RESEARCH ARTICLE

Visualizing academic descendants using modified Pavlo diagrams: Results based on five researchers in biomechanics and biomedicine

W. Brent Lievers

Bharti School of Engineering & Computer Science, Laurentian University, Sudbury, Ontario, Canada

Keywords: academic genealogy, doctoral descendants, mentorship indices, visualization

ABSTRACT

Visualizing the academic descendants of prolific researchers is a challenging problem. To this end, a modified Pavlo algorithm is presented and its utility is demonstrated based on manually collected academic genealogies of five researchers in biomechanics and biomedicine. The researchers have 15–32 children each and between 93 and 384 total descendants. The graphs generated by the modified algorithm were over 97% smaller than the original. Mentorship metrics were also calculated; their h_m-indices are 5–7 and the g_m-indices are in the range 7–13. Of the 1,096 unique researchers across the five family trees, 153 (14%) had graduated their own PhD students by the end of 2021. It took an average of 9.6 years after their own graduation for an advisor to graduate their first PhD student, which suggests that an academic generation in this field is approximately one decade. The manually collected data sets used were also compared against the crowd-sourced academic genealogy data from the AcademicTree.org website. The latter included only 45% of the people and 34% of the connections, so this limitation must be considered when using it for analyses where completeness is required. The data sets and an implementation of the algorithm are available for reuse.

1. INTRODUCTION

Mentorship is a foundational component of academia. Although it can take different forms, many of which are unofficial and uncredited, the formal mentoring relationship between a doctoral student and their advisor(s)[1] is arguably the most important. It is certainly one that has received has received a great deal of research, most of which can be be divided into one of two categories.

One approach is to focus on the student side of the advisor–advisee relationship. For example, various studies have examined the effects that advisors can have on a student's mental health (Levecque, Anseel et al., 2017; Mackie & Bates, 2019), their productivity (García-Suaza, Otero, & Winkelmann, 2020), and their career outcomes (Gaule & Piacentini, 2018; Malmgren, Ottino, & Nunes Amaral, 2010).

Another approach is to consider what these relationships reveal about the advisor. To this end, various ways of quantifying the mentoring productivity—or in biological terms, the

[1] The specific titles applied to this role vary by jurisdiction and institution (e.g., advisor, chair, director, supervisor), but the term *advisor* will be used throughout this paper for consistency.

an open access journal

Citation: Lievers, W. B. (2022). Visualizing academic descendants using modified Pavlo diagrams: Results based on five researchers in biomechanics and biomedicine. *Quantitative Science Studies*, *3*(3), 489–511. https://doi.org/10.1162/qss_a_00205

DOI: https://doi.org/10.1162/qss_a_00205

Peer Review: https://publons.com/publon/10.1162/qss_a_0025

Received: 28 March 2022
Accepted: 25 July 2022

Handling Editor: Ludo Waltman

Corresponding Author: W. Brent Lievers
blievers@laurentian.ca

The MIT Press

fecundity—of a researcher have also been proposed. One obvious metric is to simply count a researcher's *direct descendants* or *children*; that is, those students that a researcher has advised or coadvised. This counting can also be extended over multiple generations to sum a researcher's *descendants* (i.e., *children*, *grandchildren*, *great grandchildren*, etc.): all those who can trace their advisors' lineage back to the original researcher. Recently some have drawn inspiration from publishing metrics such as the *h*-index (Hirsch, 2005) and *g*-index (Egghe, 2006) as alternate ways to assess fecundity. Their mentoring equivalents, the h_m (Rossi, Damaceno et al., 2018) and g_m-index (Sanyal, Dey, & Das, 2020), attempt to quantify mentorship by considering the first two generations of descendants.

Besides these quantitative approaches, more qualitative analyses have also been performed. Academic genealogies have been assembled for nations (Damaceno, Rossi et al., 2019), fields (Kelley & Sussman, 2007; Russell & Sugimoto, 2009), journals (Mitchell, 1992; Montoye & Washburn, 1980), and individual researchers (Bennett & Lowe, 2005; Lv & Chang, 2021). These family trees highlight the mentoring relationships that exist among researchers and can provide insight into a researcher's influence on a field.

Despite this interest, visualizing networks of descendants remains a challenge for prolific researchers. A common approach (Rutter, VanderPlas et al., 2019) is to use a typical family tree representation such as the one shown in Figure 1. Each node in the graph represents an individual and each edge represents an advisor–advisee relationship. Although intuitive, this approach is unsuitable for large numbers of descendants because the aspect ratio of the graph is determined by the number of generations (height in Figure 1) and the number of individuals in each generation (width). As the number of descendants grows, the aspect ratio becomes more extreme, making it more difficult to understand the overall topology of the network. Various forms of radial or circular graphs have been proposed as alternatives that have smaller aspect ratios (Arce-Orozco, Camacho-Valerio, & Madrigal-Quesada, 2017; Grivet, Auber et al., 2006; Huang, Li et al., 2020). A related challenge of the family tree is that trying to pack as many nodes together to address aspect ratio issues makes it more difficult to distinguish who was advised by whom.

The radial layout algorithm proposed by Pavlo, Homan, and Schull (2006) shows potential for academic genealogies (Figure 2) because the distinction between the descendants of different children is clear. Each node is surrounded by a *containment circle* around which the child nodes are placed. The root node uses the entire circle and intermediate nodes use only an outward portion of the circle, the *containment arc*, which is bounded by the straight lines. Unfortunately, in the original algorithm proposed by Pavlo et al. (2006), the size of each containment circle is determined by its parent and the number of siblings. As pointed out by Huang et al. (2020), this approach results in the outermost descendants becoming smaller and smaller as the number of generations increases. If a suitably large initial size is not chosen, the outermost children can become unreadably small. A second issue is that equivalent subtrees have different sizes and shapes depending on the generation in which they occur and the number of siblings they have. Nevertheless, some modifications to the existing algorithm could

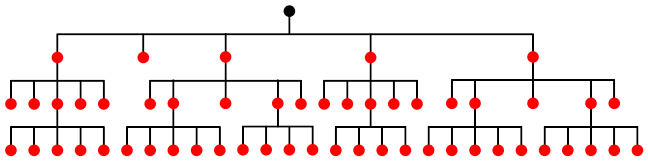

Figure 1. Example of an academic genealogy represented in a traditional family tree format.

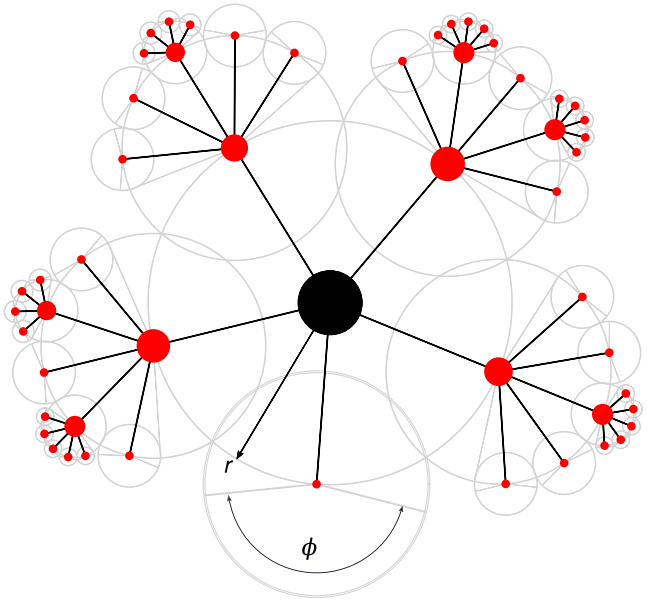

Figure 2. The same academic genealogy information from Figure 1 presented using a traditional Pavlo diagram. The grey circles are the *containment circles* shown to illustrate how the nodes are placed. The grey lines within the circles demark the *containment arc*, the portion of the circle where nodes can be placed. The size and shape of the graph are determined by two user-selected parameters: r, the radius of the containment circle for the root node, and ϕ, the included angle of the containment arc.

eliminate these shortcomings and make the resulting diagrams more compact and more usable for academic genealogies.

A challenge related to evaluating visualization methods is that the data used for assessment are often artificially generated and simplified compared to their real-world equivalents. Real data are preferred to ensure that any characteristic features are present to reveal any shortcomings of an algorithm for that desired application. For example, academic genealogies are highly asymmetric; successful researchers may have many doctoral students, but only a fraction of those will go on to have PhD students themselves. The fecundity of those students will also vary dramatically, both as a result of their individual careers and also due to birth-order effects. We think of a human generation in terms of the 20–30 years needed for a child to be born, mature, and then reproduce. Because an equivalent (albeit shorter) time is needed for academic reproduction, a researcher's first few doctoral descendants will have had longer to reproduce, and will likely have more descendants, than those who graduated near the end of the researcher's career. Finally, a student may also have multiple coadvisors and these relationships must be represented clearly. These unique characteristics underscore the need for comprehensive data to test the usability of different visualization methods. Although there are public sources of academic genealogy data available, the quality of these data remains unclear and must be assessed.

