Genealogy Tree: Understanding Academic Lineage of Authors via Algorithmic and Visual Analysis

Sandra Anil, Abu Kurian, Sudeepa Roy Dey*, Snehanshu Saha, Ankit Sinha

Department of Computer Science and Engineering, PESIT Bangalore South Campus, Bangalore, Karnataka, INDIA.

ABSTRACT

Ancestry and Genealogy tree are proven tools to determine the lineage of any person and establish dependencies among individuals. Genealogy tree can be exploited further to gain information about the researcher and his scholastic lineage which is of paramount importance in today's world of computer technology. This insight into academic genealogy could be ways of helping PhD students achieve academic socialization within the discipline, by making explicit connections that may be influential. Awareness of his scientific heritage, gives the user a broader perspective of his own research project. This paper also highlights and investigates how this academic network is exploited by certain researchers using various visualization tools. It was observed during this work that the credibility and influence factor is determined by the various citations obtained by an author and to improve their rankings in various forums, they tend to collaborate in their academic circle and boost their citation count. A recent trend among researchers is to form communities based on their academic relationships and rely on copious citations for their mutual benefit. Tracing the genealogical relationships can be helpful in detecting such communities and also create a more quality aware metrics using a lineage independent model for computation of author level metrics.

Keywords: Genealogy tree, Copious Citation, Genealogical Citation, Community, Advisor, Advisee, Sibling.

Correspondence

Sudeepa Roy Dey
PESIT, South Campus, Electronic City, Bengaluru, Karnataka, INDIA.
E-mail: sudeepar@pes.edu

INTRODUCTION

Genealogy is an account of descent of a person, family or group from an ancestor or from older forms. It is the study of the history of the past and present members of a family or families. Historical records are used for genealogical research. Over the years, the number of people pursuing PhD has increased, leading to an exponential growth of the academic genealogical tree. With this rising number, keeping track and documentation of scholastic relationships between scientists has become difficult. An attempt in this direction has been made by the American Mathematical Society, by means of their Mathematics Genealogy Project. Their objective is to catalogue the complete mathematics community. It gives information of an author, his ancestry and lineage in the tree, along with his dissertation and year of being awarded the degree. A similar approach has been put forth in this paper for the discipline of Computer Science along with some new

metrics. Mining huge databases is a complex task and lot of algorithms are derived to create a database of authors.[1] The genealogy tree will hold information about all the scientists who have contributed to the field at various research–level.

A database is built from the contributions of the scientists who give inputs like their supervisors, dissertation details, year and affiliating institute. A graph database is formed which is then searched based on user query.[2] The tree obtained can be further based on two criteria: author or domain. The tree based on author describes the author's heritage and his descendants. Details about the author's degree are also provided in this genealogical tree.

Since Computer Science can be perceived as an umbrella housing many domains which has multiple research areas within them, the domain based tree traces the complete hierarchy of scientists who have contributed to it significantly.

The proposed model will be used further to trace various citation patterns among authors under the same advisor i.e. siblings to each other, who form communities to inflate their citation count by mutually citing one another consistently.[3] Such networks are also observed within an advisor and advisee as well. The model proposes a threshold, which is the ratio

of community by total citation and the authors who exceed this threshold are identified as potential lineage or community dependant authors. Our motive is to trace and highlight such communities and their patterns of citations.

MOTIVATION AND TECHNICAL CONTRIBUTION

With the exponential growth in researchers and their publication in various journals, the need to trace the quality of work and rank the emerging authors has also increased.[4] A genealogy tree will not only help explore the pedigree and the ancestry of an author and the domain, it will also be used to investigate the citations received and do the in–depth analysis. With this motivation, a similar software model has been put forth by this paper for the department of Computer Science.[5]

The project is based on a graph database which is built by contributions of the scientists who provide input details like dissertation, and year and institute of procuring degree. A genealogy tree is created with nodes as authors/supervisors and link represents advisees. Finally, a revenue model is also presented where an Author Lineage Score is calculated based on various parameters and the authors are classified regarding this.

COMMUNITY CITATION

Community of authors is defined as a group of authors who collaborate with each other and frequently cite each other's work to increase their citations count.[6] These communities can be formed as a network of an advisor and his students or alternately between students under the same advisor in which case it will be a sibling network.

