Extracting Academic Genealogy Trees from the Networked Digital Library of Theses and Dissertations*

Wellington Dores Fabrício Benevenuto Alberto H. F. Laender

Department of Computer Science
Universidade Federal de Minas Gerais
Belo Horizonte, MG, Brazil
{wellingtond, 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 give a first step towards building a large repository that records the academic genealogy of researchers across fields and countries. We crawled data from the Networked Digital Library of Theses and Dissertations (NDLTD) and develop a framework to extract academic genealogy trees from this data and provide a series of analyses that describe the main properties of the academic genealogy trees. Our effort identified interesting findings related to the structure of academic formation, which highlight the importance of cataloging academic genealogy trees. We hope our initial framework will be the basis of a much larger crowdsourcing system.

1. INTRODUCTION

Along the humanity history, science has evolved in different directions and rhythms, allowing humans to approach the main challenges of each era. For example, some disciplines like Computer Science or Neuroscience are considered to be in their infancy in comparison with others such as Physics and Biology. In this context, many researchers have played a vital role on the different research areas and are of extreme importance to this dynamics of science, not only for their findings, usually accounted by means of their publications, but also in the formation of new researchers.

The formation of researchers over the years is usually represented as an academic genealogy tree [6, 7, 12], which is a representation very similar to the well-known genealogy tree. It simply consists of a direct graph, where nodes represent researchers and relations indicate that a researcher was the advisor of another. Tracking this sort of relationship over time is important for many reasons. For example, it 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 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.

Despite its clear importance, little attention has been given to preserving the academic genealogy. The identification of researchers' ancestors is indeed a challenging task as current repositories of theses and dissertations are usually cataloged in a decentralized way through many local digital libraries [5]. As a consequence, existing efforts on this context have focused on specific fields, such as Mathematics [12] and Neuroscience [7], or on the use of well maintained repositories but restricted to a specific country [19]. Although these efforts are valuable for providing answers to important research questions, they do not provide the big picture about the academic genealogy nor allow us to comprehend many aspects behind the formation of scientists across fields and countries.

In this paper, we give a first step towards building a large network that records the academic genealogy of researchers across fields and countries. 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. Our preliminary effort here consists of constructing and analyzing academic genealogy trees from a large existing collection of electronic theses and dissertations, the Networked Digital Library of Theses and Dissertations – NDLTD[1] [8]. To do that, we crawled the entire NDLTD as it records theses and dissertations from many institutions around the world and from different disciplines. Then, we developed a basic framework to extract information from the NDLTD, identify and disambiguate authors, and identify their advisor relationships. Finally, we carried out a series of analyses that describe the main properties of the genealogy trees we were able to construct. We hope our initial framework can

*This research is funded by projects InWeb (MCT/CNPq grant 573871/2008- 6) and MASWeb (FAPEMIG/PRONEX grant APQ-01400-14), and by the authors' individual grants from CAPES, CNPq and FAPEMIG.

JCDL '16, June 19 - 23, 2016, Newark, NJ, USA

[1] http://www.ndltd.org/

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 briefly survey existing efforts in this field. Then, we describe our methodology and the data gathered from the NDLTD, and present our characterization study of academic genealogy trees and discuss our preliminary findings. Finally, we conclude the paper and provide directions for future work.

2. RELATED WORK

Since Newman's seminal work on scientific collaboration networks [16], there has been a tremendous effort aiming to understand the structure of such communities [13, 15], characterize their patterns of collaboration [9, 17], and analyze their evolution throughout the years [1, 2]. Likewise, there has also been some recent effort to document, analyze and classify the advisor-advisee academic relationship networks. For instance, Chang [6] presents a career retrospect of prominent American physicists and describes their academic genealogy trees. Jackson [12] seeks to maintain the genealogy tree of all mathematicians around the world[2], whereas David and Hyden [7] have maintained the genealogy tree of researchers in the neuroscience field[3]. In common, these projects collect data about all researchers that work in those fields in order to establish their academic relationships. Other relevant efforts are the PhDTree[4] and the Academic Family Tree[5], which try to document the academic family trees of researchers worldwide and share their information trough a Wiki. Both projects rely on a crowdsourcing system to keep their data up-to-dated.