The goals of the current work are threefold. The first is to assemble data sets of academic descendants for five biomechanical/biomedical researchers that are both as comprehensive as possible and demonstrate a variety of possible shapes and sizes. These data, which will be limited to just doctoral advisor–advisee relationships, will be made available for future visualization studies (see Data Availability). The second goal is to introduce an improved version of the Pavlo visualization algorithm and demonstrate its suitability for displaying

academic genealogies using the collected data sets. Finally, the third goal is to analyze the data sets to quantify the fecundity of the five researchers, calculate the time necessary for someone to graduate their first PhD student, and assess the coverage of a particular online repository of academic genealogy data (Academic Family Tree, n.d.). Completing these three goals will help further the study of mentorship within academia, particularly within the fields of biomechanics and biomedicine.

2. METHODS

2.1. Data Collection

Academic genealogy data is spread across a number of sources, including public databases such as Academic Family Tree (n.d.) and the Mathematics Genealogy Project (n.d.), commercial databases such as ProQuest, and university dissertation repositories, as well as the personal websites and online CVs of individual researchers. None of these sources is necessarily comprehensive, correct, or current. Yet assembling representative genealogies, ones that have the characteristic sizes and shapes, is critical to evaluating the efficacy and robustness of a visualization algorithm.

Academic descendant data were collected for five researchers from the fields of biomechanics and biomedicine: Steven A. Goldstein, Wilson C. Hayes, Van C. Mow, Lawrence E. Thibault, and Ronald F. Zernicke. Each researcher received their doctoral degree from 1960 to1980, which is long enough ago to have multiple generations of academic descendants but also recent enough to ensure that most immediate descendants can still be contacted. The five were also chosen to ensure a range of sizes and shapes in their academic trees. Beyond these selection criteria, the individual researchers represent a convenience sample. It should also be noted that all five obtained their degrees in the United States. Although they or their descendants have graduated students in institutions around the world, the vast majority of the researchers in these trees completed their degrees in North America. Therefore, the sizes and shapes of the genealogies may not be representative of those in other countries.

Information gathered from AcademicTree.org and ProQuest was first consolidated together. These data were expanded using public information on researchers' websites, CVs available online, and information in institutional dissertation repositories. When the full text of the dissertations was available electronically, the data collected were validated against the information presented on the title page or in the acknowledgments section. Finally, individual researchers who had a current or past academic appointment were also contacted via email to confirm existing information and request any missing information. Unfortunately, this was not always possible (e.g., retirement, death, lack of contact information, no response).

Descendants were limited to doctoral students for the purpose of this study. This narrow scope was adopted because a doctoral degree is typically required to advise graduate students, which makes holders of these degrees most likely to reproduce. Master's theses were excluded because they receive less coverage in databases, which makes them more difficult to track. Postdoctoral supervision was also excluded as it would require confirmation from one of the parties involved; it doesn't generate a single dissertation-like document that allows for independent verification. Therefore, limiting the scope to PhDs greatly simplified data collection. Finally, any terminal research degree that included a written thesis or dissertation was included regardless of the name (e.g., PhD, DSc, ScD, DEng, MD).

The following information was collected about each descendant: the person's name, the institution from which their doctoral degree was obtained, the year of completion, and the names of all advisors or coadvisors. These data were deemed by our institutional Research Ethics Board (REB) to be public information and not requiring formal consent forms for collection. Nevertheless, when individuals were contacted via email, they were informed that the assembled data would be made publicly available and were given the opportunity to raise concerns. None of the respondents did so. Any students who had successfully defended by the end of 2021 were included. Data entry errors or discrepancies were occasionally uncovered. When conflicts arose, information obtained from sources more closely associated with the individual (i.e., dissertation documents, lab websites, online CVs, or email) were deemed as more authoritative.

It should also be noted that, while generally straightforward, identifying who should be recognized as an "advisor" can at times be difficult to establish. For example, when an advisor moves to another institution, students still enrolled at the original institution may require a local supervisor for administrative purposes. Although they may be listed as the primary advisor, they may not actually perform any of the associated duties. Conversely, others may be actively providing mentorship and support to a doctoral student, yet not receive formal recognition as an advisor or coadvisor. For the purposes of the data collected, we have tried to limit "advisors" to those who both received formal recognition for that role and were not purely administrative. Nevertheless, when direct communication with the advisors and students was not possible, or no reply was received, we had to proceed with the best information available.

The assembled data sets are available for reuse (see Data Availability) as comma-separated value (.csv) and extended markup language (.xml) files. Although every effort was made to ensure they were complete through to the end of 2021, they are acknowledged to be imperfect. Moreover, they will quickly become outdated as new descendants are added over time. Nevertheless, they are the most comprehensive sets of data for these five researchers at the time of writing.

2.2. Modified Pavlo Algorithm

The proposed visualization algorithm is based on the work of Pavlo et al. (2006). As shown in Figure 2, the original algorithm starts with a root node surrounded by a containment circle with a radius r. The value of r is a user-specified parameter and determines the subsequent size and spacing of all other nodes. The child nodes are then equally spaced around the perimeter of the root's containment circle and given their own containment circles, whose radii are determined by geometry. The next set of nodes are then equally spaced around the containment arc, a portion of the containment circle prescribed by the user-selected angle ϕ. The process continues recursively until all nodes are processed.

As highlighted by Huang et al. (2020), a fundamental problem with the root-outward approach of the original algorithm is that the nodes and containment circles for each subsequent generation become progressively smaller. A very large value for r must be selected to ensure that there is adequate spacing between the outermost nodes, and this value is not known *a priori*. A second issue is that individuals with equivalent numbers of descendants will not be represented by equivalently sized containment circles if they occur in different generations or have different numbers of siblings (see Figure 2). This phenomenon violates a common aesthetic principal for graph drawing which holds that "a sub-tree should be drawn the

same way regardless of where it occurs in the tree" (Reingold & Tilford, 1981). More importantly, it also results in a larger graph than is necessary due to excess space being used for some nodes, particularly those without children. Therefore, the overall objectives of this revised algorithm are to ensure that equivalent nodes and subtrees are drawn consistently and the entire graph uses less space than the original algorithm.

The modified Pavlo algorithm will be presented in detail in the following subsections. This process consists of two main steps: determining the size of the containment circles for each node, and determining the orientation of each node. A third optional step will also be presented that assigns unique node and edge colors based on a hue-saturation-lightness (HSL) color wheel.

Python implementations of the original and modified Pavlo algorithms have been made available for reuse. Consult the Data Availability section for more information.

2.2.1. Determining the node containment circle sizes

A major change to the algorithm is the order in which the size of the containment circles is calculated. The original algorithm relied on a root-outward approach. The user would select the radius, *r*, of the containment circle for the root node, and the sizes of all the subsequent circles were determined recursively based on *r* and the number of children in each generation. Unfortunately, this method requires an interactive selection of *r* to ensure some minimum spacing between nodes. The modified algorithm employs a periphery-inward approach. A minimum size for the containment circles is specified for childless nodes and the subsequent

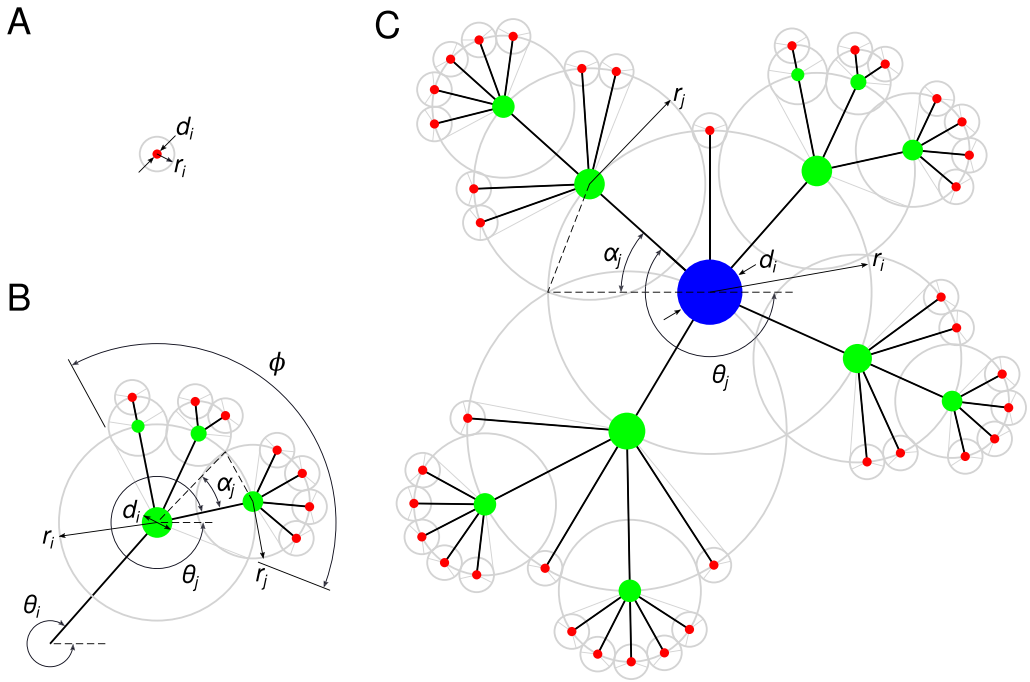

Figure 3. Size and layout parameters for (a) terminal nodes [red], (b) intermediate nodes [green], and (c) the root node [blue]. The genealogical information in (c) is the same as in Figures 1 and 2.

size calculations proceed recursively inward toward the root node. These changes ensure that a minimum spacing is maintained between nodes, ensure equivalent nodes and subtrees are drawn consistently, and pack the nodes together more tightly to reduce the total area of the graph.

