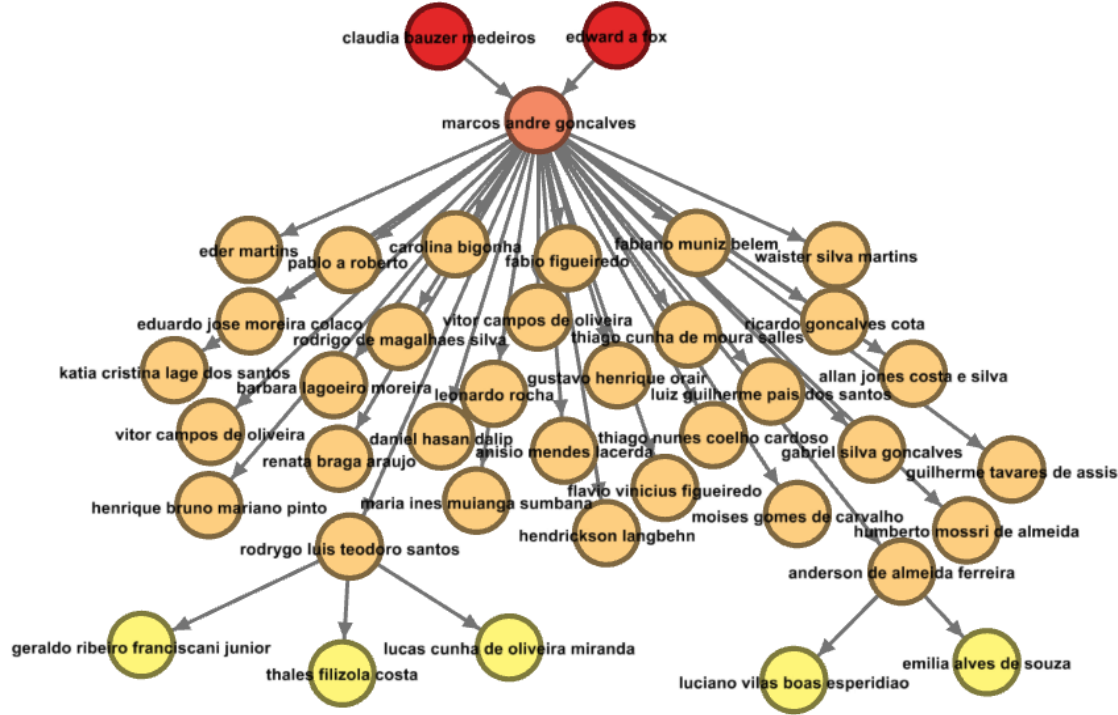
Country	PhD	Master’s
Portugal	1,179	300
USA	891	254
UK	853	219
Spain	802	162
France	660	248
Argentina	584	41

ing a large repository that records the academic genealogy of researchers across fields and countries. More importantly, we have developed a first version of a system that deploys the dataset studied here and allows users to browse the academic genealogy trees<sup>3</sup>. To briefly illustrate the potential of this system, Figure 3 shows an excerpt of the genealogy tree of Dr. Marcos André Gonçalves, a Brazilian associate professor from the Universidade Federal de Minas Gerais (UFMG), who is a well known member of the digital library community.

The colors in the figure represent the levels in the AGT. The red nodes correspond to Dr. Gonçalves’ advisors during his Master’s (Prof. Claudia Bauzer Medeiros, from UNICAMP, Brazil) and PhD (Prof. Edward A. Fox, from Virginia Tech, USA) studies. The main subtree (the one rooted by an orange node) includes the graduate (Master’s and PhD) students that have been advised by Dr. Gonçalves, which, in turn, span an additional level of subtrees (the yellow ones). Thus, by analyzing such a kind of tree we hope to be able to better understand a research lineage. More important, we believe this system represents a preliminary step towards the understanding of more important questions related to science, which we will be able to easily answer once we have a world-wide academic genealogy tree. For example, this system would allow us to identify the important researchers within areas and the role they have played on the creation and evolution of scientific communities, and even of novel fields. It would also provide a better understanding about where research areas came from, the birth and death of research communities, the identification of one’s

<sup>3</sup><http://www.sciencetree.net>

Figure 3: Example of an Academic Genealogy Tree Built from Lattes Data



academic lineage, and the role of interdisciplinary formation on the evolution of specific research fields. Ultimately, it would allow us to better comprehend the evolution of science and consequently, of our society. We note, however, that our current version of the system is still beta and its development is part of our future work.

## Acknowledgments

This research is funded by grants from CAPES, CNPq and FAPEMIG. The last author was also supported by IEAT, the UFMG Institute for Advanced Transdisciplinary Studies.

## 5. REFERENCES

- [1] A.-L. Barabási, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, and T. Vicsek. Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications*, 311(3):590–614, 2002.
- [2] F. Benevenuto, A. H. F. Laender, and B. L. Alves. The H-index paradox: your coauthors have a higher H-index than you do. *Scientometrics*, 106(1):469–474, 2016.
- [3] S. Chang. Academic genealogy of american physicists. *AAPPS Bulletin*, 13(6):6–41, 2003.
- [4] S. V. David and B. Y. Hayden. Neurotree: A collaborative, graphical database of the academic genealogy of neuroscience. *PLoS ONE*, 7(10):e46608, 2012.
- [5] W. Dores, F. Benevenuto, and A. H. F. Laender. Extracting Academic Genealogy Trees from the Networked Digital Library of Theses and Dissertations. In *Proc. of the 16th ACM/IEEE-CS Joint Conf. on Dig. Libraries*, pages 163–166, Newark, NJ, 2016.
- [6] A. A. Ferreira, M. A. Gonçalves, and A. H. F. Laender. A Brief Survey of Automatic Methods for Author Name Disambiguation. *SIGMOD Record*, 41(2):15–26, 2012.
- [7] E. A. Fox, M. A. Gonçalves, G. McMillan, J. L. Eaton, A. Atkins, and N. A. Kipp. The networked digital library of theses and dissertations: Changes in the university community. *J. Comp. in H. Educ.*, 13(2):102–124, 2002.
- [8] R. E. Hart and J. H. Cossuth. A family tree of tropical meteorology’s academic community and its proposed expansion. *Bull. of the Amer. Meteor. Soc.*, 94(12):1837–1848, 2013.
- [9] A. Jackson. A labor of love: the mathematics genealogy project. *Notices of the AMS*, 54(8):1002–1003, 2007.
- [10] J. Lane. Let’s make science metrics more scientific. *Nature*, 464(7288):488–489, 2010.
- [11] H. Lima, T. H. P. Silva, M. M. Moro, R. L. T. Santos, W. Meira Jr., and A. H. F. Laender. Assessing the profile of top Brazilian computer science researchers. *Scientometrics*, 103(3):879–896, 2015.
- [12] X. Liu, J. Bollen, M. L. Nelson, and H. Van de Sompel. Coauthorship networks in the digital library research community. *IPM*, 41(6):1462–1480, 2005.
- [13] R. D. Malmgren, J. M. Ottino, and L. A. N. Amaral. The role of mentorship in protégé performance. *Nature*, 465(7298):622–626, 2010.
- [14] E. Tuesta, K. Delgado, R. Mugnaini, L. Digiampietri, J. Mena-Chalco, and J. Pérez-Alcázar. Analysis of an Advisor-Advisee Relationship: An Exploratory Study of the Area of Exact and Earth Sciences in Brazil. *PloS One*, 10(5):e0129065–e0129065, 2014.