The proposed method is to subdivide the author citation matrix into block matrices for each individual author. A block matrix is a collection of sub matrices where each sub matrix is an independent matrix having rows and columns. Each block matrix represents the local network of an author at each level. Block matrix is an adjacency matrix i.e. if A[i, j] = 1, then j is a member of local network of i. The first row of the matrix represents the children of the author i.e. one level below the author in the genealogical tree. Similarly, the second row represents the children of the author two levels down, and the third and fourth row represent the parent and grandparent of the author respectively.

The block matrix division in Figure 1. shows A, B,C and D are the sub matrices. Each sub matrix is a dedicated adjacency matrix of an author. In the sub matrix A, author A_1 is one hop below A in the genealogical tree. A_2 is two hops below A, A_4 is two hops above A in the genealogical tree. Similarly, the local network of authors B, C and D are also represented in the block matrix as independent sub matrices.

Figure 1: Block Matrix division.

Determining the local network of an author using block matrices reduces the computational complexity, since traversal of the JSON file is eliminated. The time complexity of the algorithm, to determine the local network by parsing the JSON file is $O(n^3)$. Block matrix division gives a time complexity of $O(n)$.

Community Citation is defined as the total citations obtained by the author from his citation community. It is a subset of the Genealogical Citations of the author.

A special case of community citation is the copious citation. Copious citation refers to citations between pair of authors. The authors mutually cite each other's work and by this they both benefit from it.

A threshold value for detecting a community must be determined from historical data. An appropriate trend detection algorithm must be used to decide the threshold. The threshold will have a lower bound and upper bound i.e. the threshold will be a finite range in practice.

The total genealogical citations are obtained by adding the citations of the authors present in the local network, from the author citation matrix. The ratio of the genealogical citations with the total citation is computed and compared against the threshold and the author is labelled as lineage dependant if the ratio is greater than the threshold.

NON-GENEALOGICAL CITATION

Non–genealogy citations (NGC) of an author are defined as citations obtained from authors who are not present in the genealogy network of that author. Non–genealogy citations can be computed as the difference between total citations and genealogy citations.

The citations of authors are represented in an all author matrix. Each element A[i,j] denotes the number of times author j has cited author i. Every author is assigned a unique ID in the all author matrix.

Figure 2: All Author Matrix. In the matrix, for author A self-citations are 10. Author A has cited Author B 7 times, Author D 15 times and has not cited Author C even once. The complete matrix can be interpreted similarly.

The following cases may arise while computing the NGC from an author network:

1. Unique Name Case: The authors name is unique in the complete network.

2. Multiple Name Case: There is more than one author having the same name.

3. Two Advisor Case: The author will have more than one advisor, irrespective of the uniqueness of his name.

4. Multiple Name and Two Advisor Case: Authors sharing the same name, few of them will have multiple advisors.

As indicated in Figure 2 the complete matrix is computed. The total citations of an author represented by Y are obtained from the all author matrix. To compute the genealogical citations first the authors who are part of the genealogy network of the author are identified. Then the sum of citations of all these authors, represented by X, is obtained from all author matrices. NGC is calculated as

$$NGC = Y - X$$

TECHNICAL IMPLEMENTATION

Creation of Database

A sample graph database was created to test the Community Detection method. Each node in the database had 8 fields. These fields were name of the author, the author's teacher (labelled as Level 1), the author's students (labelled as Level 2), the author's Ph.D. thesis, the Institute at which he acquired his Ph.D., the Country of origin, domain, and the number of overall citations he has received. Each teacher or student present in the above-mentioned attribute Level 1 or Level 2 will be stored as a key–value pair. The data structure will store the name along with the number of citations to the author under consideration. This will help to detect any irregular or false citations that the author under study may be receiving by his local community.

The Visualization toolbox

A web application was created to visualize community detection. The website[7] offers 3 options. First is to display the local

network of any given author, the second option is to display the details of the author and the third option is for community detection. When the user refers to the local network it is the community around the author that includes the author his teacher, his teacher's teacher, the author's students and the student's student. When an author is queried and the third option is selected the software will check whether there is a presence of local citations within the author's local community and display the results.