As illustrated in Figure 3, there are three types of node scenarios that must be considered. The most common are what we'll refer to as *terminal nodes* because they have no children (Figure 3(a)). Each node has a containment circle, which is used to place it relative to its sibling nodes. In this case, the radius of the containment circle, r_i, is given by

$$r_i = \frac{1}{2}(d_i + g) \tag{1}$$

where d_i is the diameter of the node, and g is a user-specified parameter that prescribes the minimum gap between nodes. The ability to prescribe a minimum gap is an important improvement over the original Pavlo algorithm and eliminates the need to vary the radius of the root containment circle, r, to achieve the desired spacing.

We will assume that the diameter of the node, d_i, is related to the total number of descendants for that node, N_i, by

$$d_i = \sqrt{N_i + 1}. \tag{2}$$

The term *descendant* refers to an individual from any subsequent generation (e.g., children, grandchildren, great-grandchildren) that trace their lineage to node i. By definition, terminal nodes have zero descendants ($N_i = 0$), which means they have a diameter of $d_i = 1$.

Intermediate nodes (Figure 3(b)) are those with both parents and children. Similar to Pavlo et al. (2006), the child nodes are spaced around a containment arc of the circle prescribed by the angle ϕ, a second user-specified parameter. The length of a continuous containment arc, L_i, is given by

$$L_i = \phi r_i \tag{3}$$

but based on a finite number of child nodes, n_i, packed together along the arc, it can be discretized into a series of line segments corresponding to the radii, r_j, of the children's containment circles. The segmental length of the arc is then given by

$$L_i = 2\sum_{j=1}^{n_i} r_j. \tag{4}$$

Using the cosine rule, we know that the radius of the node's containment circle, r_i, is related to the radius of a child's containment circle, r_j, via

$$r_j^2 = r_i^2 + r_i^2 - 2r_i^2\cos\alpha_j = 2r_i^2(1 - \cos\alpha_j) \tag{5}$$

which we can rearrange as

$$\alpha_j = \arccos\left(1 - \frac{1}{2}\frac{r_j^2}{r_i^2}\right). \tag{6}$$

The minimum value of r_i that ensures all children are optimally packed is determined by solving the equation

$$\left|\phi - 2\sum_{j=1}^{n_i} \alpha_j\right| = 0. \tag{7}$$

A numerical approach must be used to solve this equation for r_i as no direct solution exists. We can rearrange Eq. 3 to obtain an initial estimate of $r_i = L_i/\phi$.

There are certain scenarios where it is mathematically possible that a parent node could have a smaller containment circle radius than its descendants, such as when an intermediate node has only one child. This problem only compounds when a chain of nodes with a single child occurs. To avoid this issue, we define a minimum size for the containment circle, $r_{\min}$ given by

$$r_{\min} = g + \frac{1}{2}\left(d_i + \max\left(d_j\right)\right), \tag{8}$$

which ensures the minimum gap size is maintained between the parent and the largest child. We set the containment circle radius, r_i, to be the maximum of the two values given by Eqs. 7 and 8. The arc will be larger than necessary for the children in such cases, such as the intermediate nodes with one and two children shown in Figure 3(b).

The size of the *root node* (Figure 3(c)) is calculated in a similar manner to the intermediate nodes, except that the entire circumference of the containment circle can be used. Therefore, we determine r_i by finding a solution to

$$\left|2\pi - 2\sum_{j=1}^{n_i} \alpha_j\right| = 0. \tag{9}$$

Again, the r_i is set to the maximum of Eqs. 8 and 9 to avoid problems from small numbers of children.

Because determining a node's containment circle depends on all its descendants, the size calculations must be performed beginning with the terminal nodes and ending with the root node. Reversing the order of size calculations from root-outward to periphery-inward results in much more compact graphs and is a major improvement to the original Pavlo algorithm.

2.2.2. *Determining the node orientation*

Once the sizes of the containment circles have been calculated, we must determine the angular position of each child node relative to its parent. Therefore, the angular assignments must begin with the root node and work outward to the terminal nodes.

The general case is the one given by the intermediate nodes (Figure 3(b)), so we will consider it first. We know that half the angle covered by a child is given by Eq. 6. Therefore, the angle θ_j for each child is given by

$$\theta_j = \theta_i - \frac{1}{2}\phi' + \alpha_j + 2\sum_{k=1}^{j-1} \alpha_k, \tag{10}$$

where θ_i is the orientation of the parent node and

$$\phi' = 2\sum_{j=1}^{n_i} \alpha_j \tag{11}$$

to account for the fact that the children may not fill the entire containment arc length.

For the root node, Eq. 10 simplifies to

$$\theta_j = -\frac{1}{2}\phi' + \alpha_j + 2\sum_{k=1}^{j-1}\alpha_k \tag{12}$$

by assuming that $\theta_i = 0$.

2.2.3. Determining node and edge colors

Every child node has had only a single parent in the examples shown thus far; however, it is possible for a doctoral student to have two or more advisors. For this study, only coadvisors already within the academic tree—that is, someone who is a descendant of the root researcher—will be included in the visualizations. Nevertheless, these extra edges in the graph can still cause some confusion.

A feature common to both the original and the modified Pavlo algorithm is that each child has to be assigned to the containment circle of a single parent. When multiple advisors exist in the tree, assignment was made based on the order of recognition in the doctoral dissertation, either on the title page or in the acknowledgements. It was assumed that the advisor mentioned first should be given priority. When the dissertation was unavailable, we relied on information provided by those who responded to our email requests for information.

A second issue is that multiple edges crossing through the graphs may make it difficult to identify who is advising whom. To reduce confusion, each node was assigned a unique color and all edges originating from that node were given the same color. It is assumed that all nodes can be considered on a circle centered at the root node (Figure 4). The radial distance

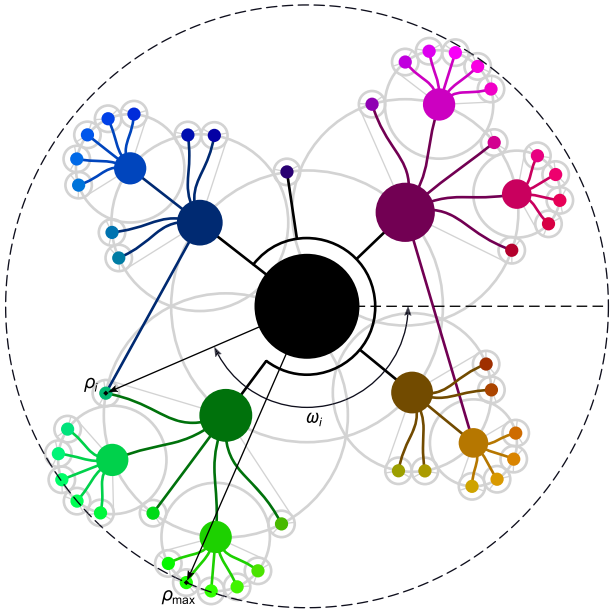

Figure 4. Node colors are assigned by determining the smallest circle, centered at the root node, that contains all the descendants. Locations with an equivalent hue-saturation-value circle are used to calculate individual colors. The nodal information is the same as in Figures 1–3; however, some extra edges have been added to indicate coadvisors.

from the root to the center of the furthest node, $\rho_{\max}$, is treated as the radius of this circle. Colors are then assigned to the nodes based on the angle and radius of an HSL (hue, saturation, lightness) color wheel (HSL and HSV, n.d.). For each node i, with a radial distance ρ_i and an angular position ω_i, the (r, g, b) components for that node are given by

$$(\mathrm{r}_i, \mathrm{g}_i, \mathrm{b}_i) = \begin{cases} (\lambda, \chi, 0), & 0^\circ \le \omega_i \le 60^\circ \\ (\chi, \lambda, 0), & 60^\circ \le \omega_i \le 120^\circ \\ (0, \lambda, \chi), & 120^\circ \le \omega_i \le 180^\circ \\ (0, \chi, \lambda), & 180^\circ \le \omega_i \le 240^\circ \\ (\chi, 0, \lambda), & 240^\circ \le \omega_i \le 300^\circ \\ (\lambda, 0, \chi), & 300^\circ \le \omega_i \le 360^\circ \end{cases} \tag{13}$$

where $\lambda = \rho_i/\rho_{\max}$ and

$$\chi = \lambda\left(1 - \left|\frac{\omega_i}{60^\circ} \bmod 2 - 1\right|\right). \tag{14}$$

The order in which children are drawn is important because it can be used to indicate the order in which they completed their doctoral studies. For intermediate nodes, this means the oldest child (i.e., earliest completion date) is drawn first and subsequent children are drawn, in order, in a clockwise direction. However, because the children of the root node are placed around a circle with no obvious beginning or end, the edge to the oldest child is drawn directly from the root node (Figure 4). All other edges from the root node are drawn from around a central arc to indicate the order of completion. Edges from intermediate nodes to their children are drawn using Bezier curves if that child is on its own ring; however, a straight line is used if the child is on another ring to better distinguish the two scenarios (Figure 4).

