# Building the Brazilian Academic Genealogy Tree

Wellington Dores, Elias Soares, Fabrício Benevenuto,
and Alberto H.F. Laender(✉)

Computer Science Departament,
Universidade Federal de Minas Gerais, Belo Horizonte, 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. 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 for managing information about individual researchers and research groups in Brazil.

**Keywords:** Academic genealogy trees · Academic mentorship · Lattes

## 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, citations received and collaborations established [1,9], 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 [2,3,7], in which nodes represent researchers and relations indicate that a researcher was the advisor of another one. However, these efforts have focused on specific scientific fields, such as Mathematics [7] and Neuroscience [3], or have been restricted to a specific community as it is the case of a career retrospect of prominent American physicists [2]. Although limited to specific contexts, overall these efforts show that the analysis of such relationships in the form of a genealogy structure contributes to a greater understanding


J. Kamps et al. (Eds.): TPDL 2017, LNCS 10450, pp. 537–543, 2017.
DOI: 10.1007/978-3-319-67008-9_43

of a scientific community and of its individual values, allowing us to identify the impact generated by individuals in the formation of a community [10].

Complementary to all of them, we have started an ambitious project towards building a large network that records the academic genealogy of researchers across fields and countries [4]. Our preliminary work used data from NDLTD, the Networked Digital Library of Theses and Dissertations[1] [6], 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[2]. Maintained by CNPq, the Brazilian National Council for Scientific and Technological Development, this platform is an internationally renowned initiative [8] that provides a repository of researchers' curricula vitae and research groups, all integrated into a single system. In order to be able to submit any research grant proposal, 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 the 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.

## 2 Building the Academic 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 curriculum 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. Thus, for each one of these two sections, we wrote specific XPath queries to extract each required piece of data (e.g., the researcher's name and the names of her advisees). Note that the output of this procedure is actually a directed acyclic graph, since in her academic life

[1] http://www.ndltd.org.
[2] http://lattes.cnpq.br.

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.

According to Algorithm 1, in order to build the individual 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 such trees, 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 17), 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 6); (ii) search G for the nodes of the researcher's PhD and Master's advisors, creating them if they do not yet exist or updating them otherwise, and then connect them to the researcher's node (lines 7 to 10); (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 11 to 16).

A critical component of our algorithm is the search function in lines 4, 7 and 12. 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 researcher's curriculum, since this requires some action from the researcher when updating her curriculum, which is not always done. Thus, to overcome this problem, we have implemented a simple, but quite effective strategy to handle this typical name disambiguation problem [5], which considers the following parameters as input for a similarity function: the researchers' names, the names of their institutions, the titles of their theses or dissertations, and the respective years of defense. However, a detailed discussion of this similarity function is out of the scope of this paper.

## 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 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 has advised on average 3.81 PhD or Master's students. Despite the average size of the trees

**Algorithm 1.** AGT Bulding Procedure

```
Input: A set C of Lattes Curricula;
Output: A graph G with all AGT's built;
Sort C by the researchers' PhD degree year;
Set G empty;
foreach Curriculum c in C do
    Search G for the researcher's node n;
    if there is no such a node in G then Create node n;
    else Update the academic attributes of n;
    Search G for the nodes p and m of the researcher's PhD and Master
     advisors;
    if either p or m are not found then Create them;
    else Update the academic attributes of p and m;
    Connect p and m to n;
    foreach advisee in c do
        Search G for the advisee's node a;
        if there is no such a node in G then Create node a;
        else Update the academic attributes of a;
        Connect a to n
    end
end
```

being 40.19, the 10 largest trees have more than 5,000 nodes, although 80% of them have less than 20 nodes, as shown in Fig. 1(left graph). On the other hand, almost half of the trees have depth 1, as also shown in the same figure (right graph). We also noted that Brazilian trees are about 6.77 times wider than deeper. This number is much higher in comparison with the same ratio for trees built from NDLTD [4], 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.