Other works aim to analyze, understand, and model the structures and properties of such specific academic networks. For instance, Tuesta *et al.* [19] have analyzed the advisor-advisee relationship in the Brazilian exact and earth science field. Their intention was to explore the correlation between time and productivity throughout the advising relationship. Malmgreen *et al.* [14] have investigated the role of mentorship in protégé performance by studying mentorship fecundity using data from the Mathematics Genealogy Project. On the other hand, Rossi and Mena-Chalco [18] have introduced some topological metrics to characterize the individuals in academic genealogy trees, whereas Griffiths [10] has shown that a class of genealogy trees is related to unlabelled graph-theoretical trees, which allows to solve some counting problems associated to such trees.

Differently from the above studies, our paper gives the first, yet preliminary, step towards building a large repository that records the academic genealogy of researchers across fields and countries. Thus, to the best of our knowledge, our effort is complementary to the existing ones.

3. THE NDLTD GENEALOGY TREES

To construct the academic genealogy trees from NDLTD data, we first gathered data from all researchers with records on this digital library. The NDLTD is formed by collections of Eletronic Thesis or Dissertation (ETD) records from hundreds of academic institutions around the world. Its repository is mainly maintained by harvesting individual ETD records from other sources by using the Open Archives Metadata Harvesting Protocol (OAI-PMH)[6], which are then encoded in XML. Calado *et al.* [5], for instance, describe one of such efforts, in which ETD records were created by automatically extracting data from thousands of webpages.

3.1 Dataset

After collecting all ETD records stored in the NDLTD, the respective XML documents were parsed and transformed into a CSV file keeping only the ETD fields showed in Table 1. In total, 4,588,474 ETD records were collected. Despite the Dublin Core initiative[7] to standardize the set of adopted metadata, many collections do not follow the proposed specification causing the lost of important information. For instance, there were cases in which some metadata did not follow the required format, or were simply not present in the ETD records. Additionally 279,811 records returned the status deleted.

Table 1: Metadata gathered from NDLTD.

Field	Records with values
Title	4,166,668
Creator	4,116,325
Subject	2,222,814
Description	3,588,628
Publisher	2,451,501
Contributor	1,737,371
Date	3,986,625
Type	2,973,366
Format	1,683,547
Identifier	4,162,019
Language	3,550,054
Coverage	125,804
Rights	877,778
Thesis.degree	1,532

3.2 Data Extraction

Finished the data collection, the second step was to extract specific data from the EDT records. However, to find a general solution to clean such data is not a simple task. Thus, we adopted an intermediate solution between an automatic process and a totally manual intervention. For this, we first removed all fields whose content included text in non-occidental characters. Then, we applied some data cleaning procedures to eliminate inconsistent content (e.g., general abbreviations such as "s.n." and "s.l.", emails and general comments such as "Text Here", among others). In this process, the fields *Creator* and *Contributor* were the most important ones because they would be used to link the theses' authors with their respective advisors. This task, however, presented a major challenge to our purposes due to the limited number of records (1,737,371, which is about 38% of the total) containing the field *Contributor*. Finished the cleaning process, only 638,812 records were considered to construct the genealogy trees, resulting a forest with 95,169 components.

3.3 Name Disambiguation

The main task in the construction of the genealogy trees is to link the researcher's name found in the *Contributor* field

[2]http://genealogy.math.ndsu.nodak.edu/
[3]http://neurotree.org/
[4]http://phdtree.org/
[5]http://academictree.org/
[6]Proposed by the Open Archives Initiative (OAI)
[7]http://dublincore.org/

Figure 1: Excerpt of the genealogy tree constructed from NDLTD data.

of each EDT record with the name of some researcher in the *Creator* field of another record, since the field *Contributor* is primarily used to store the advisor's name. To achieve such a goal, we adopted a simple solution based on the BK-tree data structure [4]. A BK-tree is a metric tree specifically designed to discrete metric spaces. Thus, it provides a simple and effective solution to search for the most similar names in our dataset, since our major difficulty here is the lack of information to help correctly matching two names. The BK-tree allows one to search for strings that are similar to the query by using a Jaro string comparator. In our case, we used a similarity threshold of 95%.