2.3. Analyses

In addition to creating the visualizations themselves, five groups of analyses were performed on the five data sets: determining the reductions in graph size achieved by the modified algorithm, investigating the effects of the user-selected parameters on the generated graphs, quantifying researcher fecundity using mentorship metrics, calculating the length of an academic (doctoral) generation, and assessing the completeness of the data available via AcademicTree.

2.3.1. Improved performance of the modified algorithm

One of the main objectives of the modified algorithm was to decrease the space required to display the genealogies. Family trees for the five researchers were generated using both algorithms (see Data Availability for Python implementations). Each graph was rotated to the portrait orientation that used the smallest area as calculated by a rectangular bounding box. The reduction in area was calculated as

$$\Delta A = \left(1 - \frac{A_{MP}}{A_P}\right) \times 100\% \tag{15}$$

where A_P is the rectangular area for the original Pavlo algorithm and A_{MP} is the rectangular area for the modified algorithm. Decreased size is reported as a positive percentage.

It should be noted that the size of the original Pavlo diagram will be determined by the initial radius r, whereas the modified algorithm prescribes a minimum gap between nodes, g. To ensure a fair comparison, the values of r were scaled to ensure an equivalent gap size. The value of ϕ was kept constant for both algorithms.

2.3.2. *Effects of user-selected parameters* (g and ϕ)

The two user-selected parameters (g and ϕ) will control the size, shape, and quality of the graphs generated by the modified algorithm. Therefore, these parameters were investigated independently to understand their effects. Because a constant value of g ($g_0 = 1$) was used throughout, and an optimal value of ϕ (ϕ_0) was determined for each genealogy, these parameters were used to calculate a reference area (A_0) for each graph. The values of g were then varied from 0.5 to 3 and the ratio of the resulting graph area (A) relative to the reference area (A/A_0) was used to calculate the effect on size. Similarly, ϕ was varied between 90° and ϕ_0.

2.3.3. *Mentorship statistics and metrics*

Various summary statistics and metrics were calculated for the five individual researchers. The first involved counting the number of descendants a researcher had in each generation. Those supervised directly by the researcher were the first generation (children), those supervised by the first generation were the second generation (grandchildren), and so on. The total of all descendants was also calculated. When an individual has two or more advisors, it is possible for them to be considered part of multiple generations. As with the visualizations, the primary supervisor was used to determine the generation to which they belonged.

Two researcher fecundity metrics were also calculated and reported, both of which were inspired by bibliometric indices (Hirsch, 2005; Egghe, 2006). The mentoring h-index (h_m) proposed by Rossi et al. (2018) is defined as the number of direct descendants n who themselves have at least n descendants. However, Sanyal et al. (2020) noted that this metric was insensitive to the fact that an individual child may have a large number of descendants. They proposed the mentoring g-index (g_m) which is defined as the largest number n, for which a researcher has n academic children and n^2 grandchildren.

2.3.4. *Academic generation length calculation*

For each researcher, the time between their own graduation and the completion of their first doctoral student was calculated in years. This time to reproduce within academia, an *academic generation*, is the research equivalent of a human generation. Given that other forms of progeny such as master's or postdoctoral students have not been considered in this study, it might more accurately be termed a *doctoral generation*. Nevertheless, this distinction may be unnecessary because those with master's degrees are typically ineligible to advise graduate students, and postdoctoral students, by definition, already have the qualifications necessary.

2.3.5. *Assessment of AcademicTree data*

Online databases are frequently used by researchers interested in understanding academic genealogical patterns. These databases tend to be focused on researchers in specific domains

such as mathematics (Mathematics Genealogy Project, n.d.) or biological anthropology (Barr, Nachman, & Shapiro, n.d.). Although it began as NeuroTree, and was initially focused on researchers in neuroscience (David & Hayden, 2012), AcademicTree.org has since expanded to other areas and has become the most generalized repository available. With almost one million entries and connections, and because researchers routinely use this data for analysis, it is of interest to assess the comprehensiveness of these community-provided data compared to the manually tracked data collected for this project.

AcademicTree data consist of two types of information: a person and a connection. Snapshots of the entire data set (David, 2021)—the most recent of which is from January 14, 2021—are publicly available for processing and analysis (Liénard, Achakulvisut et al., 2018). First, the researchers identified in the five data sets were checked against those in AcademicTree to confirm whether they were present. Second, whether the connection between advisor and student was present in the database was also verified. Only the connections between people in the individual trees, those represented by edges in the visualizations, were evaluated. The values for both people and connections are reported as a percentage of the data collected in this study which are correctly contained in AcademicTree. Incorrect or additional information in AcademicTree was not evaluated, so the reported values represent an upper-bound estimate of the data coverage.

3. RESULTS

Academic genealogies for the five researchers were assembled from a variety of online sources and from information provided by individuals within each tree. Data files containing this information are available for those who wish to reuse them (see Data Availability).

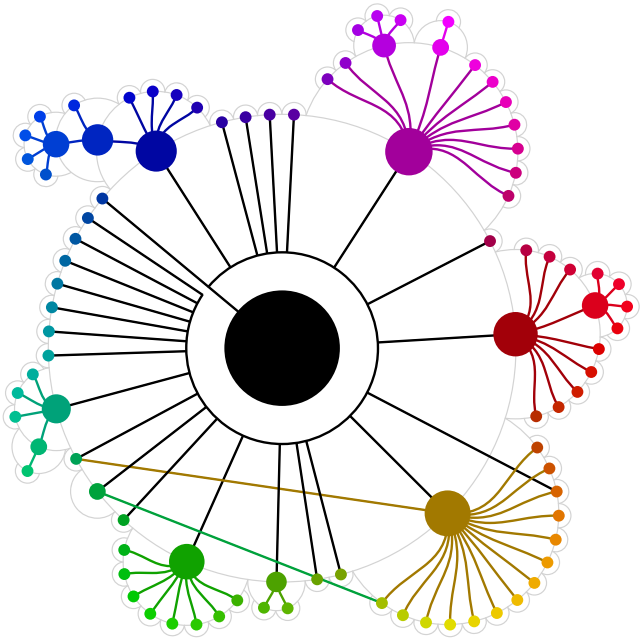

Figure 5. Modified Pavlo diagram showing 93 doctoral descendants of Ron E. Zernicke ($\phi = 175°$).

The modified Pavlo diagrams showing the academic descendants of the five selected researchers are given in Figures 5–9. Siblings will never overlap due to the nature of the algorithm; however, interactions between more distantly related individuals are possible. The largest value of ϕ that eliminated intersection of the containment rings was determined for each graph via trial and error; the specific value used is indicated in the caption. Note that only the outer portions of the containment rings have been drawn, and the ϕ lines have been eliminated altogether, to reduce visual clutter. Note also that the graphs have been rotated into the portrait orientation that makes the most efficient use of the page. Python implementations of the original and modified algorithms are available for reuse (see Data Availability).

The modifications to the algorithm were able to reduce the total area needed to present the genealogies. These reductions were quantified after adjusting the r value of the original Pavlo algorithm to ensure equivalent node size and spacing, and after rotating both genealogies to their optimal portrait orientation. As shown in Figure 10, the sample data used in Figures 1–3 were just one quarter of the size of the original when plotted with the modified algorithm ($\Delta A = 76.5\%$). Even larger reductions were observed for the genealogies of the five

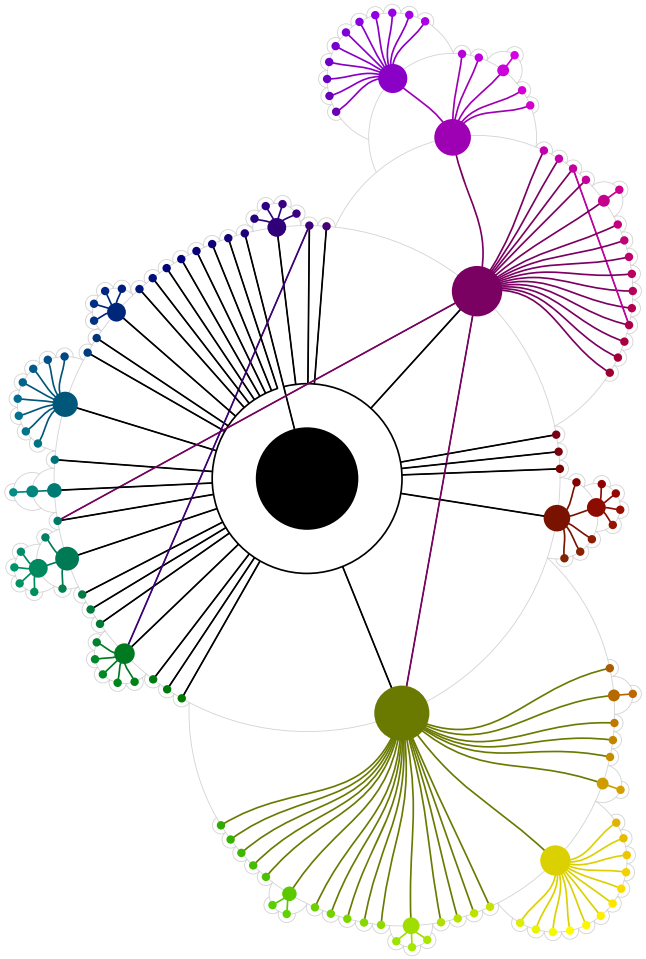

Figure 6. Modified Pavlo diagram showing 147 doctoral descendants of Steven A. Goldstein ($\phi = 165°$).