Figure 3 and 4 depict the pages where an author Joseph Cook is search and Figure 4. Genealogy tree of Joseph Cook is shown. The authors represented by green nodes are the nodes one level below, Joseph Cook i.e his advisees and the authors represented by blue nodes are the nodes two levels below i.e advisees of his advisees (Figure 5 represents the interactive page of the author).

The visualization shown can be viewed and interacted with, by visiting the website http://gt.sahascibase.org/

For the creation of a graphical view, a JavaScript based library,[8] vis.js, was used. When the user clicks the network display button, a GET request is sent to our backend server which in turn requests the Neo4J database with recursive queries.
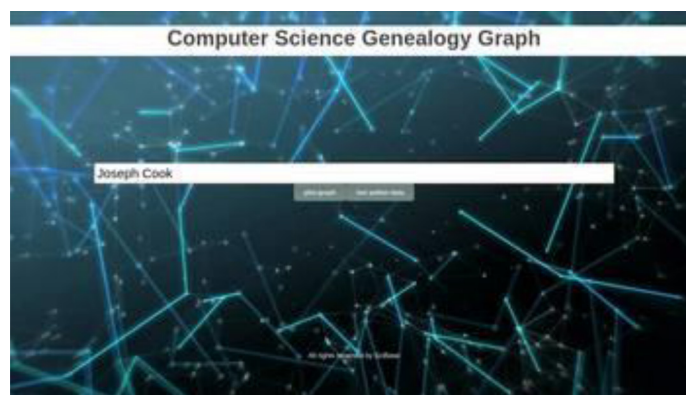
Figure 3: Home Page of the website.Search an author page
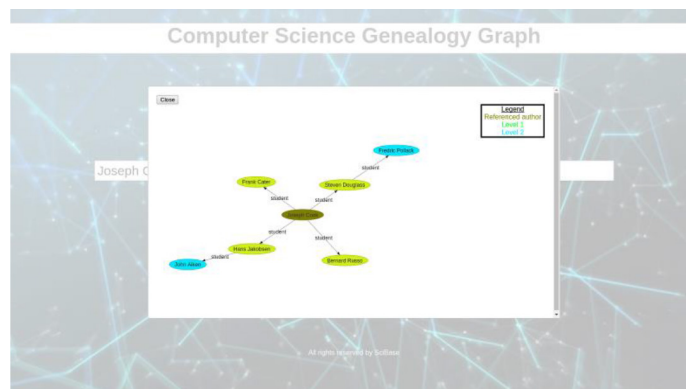
Figure 4: Genealogy tree of the author.
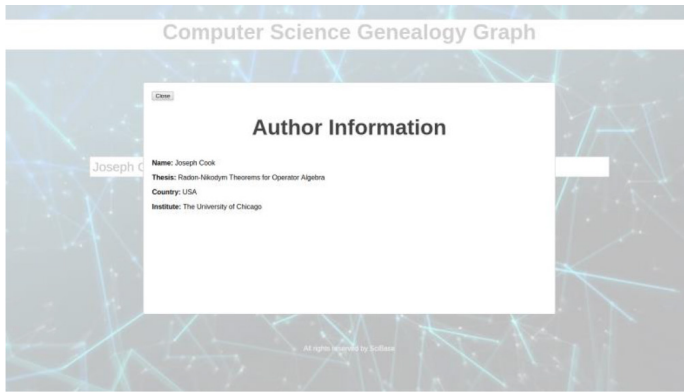
Figure 5: Author information. The profile of Joseph Cook i.e institute, thesis and country are displayed.

Initially, 2 empty arrays NODES and EDGES are defined and the name of author for whom the graph is being generated, is pushed in the NODES array.

Each recursive query then returns the next level of related authors in the community.[8] For each query result, the name of all the next level authors is pushed in the NODES array, and a JavaScript object in the form

{

'from' : currentAuthor,

'to' : nextLevelAuthor

}

is created and pushed in the EDGES array.At the end of recursive query calls, the formed NODES and EDGES arrays are sent back to the frontend, where with a call to vis.js library function along with providing it the NODES and EDGES array in proper format, a graph as shown in Figure 4 is created.