**Table 1.** Characterization of the AGT's

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

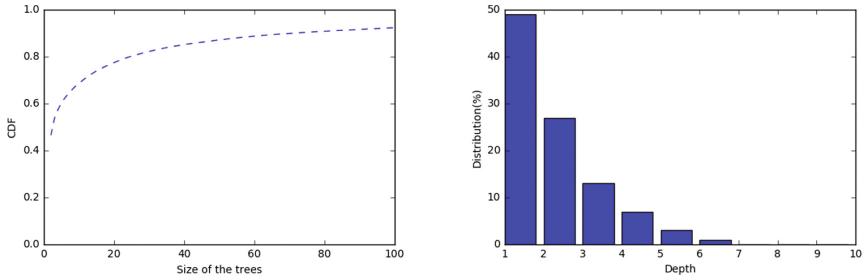


**Fig. 1.** Cumulative distribution function of the tree sizes and tree depth distribution

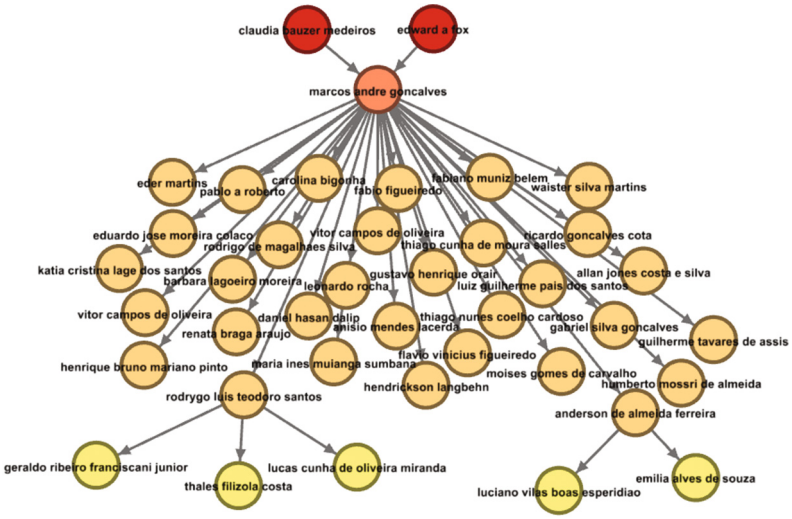


**Fig. 2.** Example of an academic genealogy tree built from Lattes (Color figure online)

## 4 Conclusions and Future Work

In this work, we used data crawled from the Lattes Platform to build a preliminary version of the Brazilian academic genealogy tree. 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 NDLTD [4], allowed us to identify many challenges that we need to tackle towards developing a large repository that records the academic genealogy of researchers across fields and countries. More importantly, we have developed the first version of a system[3] that deploys the dataset studied here and allows users to browse the trees.

[3] Available at http://www.sciencetree.net.

To briefly illustrate the potential of this system, Fig. 2 shows an excerpt of the genealogy tree of Dr. Marcos André Gonçalves, a Brazilian associate professor at the Universidade Federal de Minas Gerais (UFMG), who is a well-known researcher in the digital library community. As we can see, the node colors represent the levels in his tres. The red nodes correspond to Dr. Gonçalves' advisors during his Master's and PhD studies, respectively Prof. Claudia Bauzer Medeiros, from UNICAMP, Brazil, and Prof. Edward A. Fox, from Virginia Tech, USA. The main subtree (the orange one) 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 this kind of tree we hope to better understand a research lineage. Moreover, we believe this system represents a preliminary step towards the understanding of more important questions related to science, which we will be able to 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. 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 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 the current version of our system is still beta and its development is part of our future work.

**Acknowledgments.** This research is funded by projects InWeb (grant MCT/CNPq 573871/2008-6) and MASWeb (grant FAPEMIG/PRONEX APQ-01400-14), and by the authors' individual grants from CAPES, CNPq and FAPEMIG. Fabrício Benevenuto and Alberto H.F. Laender are also supported by the Humboldt Foundation and IEAT, respectively.

## References

1. Benevenuto, F., Laender, A.H.F., Alves, B.L.: The H-index paradox: your coauthors have a higher H-index than you do. Scientometrics **106**(1), 469–474 (2016)
2. Chang, S.: Academic genealogy of american physicists. AAPPS Bull. **13**(6), 6–41 (2003)
3. David, S.V., Hayden, B.Y.: Neurotree: a collaborative, graphical database of the academic genealogy of neuroscience. PLoS ONE **7**(10), e46608 (2012)
4. Dores, W., Benevenuto, F., Laender, A.H.F.: Extracting academic genealogy trees from the networked digital library of theses and dissertations. In: Proceedings of JCDL, pp. 163–166 (2016)
5. Ferreira, A.A., Gonçalves, M.A., Laender, A.H.F.: A brief survey of automatic methods for author name disambiguation. SIGMOD Rec. **41**(2), 15–26 (2012)
6. Fox, E.A., Gonçalves, M.A., McMillan, G., Eaton, J.L., Atkins, A., Kipp, N.A.: The networked digital library of theses and dissertations: changes in the university community. J. Comp. H. Educ. **13**(2), 102–124 (2002)

7. Jackson, A.: A labor of love: the mathematics genealogy project. Not. AMS **54**(8), 1002–1003 (2007)
8. Lane, J.: Let's make science metrics more scientific. Nature **464**(7288), 488–489 (2010)
9. Liu, X., Bollen, J., Nelson, M.L., Van de Sompel, H.: Coauthorship networks in the digital library research community. IPM **41**(6), 1462–1480 (2005)
10. Malmgren, R.D., Ottino, J.M., Amaral, L.A.N.: The role of mentorship in protégé performance. Nature **465**(7298), 622–626 (2010)