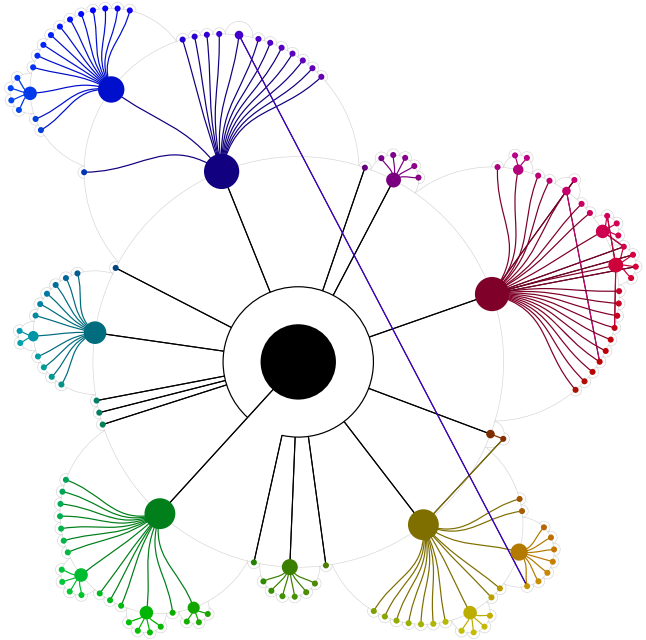

Figure 7. Modified Pavlo diagram showing 150 doctoral descendants of Lawrence E. Thibault (ϕ = 142°).

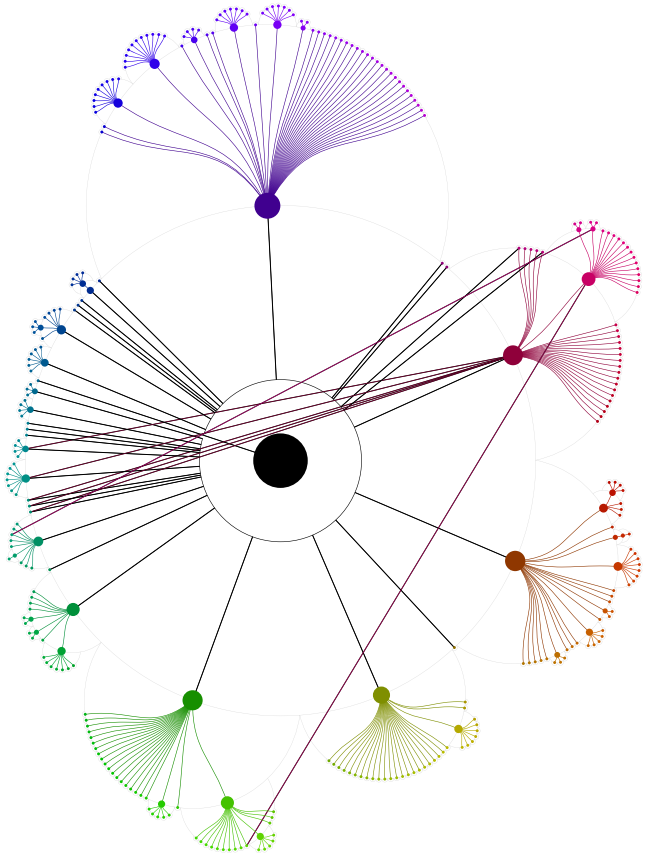

Figure 8. Modified Pavlo diagram showing 343 doctoral descendants of Van C. Mow (ϕ = 128°).

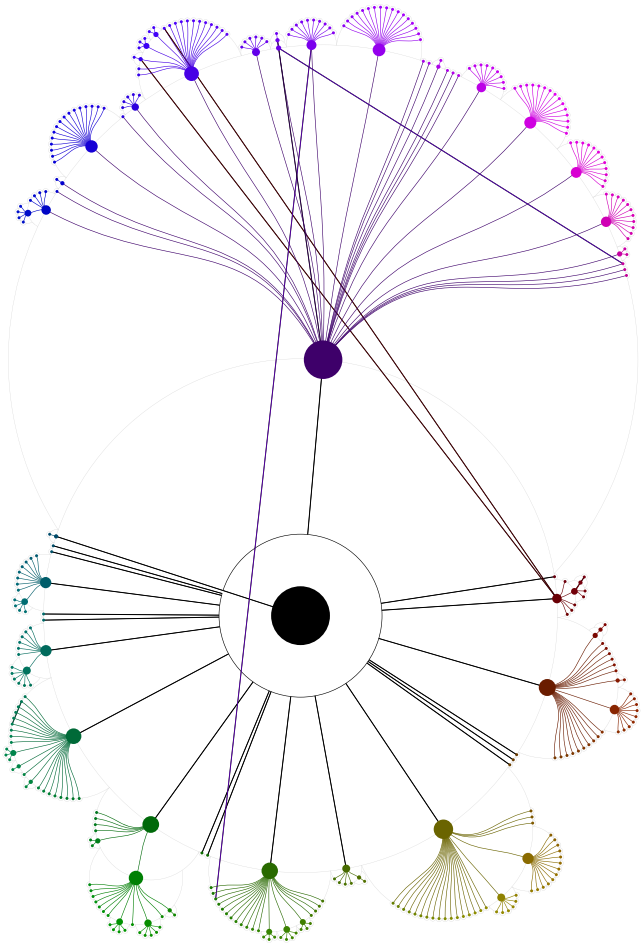

Figure 9. Modified Pavlo diagram showing 384 doctoral descendants of Wilson C. Hayes (ϕ = 140°).

researchers. The Zernicke data was reduced in area by 97.4% (Figure 11). The graphs of the other four researchers had $\Delta A > 99.9\%$ but are not shown due to the very sparse trees produced by the original algorithm.

The areas of the graphs generated by the modified Pavlo algorithm will be affected by the user-selected g parameter. Three values of g are shown in Figure 12 applied to the Zernicke genealogy; the overall layout of the nodes is unchanged by g, and only the scale is affected. The change in size might be expected to follow a trend where $A/A_0 \propto (g/g_0)^2$ as a doubling of g might be expected to double both the width and height of the graph; however, Figure 13 indicates that A/A_0 increases more slowly. This behavior results from the graph-specific path by which the outermost nodes approach the bounding box.

The optimal angle (ϕ_0) was selected iteratively for each graph based on the values at which two or more containment rings began to overlap. The plot in Figure 13 indicates that the area (A/A_0) increases nonlinearly for decreasing values of ϕ. The values of ϕ_0 varied from 128–175° for the five genealogies. Although ϕ_0 tends to decrease with the total number of nodes, the value is dependent on the specific shape of the graph. An alternative to iteratively selecting an optimized ϕ is to select a small angle unlikely to result in collisions of the containment rings, albeit with a resulting increase in area. For example, if a conservative value of ϕ = 120° had been chosen

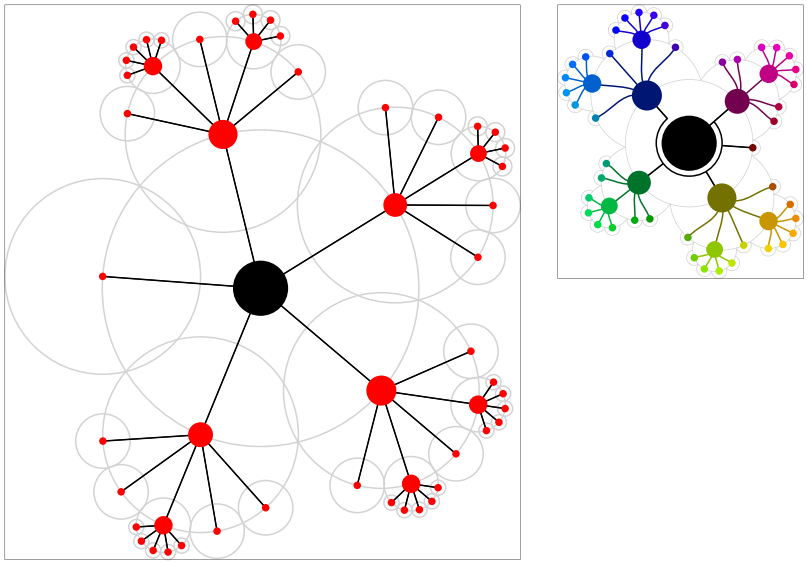

Figure 10. Comparison of the sample family tree shown in Figures 1–3 generated using the original and the modified Pavlo algorithms. The bounding boxes indicate the areas used to compare the relative size of each graph. The same node size, minimum gap size, and $\phi = 160°$ were used in both cases, but the modified algorithm is smaller ($\Delta A = 76.5\%$).

a priori for all graphs, their areas would have increased between 1.25 and 2.25 times (Figure 13). It should also be noted that, because there tends to be one region that determines the ϕ_0, the graphs are relatively insensitive to small deviations from the optimal value. Figure 14 illustrates the resulting changes when adjusting ϕ_0 for the Goldstein genealogy by ±10°.