3.4 Characterizing the Genealogy Trees

The legacy of a researcher can be measured not only in terms of her publications and scientific discoveries, but also in terms of the formation of other researchers. Next, we analyze a small example of an academic genealogy tree as an attempt to visualize this second part of a researcher's legacy, which is the research families and communities that emerge around a particular researcher. Figure 1 shows an excerpt of the genealogy tree of researchers from the Graduate Program in Animal Science of the Universidade Federal de Minas Gerais (UFMG) in Brazil. The colors in the figure represent the graph modularity, which can be understood as a "family" core of researchers. The tree includes a main subtree (the red one), which includes the graduate (PhD and MSc) students that have been advised by Prof. Élvio Carlos Moreira, a senior faculty member in that program. His tree spans seven other trees which, in turn, span three additional subtrees. Thus, by analyzing such a kind of tree we hope to be able to better understand the role of these families on coauthorship and community formation. More important, this example elucidates that the system we aim at developing can be helpful for those interested in understanding the impact that individual researchers have in a community in terms of scientific formation.

We now investigate some metrics that describe the structure of the trees we have been able to construct. The *width* is the number of advisees a researcher has advised (the researcher's out-degree), whereas the *depth* represents her lineage size. Together, these two metrics provide an overview of the legacy of a researcher in terms of academic formation. In our example in Figure 1, Prof. Moreira's tree has width 60, which is the number of all his advisees, and depth 5, which is the size of his largest lineage. In our dataset, an average researcher has an advising rate of 0.30 researchers. In contrast, the advisor with the highest advising rate formed 169 students. In fact, the 100 most prolific researchers advised 5,948 students, which corresponds to 7% of all nodes in the trees we analyzed. Figure 2 shows the distribution of these two metrics in our dataset.

These results also suggest that academic genealogy trees are much wider than deeper. In fact, if we consider the width and the depth of a tree as its largest width and depth, respectively, we noted that trees are on average 2.48 wider than deeper. The Pearson correlation coefficient between the width and the depth of a tree is 0.60, which suggests that the largest trees are also the deepest ones. In order to better understand this correlation between width and depth, we have considered a variation of the well known *h-index*, adapted to the context of academic genealogy trees.

The h-index [3, 11] is a metric originally proposed to measure a researcher's scientific output. Its calculation is quite simple as it is based on the researcher's set of most cited publications and the number of citations they have received. More specifically, a researcher has an h-index h if she has at least h publications that have received at least h citations.

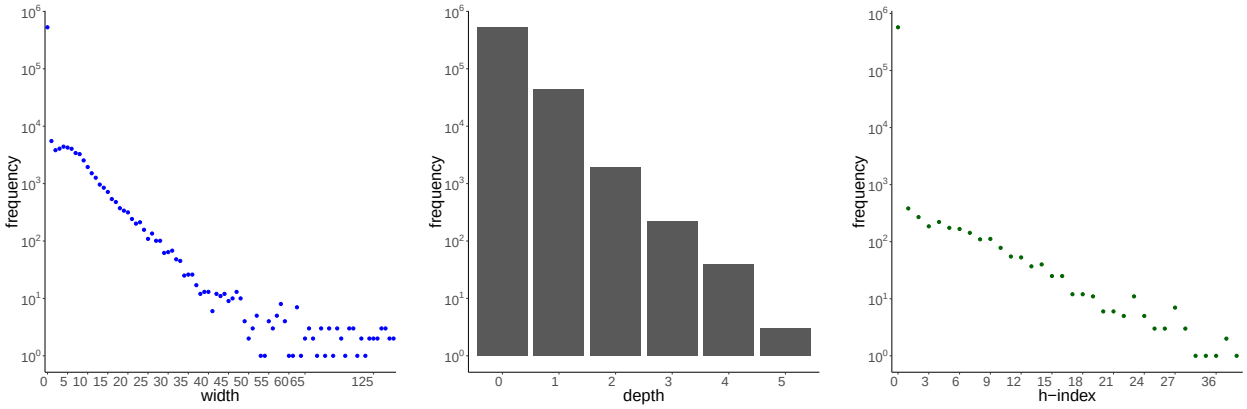

Figure 2: Width, depth, and h-index distributions of genealogy trees.