1. Final Neo4j database

There are two types of nodes in the database as shown in Figure 6. The purple nodes are the author nodes and the pink nodes are the article nodes. There are three types of relations PARENT_OF, CITED_BY and AUTHORED_BY. The
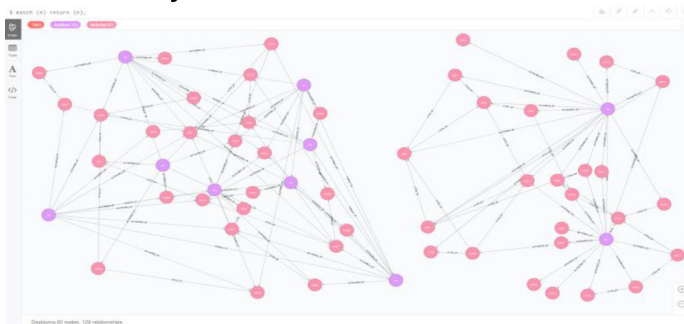
Figure 6: Neo4j database.

author nodes are connected by the PARENT_OF relation. The author node which is connected to another author node by this relation is the advisor of that author node. The articles are connected by the CITED_BY relation, indicating articles which have been cited by other articles. The author and article nodes are connected by AUTHORED_BY relation indicating the author of each article

CONCLUSION AND FUTURE WORK

This paper has highlighted certain key aspects of the author metrics and investigated the citations based on the source and not just the count. The genealogy tree structure is exploited to generate the adjacency matrix, which is then used to derive various metrics such as Genealogy citations (GC), Non-genealogy citations (NGC) which have potential to filter all the contribution towards unethical citation boosting. The authors of this paper humbly put across the following contributions:

- A sample database of authors and fields were name of the author, the author's teacher (labeled as Level 1), the author's students (labeled as Level 2), and the author's Ph.D. thesis, the Institute at which he acquired his Ph.D., the Country of origin, domain, and the number of overall citations he has received.

- Author level metrics derived exclusively with genealogy tree as the backbone. These metrics are further validated through the web application.

- Community detection Algorithms designed to exploit the genealogy tree to the optimum and trace the susceptible population that influence the citation counts of an author in various ways.

- A web application is created to visualize community detection as well.

- Devise a scoring model to highlight an author's influence across his /her lineage and community.

- The visualization shown can be viewed and interacted with, by visiting the website http://gt.sahascibase.org/

Investigating the pattern of citations[7] and measuring the true impact of a researcher has always been a debatable yet important aspect of scientometrics. With this software, the authors hope to create a justified visualization model and put forth new author level metrics that can quantify the influence of an author in terms of pure quality research work, thereby converging quantity and quality of research. The software can be used for further analysis of metrics like sibling citations and recursive copious citations. In future, this model can serve as a visual tool to calculate and perform in-depth analysis of citation network.

REFERENCES

1. Ginde G, Saha S, Balasubramaniam C, Harsha RS, Mathur A, Dayasagar BS, *et al.* (2015). Mining massive databases for computation of scholastic indices: Model and quantify internationality and influence diffusion of peer-reviewed journals. In Proceedings of the fourth national conference of Institute of Scientometrics, SIoT.

2. Saha S, Dwivedi A, Dwivedi N, Ginde G, Mathur A. (2015). JIMI, journal internationality modelling index: An analytical investigation. In Proceedings of the fourth national conference of institute of scientometrics, SIoT.

3. Moed HF. Measuring contextual citation impact of scientific journals. Journal of informetrics. 2010;4(3):265-77.

4. Saha S, Jangid N, Mathur A, Narsimhamurthy AM. DSRS: Estimation and forecasting of journal influence in the science and technology domain via a lightweight quantitative approach. Collnet Journal of Scientometrics and Information Management. 2016;10(1):41-70.

5. Ginde G, Saha S, Mathur A, Venkatagiri S, Vadakkepat S, Narasimhamurthy A, *et al.* ScientoBASE: a framework and model for computing scholastic indicators of non-local influence of journals via native data acquisition algorithms. Scientometrics. 2016;108(3):1479-529.

6. Haddow G, Genoni P. Citation analysis and peer ranking of Australian social science journals. Scientometrics. 2010;85(2):471-87.

7. SciBase project: http://gt.sahascibase.org

8. Neo4j Data Structure: https://neo4j.com