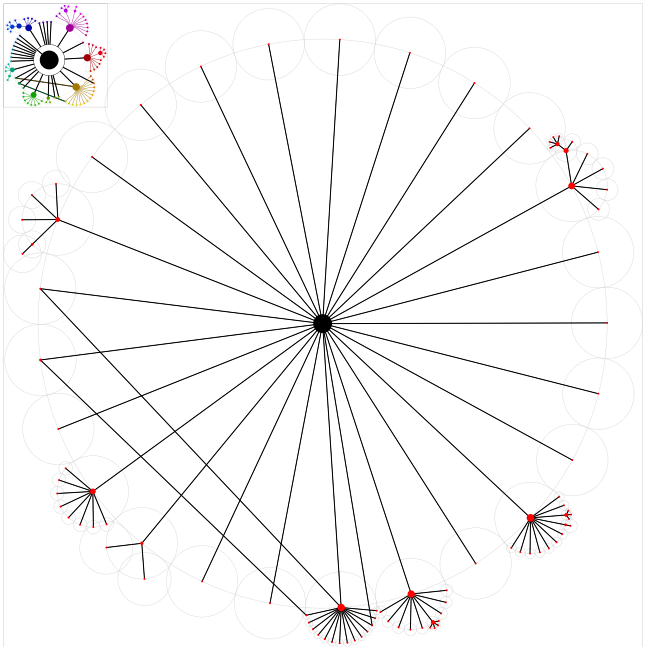

Figure 11. Comparison of the Zernicke family trees generated by the original and the modified Pavlo algorithms (inset). The bounding boxes indicate the areas used to compare the relative size of each graph. The same node size, minimum gap size, and $\phi = 175°$ were used in both cases, but the modified algorithm is much smaller ($\Delta A = 97.4\%$).

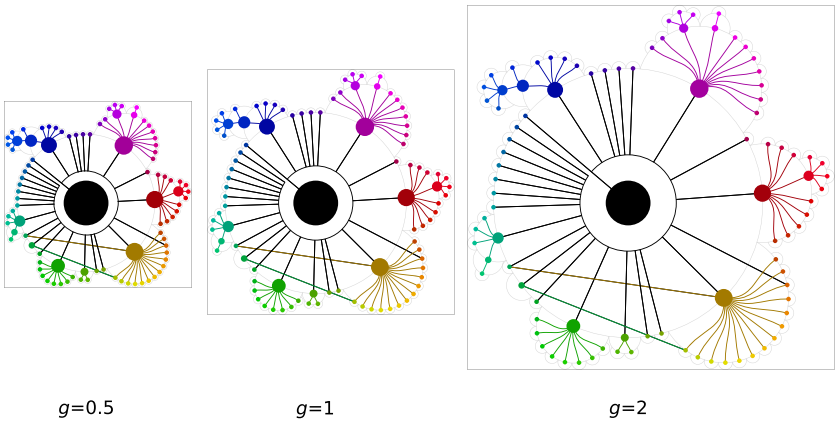

Figure 12. Modified Pavlo diagram showing 93 doctoral descendants of Ron E. Zernicke (ϕ = 175°) for three values of g.

The numbers of descendants in each generation for the five researchers are shown in Table 1. The researchers had between 15 and 32 direct descendants and their total number of descendants ranged from 93 to 384. Some individuals and their descendants appear in two family trees because of cosupervision. Therefore, the 1,118 descendants calculated by adding up the totals of the five researchers consist of only 1,091 *unique descendants* when duplicates are removed. When the five original researchers are included, there are 1,096 unique researchers across the five trees. The h_m-index was 5–7 for each of the researchers, despite the very different genealogical trees. The g_m-index showed more sensitivity and varied in the range 7–13. Based on the analysis of AcademicTree.org data reported by Sanyal et al. (2020), these values place them in the top 1% of researchers with

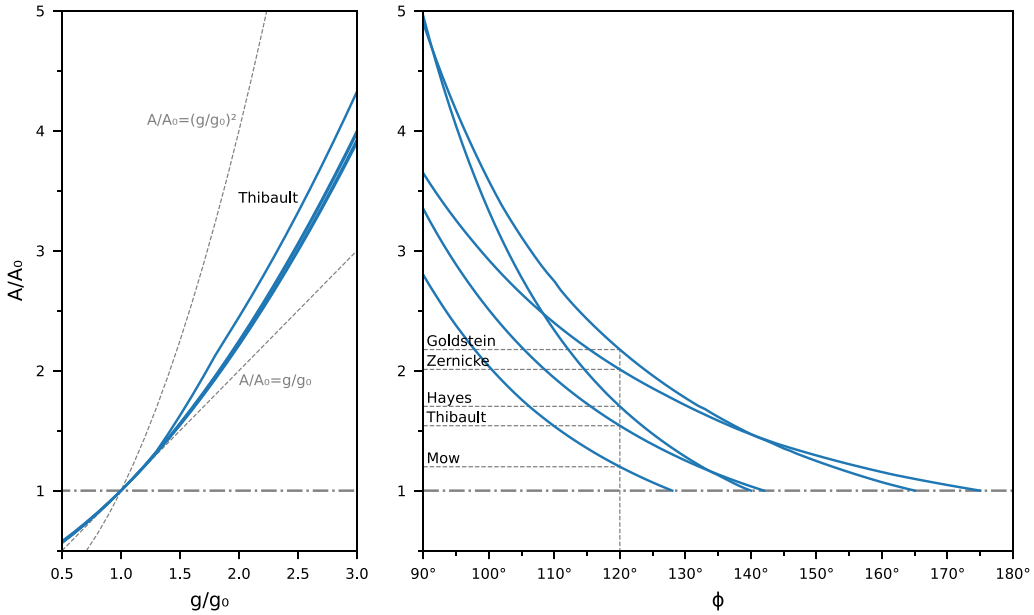

Figure 13. Change in area (A/A_0) for a range of g and ϕ values for the five genealogies. The dot-dashed line in each graph indicates the reference condition for each genealogy.

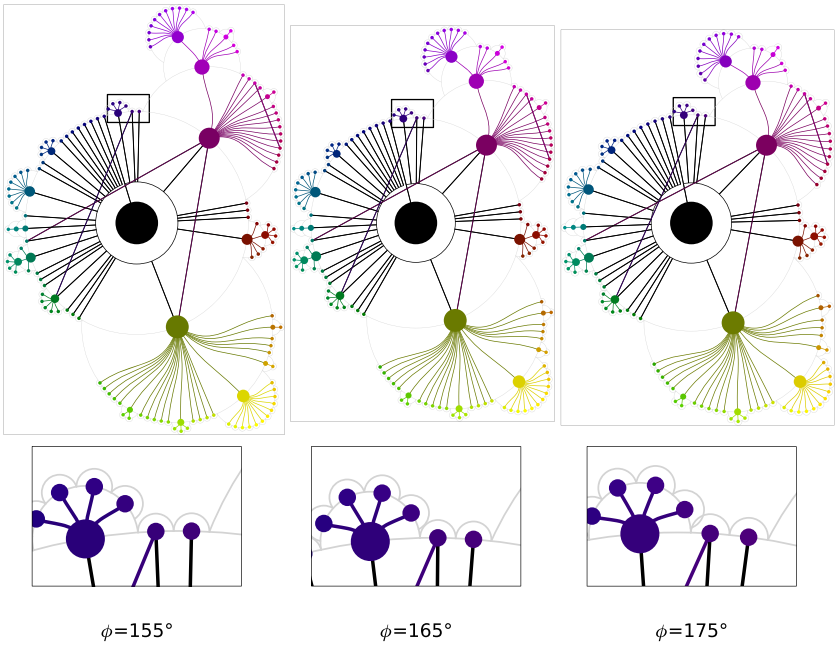

Figure 14. Effects of a ±10° deviation from the optimal ϕ value (ϕ_0 = 165°) on graph size and quality for the 147 doctoral descendants of Steven A. Goldstein.

at least one descendant. This comparison is reported for context but should be interpreted with caution given the differences in the data sets used.

There were 153 individuals among the 1,096 unique researchers (14%) who had graduated at least one PhD student of their own by the end of 2021. The difference (in years) between the graduation date of each advisor and that of their first doctoral student is shown in Figure 15. The distribution is right skewed with an average time of 9.6 years. The median (9 years) and mode (7 and 8 years) were both slightly faster than the average.