In our context, the metric is computed slightly different as proposed by Rossi and Mena-Chalco [18]. A researcher has an genealogy h-index h if she has at least h advisees and, at least one of them, has advised at least h advisees as well. Thus, if a researcher has at least 10 advisees and one of them has advised at least 10 other advisees, her genealogy h-index is 10. We can note from Figure 2 that most researchers have a low h-index, but some of them reach really high values. For example, the largest h-index in our dataset is 76.

4. CONCLUSIONS AND FUTURE WORK

In this work, we used data crawled from the Networked Digital Library of Theses and Dissertations (NDLTD) to construct academic genealogy trees. Although still preliminary, our effort identified a number of interesting findings related to the structure of academic formation, which highlight the importance of cataloging academic genealogy trees. Our effort showed that NDLTD is a valuable collection for this purpose and that it also 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. First, we were able to identify researcher names and advisor relationships of a relatively small amount of records. This is because the content on these entries is free text and present many challenges for being properly processed. Proposing an algorithm able to unveil more nodes for our trees is in our research agenda. Second, we aim at identifying the research disciplines of the researchers based on specific EDT fields and also incorporate data from other sources in addition to NDLTD. Finally, we plan to develop our system in a way that researchers and other interested people can help us to curate our genealogy trees, which may also pose other challenges.

5. REFERENCES

[1] B. L. Alves, F. Benevenuto, and A. H. F. Laender. The Role of Research Leaders on the Evolution of Scientific Communities. In *Proc. of WWW (Companion Volume)*, pages 649–656, Rio de Janeiro, Brazil, 2013.

[2] A.-L. Barabâsi, 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.

[3] 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.

[4] W. A. Burchard and R. M. Keller. Some approaches to best-match file searching. *CACM*, 16(4):230–236, 1973.

[5] P. Calado, M. A. Gonçalves, E. A. Fox, B. A. Ribeiro-Neto, A. H. F. Laender, A. S. da Silva, D. de Castro Reis, P. A. Roberto, M. V. Vieira, and J. P. Lage. The Web-DL Environment for Building Digital Libraries from the Web. In *Proc. of JCDL*, pages 346–357, Houston, USA, 2003.

[6] S. Chang. Academic genealogy of american physicists. *AAPPS Bulletin*, 13(6):6–41, 2003.

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

[8] 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.

[9] W. Glänzel. National characteristics in international scientific co-authorship relations. *Scientometrics*, 51(1):69–115, 2001.

[10] R. C. Griffiths. Counting genealogical trees. *J. of Math. Biol.*, 25(4):423–431, 1987.

[11] J. E. Hirsch. An index to quantify an individual's scientific research output. *PNAS*, 102(46):16569–16572, 2005.

[12] A. Jackson. A labor of love: the mathematics genealogy project. *Notices of the AMS*, 54(8):1002–1003, 2007.

[13] 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.

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

[15] G. V. Menezes, N. Ziviani, A. H. F. Laender, and V. A. F. Almeida. A Geographical Analysis of Knowledge Production in Computer Science. In *Proc. of WWW*, pages 1041–1050, Madrid, Spain, 2009.

[16] M. E. Newman. The structure of scientific collaboration networks. *PNAS*, 98(2):404–409, 2001.

[17] M. E. Newman. Coauthorship networks and patterns of scientific collaboration. *PNAS*, 101(s. 1):5200–5205, 2004.

[18] L. Rossi and J. P. Mena-Chalco. Caracterização de árvores de genealogia acadêmica por meio de métricas em grafos. *Anais do XXXIV CSBC*, pages 21–32, 2014 (in Portuguese).

[19] 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.