Finally, the coverage of the crowd-sourced AcademicTree data, relative to the data collected for this study, are shown in Table 2. The percentage of people in the five individual genealogies varied between 23% and 70%, with 45% of the unique researchers included. The numbers of connections included in AcademicTree was lower: 34% overall, with a

Table 1. Number of descendants for the five researchers, broken down by generation, along with their mentoring metrics (h_m and g_m)

	Descendants by generation					Metrics	
Researcher	1st	2nd	3rd	4th	Total	h_m	g_m
Zernicke	25	53	11	4	93	5	7
Goldstein	32	69	35	12	148	5	8
Thibault	15	83	48	4	150	6	9
Mow	29	176	124	14	343	7	13
Hayes	21	139	201	23	384	7	11

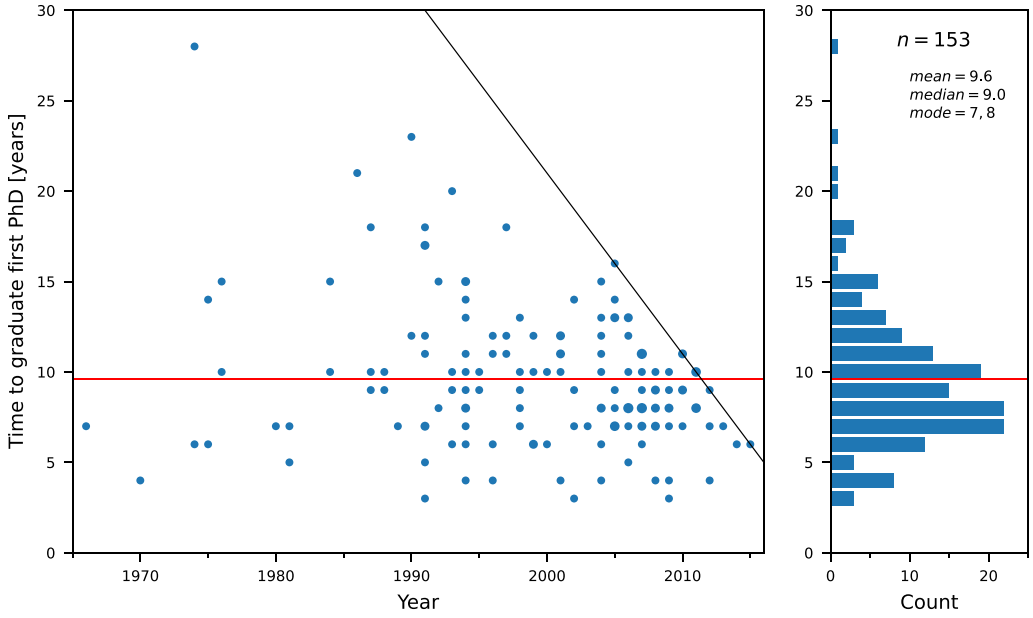

Figure 15. Time between a researcher completing their own PhD and graduating their first doctoral student, based on their year of graduation. The angled black line indicates the maximum possible time for a given year. The red line indicates the average. The frequency distribution and statistics are also shown.

Table 2. Percentage of the people and connections in the current study (CS) found in AcademicTree (AT) for the five trees and when only unique researchers are considered

	People			Connections		
Researcher	AT	CS	%	AT	CS	%
Zernicke	22	94	23.4	16	96	16.7
Goldstein	64	149	43.0	47	152	31.0
Thibault	61	151	40.4	48	160	30.0
Mow	241	344	70.0	202	352	57.4
Hayes	130	385	33.8	89	391	22.8
Unique	496	1,096	45.3	391	1,132	34.3

range of 17–57%. These numbers represent an upper-bound estimate given that we expect that the current data are incomplete and because any erroneous connections in the AcademicTree data set were also not evaluated.

4. DISCUSSION

Circular or radial graphing algorithms result in academic genealogies with smaller aspect ratio layouts, as compared to a standard family tree, for large numbers of descendants. A modified Pavlo layout algorithm has been presented herein that corrects some of the shortcomings of the original. It has been shown to be useful on a range of academic trees with up to four generations and over 380 descendants. The data sets and a reference implementation of the algorithm are available as open data (see Data Availability).

The modified algorithm succeeded in reducing the area occupied by each genealogy. A 77% reduction was obtained for the simple example tree in Figure 10, with reductions of 97% or greater for the genealogies of the five researchers. It could be argued that including the containment rings in the bounding boxes used to calculate area inflated these values in some cases (e.g., Figure 11). Nevertheless, substantive reductions were obtained in this study with the modified algorithm. Other use cases would have to be studied to confirm whether similar performance can be expected; however, the approach appears to be robust.

The algorithm has two user-specified parameters: the minimum gap length (g) and the included angle for the containment arc onto which children are fit (ϕ). Values of $g = 1$ and $\phi = 128–175°$ have been used successfully herein. Because the g term controls the scale of the resulting graph, it can be chosen to alter the spacing between nodes without altering the overall shape (Figure 12). Conversely, the value of ϕ was manually selected to obtain the largest value that ensured that no overlap in the containment rings occurred. Smaller values of ϕ tended to be needed as the number of descendants grew, but the exact value depends on the specific shape of the graph. Based on the range of values determined in the current study, an initial estimate of $\phi = 150°$ is recommended for an iterative search. Alternatively, a constant value of 120° would have yielded satisfactory graphs in all cases, albeit with up to a 2.25× increase in area (Figure 13). Such increases in size may be undesirable in some applications, but the resulting graphs would still be much smaller ($\Delta A > 90\%$) than those produced by the original Pavlo algorithm. The value of ϕ in the current algorithm is both manually selected and constant across all nodes of the graph. Future improvements could be made to the algorithm to either recursively adjust a constant ϕ value or to determine unique ϕ_i values for each containment ring to eliminate overlap and minimize the area used; however, these changes would come with increased computational costs. The current algorithm has shown to be applicable to the unique shape and sizes of academic genealogies, but may also have application to representing a broader range of trees. Different guidelines for ϕ values may be needed in such cases.

The data sets assembled for five biomedical researchers relied on a variety of public and commercial resources. Individual researchers were then contacted to confirm the collected data and gather additional information. Ensuring that the data sets were as current, correct, and comprehensive as possible was important to properly demonstrate the suitability of the algorithm and justified the extra effort involved. Having shown that the algorithm can perform well when handling these large, real-world data sets, it should have no issues with smaller, more sparse graphs. Plus, it is important that the data used exhibit the unique characteristics of academic genealogies. For example, the Hayes tree (Figure 9) has one child (Dennis R. Carter) who himself has a very large number of descendants. Such a feature would not necessarily be found in artificially generated data, or even when using incomplete data. Although every effort was made to ensure the completeness of the data, it must be acknowledged that they are imperfect. Not everyone could be contacted, and not everyone who was contacted replied (the response rate was roughly 45%). Nevertheless, these data are the most exhaustive academic genealogies for these five researchers currently available.

It was interesting to compare the results of the manually traced genealogies created for this study with the crowd-sourced data available. Roughly 45% of the people and 34% of the connections identified were found in the Academic Family Tree (n.d.) data. It should also be noted that this evaluation only considered the people and connections *within* the researchers' genealogies; advisors (and connections to those advisors) outside the tree were not considered or counted, nor were any erroneous connections within the AcademicTree data. Because of this

methodology, and because the current data are known to be incomplete, these percentages represent an upper-bound estimate of the true coverage. Given that the five researchers were within biomechanics and biomedicine, it is unclear how coverage might differ for researchers in other domains. Nevertheless, some incompleteness should be expected and accounted for by researchers performing analyses using the AcademicTree data.

Less complete data are to be expected in crowd-sourced resources, as they rely on continuous participation to provide the necessary information. In this context, it is noteworthy that one researcher had much higher coverage than the other four (Table 2). The reason for this discrepancy is that Dr Mow was awarded the *2017 Alfred R. Shands, Jr., MD Award* by the Orthopaedic Research Society (ORS) for significant contributions to the field. As part of the awards ceremony, some of his descendants presented an academic lineage that they had compiled and uploaded to the AcademicTree website. This detail further underscores both the diligence needed to assemble exhaustive data and the challenge of keeping it updated.

There were 1,096 unique researchers across the five data sets. As of the end of 2021, 153 of them had gone on to have a PhD student of their own (14%) and it took an average time of 9.6 years to do so. The distribution of these times to graduate a first PhD is right skewed. This behavior likely reflects uneven sampling, as the increasing numbers of descendants with time means that graduation dates skew towards the present, and the fact that there is maximum length of time for recent graduates to have graduated their own PhD students (the angled line of Figure 15). Given that most of the researchers studied are in biomechanics or biomedicine, and given that most degrees were earned at institutions in the United States, those durations might not be reflective of other contexts. Nevertheless, it is helpful to think of an academic (or doctoral) generation as being roughly a decade in length.

Two mentorship metrics were calculated for the five researchers. The h_m-index varied from 5–7, whereas the g_m index ranged from 7–13. These results agree with the observations of Sanyal et al. (2020) that the h_m-index was a less sensitive metric. Based on an analysis of AcademicTree by Sanyal et al. (2020), these g_m values would place each of the researchers in elite territory; however, direct comparison between the two is difficult because of differences in the data used. For example, AcademicTree includes all graduate students and postdoctoral researchers in its mentorship data, not just the doctoral students considered herein, which would lead to higher metrics than those reported in the current study. Conversely, the incompleteness of the AcademicTree data already discussed could also skew their metrics downward. The g_m-index offers improved discretization but care is needed when evaluating different researchers to ensure that equitable comparisons are being made.

Finally, it should be recognized that the graphs and indices only capture a particular form of "success" with regard to mentorship. Quantity and quality are orthogonal concepts. These numbers focus on the former and, although it is tempting to view those with smaller numbers as being less successful, it is important to recall the distinction between student- or advisor-centric methods of assessment. A student entering a doctoral program may do so to pursue a career in industrial research, with the goal of launching a start-up company, or to complement future training in other professions such as law or medicine. The extent to which an advisor equips that student to obtain these goals is a different metric of success altogether. Other important definitions of success, such as the way in which an advisor treats their trainees, are equally difficult to quantify. Therefore, although the work presented herein certainly provides insight into mentoring fecundity, it should be balanced by the recognition that "not everything that can be counted counts, and not everything that counts can be counted" (Cameron, 1963).

In conclusion, the current work has proposed a modified Pavlo algorithm for representing compact depictions of academic genealogies. The utility of the approach has been demonstrated using data sets showing the doctoral descendants of five prolific researchers in biomechanics and biomedicine. A number of different analyses have also been performed on the data, which show that the g_m-index is a more sensitive measurement of fecundity, roughly 45% of people and 34% of connections were covered in AcademicTree, and the average time to graduate one's first PhD student was roughly a decade.

ACKNOWLEDGMENTS

The author would like to thank all the respondents for their assistance with, interest in, and enthusiasm for this project. Interacting with you has been a wonderful reminder of the best aspects of academia. The author would also like to acknowledge his father, the keeper of our family tree, from whom he has inherited an interest in genealogies.

COMPETING INTERESTS

The author has no competing interests.

FUNDING INFORMATION

No funding was received for this research.

DATA AVAILABILITY

The data associated with this paper are available for reuse from Borealis (formerly Scholars Portal Dataverse): https://doi.org/10.5683/SP3/MDGUTK. The data for the five genealogies are available as comma-separated value (CSV) and extended markup language (XML) files, while the diagrams themselves are provided as scalable vector graphics (SVG). The Python code used to generate the original and modified Pavlo diagrams is also provided as a reference implementation. All files are available under a Creative Commons CC0 "Public Domain Dedication" license.

REFERENCES

Academic Family Tree. (n.d.). https://academictree.org/.

Arce-Orozco, A., Camacho-Valerio, L., & Madrigal-Quesada, S. (2017). Radial tree in bunches: Optimizing the use of space in the visualization of radial trees. In *2017 International Conference on Information Systems and Computer Science* (pp. 369–374). https://doi.org/10.1109/INCISCOS.2017.32

Barr, W. A., Nachman, B., & Shapiro, L. (n.d.). *The academic phylogeny of biological anthropology*. https://bioanthtree.org.

Bennett, A. F., & Lowe, C. (2005). The academic genealogy of George A. Bartholomew. *Integrative and Comparative Biology*, *45*(2), 231–233. https://doi.org/10.1093/icb/45.2.231, PubMed: 21676766

Cameron, W. B. (1963). *Informal sociology: A casual introduction to sociological*. New York: Random House.

Damaceno, R. J. P., Rossi, L., Mugnaini, R., & Mena-Chalco, J. P. (2019). The Brazilian academic genealogy: Evidence of advisor–advisee relationships through quantitative analysis. *Scientometrics*, *119*(1), 303–333. https://doi.org/10.1007/s11192-019-03023-0

David, S. V. (2021). *Academic Family Tree data export (1.0) [data set]*. https://doi.org/10.5281/zenodo.4441298

David, S. V., & Hayden, B. Y. (2012). Neurotree: A collaborative, graphical database of the academic genealogy of neuroscience. *PLOS One*, *7*(10), e46608. https://doi.org/10.1371/journal.pone.0046608, PubMed: 23071595

Egghe, L. (2006). Theory and practise of the *g*-index. *Scientometrics*, *69*(1), 131–152. https://doi.org/10.1007/s11192-006-0144-7

García-Suaza, A., Otero, J., & Winkelmann, R. (2020). Predicting early career productivity of PhD economists: Does advisor-match matter? *Scientometrics*, *122*(1), 429–449. https://doi.org/10.1007/s11192-019-03277-8

Gaule, P., & Piacentini, M. (2018). An advisor like me? Advisor gender and post-graduate careers in science. *Research Policy*, *47*(4), 805–813. https://doi.org/10.1016/j.respol.2018.02.011

Grivet, S., Auber, D., Domenger, J. P., & Melancon, G. (2006). Bubble tree drawing algorithm. *Computer Vision and Graphics*, 633–641. https://doi.org/10.1007/1-4020-4179-9_91

HSL and HSV. (n.d.). https://en.wikipedia.org/wiki/HSL_and_HSV.
Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(46), 16569–16572. https://doi.org/10.1073/pnas.0507655102, PubMed: 16275915
Huang, G., Li, Y., Tan, X., Tan, Y., & Lu, X. (2020). PLANET: A radial layout algorithm for network visualization. *Physica A*, *539*, 122948. https://doi.org/10.1016/j.physa.2019.122948
Kelley, E. A., & Sussman, R. W. (2007). An academic genealogy on the history of American field primatologists. *American Journal of Physical Anthropology*, *132*(3), 406–425. https://doi.org/10.1002/ajpa.20532, PubMed: 17154360
Levecque, K., Anseel, F., De Beuckelaer, A., Van der Heyden, J., & Gisle, L. (2017). Work organization and mental health problems in PhD students. *Research Policy*, *46*(4), 868–879. https://doi.org/10.1016/j.respol.2017.02.008
Liénard, J. F., Achakulvisut, T., Acuna, D. E., & David, S. V. (2018). Intellectual synthesis in mentorship determines success in academic careers. *Nature Communications*, *9*, 4840. https://doi.org/10.1038/s41467-018-07034-y, PubMed: 30482900
Lv, R., & Chang, H. (2021). Bibliometric-based study of scientist academic genealogy. *Journal of Data and Information Science*, *6*(3), 146–163. https://doi.org/10.2478/jdis-2021-0021
Mackie, S. A., & Bates, G. W. (2019). Contribution of the doctoral education environment to PhD candidates' mental health problems: A scoping review. *Higher Education Research and Development*, *38*(3), 565–578. https://doi.org/10.1080/07294360.2018.1556620
Malmgren, R. D., Ottino, J. M., & Nunes Amaral, L. A. (2010). The role of mentorship in protégé performance. *Nature*, *465*(7298), 622–626. https://doi.org/10.1038/nature09040, PubMed: 20520715
Mathematics Genealogy Project. (n.d.). https://www.mathgenealogy.org/.
Mitchell, M. F. (1992). A descriptive analysis and academic genealogy of major contributors to *JTPE* in the 1980s. *Journal of Teaching in Physical Education*, *11*(4), 426–442. https://doi.org/10.1123/jtpe.11.4.426
Montoye, H. J., & Washburn, R. (1980). *Research Quarterly* contributors: An academic genealogy. *Research Quarterly for Exercise and Sport*, *51*(1), 261–266. https://doi.org/10.1080/02701367.1980.10609287
Pavlo, A., Homan, C., & Schull, J. (2006). A parent-centered radial layout algorithm for interactive graph visualization and animation. *arXiv:cs/0606007*. https://doi.org/10.48550/arXiv.cs/0606007
Reingold, E. M., & Tilford, J. S. (1981). Tidier drawings of trees. *IEEE Transactions on Software Engineering*, *7*(2), 223–228. https://doi.org/10.1109/TSE.1981.234519
Rossi, L., Damaceno, R. J. P., Freire, I. L., Bechara, E. J. H., & Mena-Chalco, J. P. (2018). Topological metrics in academic genealogy graphs. *Journal of Informetrics*, *12*(4), 1042–1058. https://doi.org/10.1016/j.joi.2018.08.004
Russell, T. G., & Sugimoto, C. R. (2009). MPACT family trees: Quantifying academic genealogy in library and information science. *Journal of Education for Library and Information Science*, *50*(4), 248–262.
Rutter, L., VanderPlas, S., Cook, D., & Graham, M. A. (2019). ggenealogy: An R package for visualizing genealogical data. *Journal of Statistical Software*, *89*(13), 1–31. https://doi.org/10.18637/jss.v089.i13
Sanyal, D. K., Dey, S., & Das, P. P. (2020). g_m-index: A new mentorship index for researchers. *Scientometrics*, *123*(1), 71–102. https://doi.org/10.1007/s11192-020-03384-x