Author manuscript, published in IEEE Transactions on Visualization and Computer Graphics, 24(1), pp. 205-214, 2018.
DOI: 10.1109/TVCG.2017.2744080

How Do Ancestral Traits Shape Family Trees over Generations?

Siwei Fu, Hao Dong, Weiwei Cui, Jian Zhao, and Huamin Qu

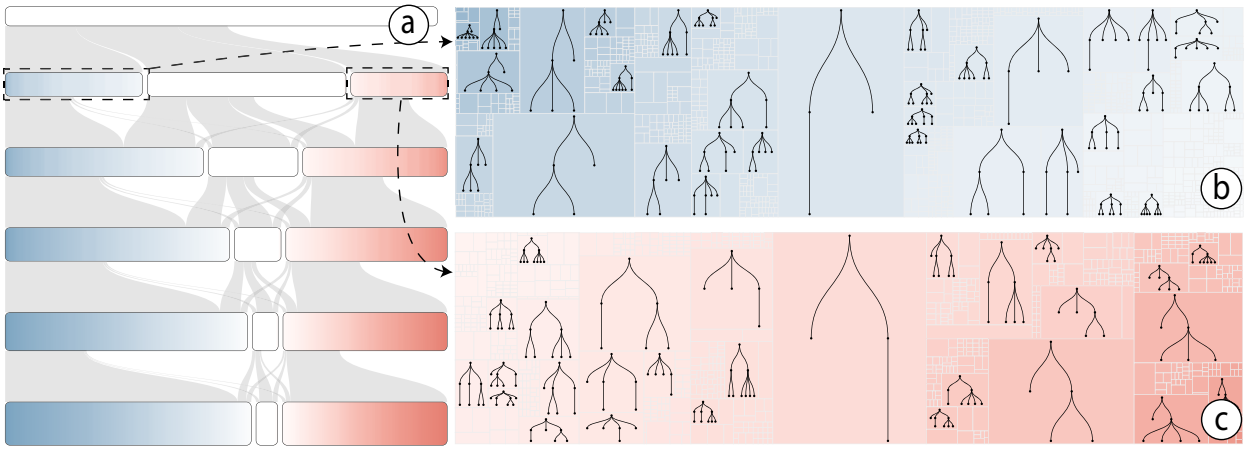

Figure 1. (a) TreeEvo organizes and demonstrates the entire collection of family trees by growth and continuity in a Sankey diagram like visualization. In this example, Sankey nodes in each row represent all trees with the same depth, which are categorized into three groups: left-inclined (blue), balanced (white), and right-inclined (red). (b) After the blue Sankey node is selected, detailed composition of the node, i.e., a set of trees, is displayed in a space-efficient layout. Trees of each specific structure are represented by a rectangle, of which the color indicates inclination and area encodes the number trees. The node-link structure of family trees is displayed if the rectangle is large enough. (c) Family trees included in the red Sankey node, which are right-inclined, are displayed upon selection.

Abstract— Whether and how does the structure of family trees differ by ancestral traits over generations? This is a fundamental question regarding the structural heterogeneity of family trees for the multi-generational transmission research. However, previous work mostly focuses on parent-child scenarios due to the lack of proper tools to handle the complexity of extending the research to multi-generational processes. Through an iterative design study with social scientists and historians, we develop TreeEvo that assists users to generate and test empirical hypotheses for multi-generational research. TreeEvo summarizes and organizes family trees by structural features in a dynamic manner based on a traditional Sankey diagram. A pixel-based technique is further proposed to compactly encode trees with complex structures in each Sankey Node. Detailed information of trees is accessible through a space-efficient visualization with semantic zooming. Moreover, TreeEvo embeds Multinomial Logit Model (MLM) to examine statistical associations between tree structure and ancestral traits. We demonstrate the effectiveness and usefulness of TreeEvo through an in-depth case-study with domain experts using a real-world dataset (containing 54,128 family trees of 126,196 individuals).

Index Terms—Quantitative social science, Design study, Multiple tree visualization, Sankey diagram.

1 Introduction

In social sciences, increasingly available multi-generational datasets enable new research opportunities on the transmission of socio-economic and behavioral traits over generations [25,41,42]. However, due to data complexity and lack of efficient tools, most previous studies only focus on the transmission between parents and children and, only recently, across three generations, leaving one fundamental question unanswered: whether and how does the structure of family trees differ by ancestral traits over generations? Because all traits transmit through family trees, the answer is critical for evaluating and improving the current design of multi-generational transmission research.

While the two-generational analysis is straightforward in research design and estimation, extending it to multi-generational processes is challenging for three primary reasons. First, given limited prior knowledge of the multi-generational socio-economic and demographic processes of human populations, to assume the transmission or influence of individual traits transmitted in any specific form across multiple generations is virtually arbitrary without empirical ground. Incorrect assumptions may introduce biases in research design and statistical estimation. Thus, it is desirable to have a visual analytics tool to generate and verify/reject hypotheses based on newly available data.

Second, most statistics tools used in social science, such as SPSS [5], STATA [1], and R [4], are limited in visual analytics. These tools either report numeric analytical results with limited graphic options, or require advanced programming skills to generate desired diagrams, which hinders experts from understanding and processing the hierarchical data structure of family trees or other social networks.

Third, due to the large data volume and heterogeneity in family tree structures, experts have difficulties understanding the link between the specific structure of each family tree at the micro level and the general topology of a collection of family trees at the aggregated level. Although conventional statistical tools may also assist describing and categorizing family trees by statistics of *known* variables (e.g., number of generations), they take little advantage of structural information in multi-generational data as well as the entire collection of family trees to help discover important *unknown* patterns and factors (e.g., inclination; see Section 3.4).

- *S. Fu and H. Qu are with Hong Kong University of Science and Technology. E-mail: {sfuaa, huamin}@cse.ust.hk.*
- *H. Dong is with Princeton University. E-mail: haodong@princeton.edu.*
- *W. Cui is with Microsoft Research. E-mail: Weiwei.Cui@microsoft.com.*
- *J. Zhao is with FX Palo Alto Laboratory. E-mail: zhao@fxpal.com.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

To address the above challenges, we conduct a design study with social demographers and historians to explore visualization designs for analyzing the associations between individual traits of founding ancestors (e.g., life span) and the structure of family trees in later generations. We derive a set of analytical questions based on discussions and interviews with a group of six experts. Guided by these questions, we develop TreeEvo (**Tree+Evo**lution), an interactive visual analytics system that supports analysis of the association between ancestral traits and family tree structures. Based on a traditional Sankey diagram, TreeEvo employs a dynamic visualization for organizing and demonstrating a large collection of family trees (in our case, 54,128 family trees) by growth and continuity (Figure 1(a)). Moreover, to describe structures of family trees, TreeEvo simplifies trees with a novel pixel-based technique in each Sankey node, and maps structural features to colors. Detailed information of trees can be further explored in a multi-scale space-filling representation (Figure 1(b)). In addition, we employ a Multinomial Logit Model (MLM) [19] to provide quantitative analysis for revealing the association between life history traits, socio-economic status of male founders and the structure of family trees. During the entire study, we use the China Multi-Generational Panel Dataset, Liaoning (see Section 3.2) as a testbed for our design in realistic scenarios and show that our system enables experts to unveil multi-generational implications of reproductive strategies, which has never been studied in relevant domains.

2 Related Work

Our work is related to multi-generational studies done in social sciences, genealogy visualization, and more generally the visualization of multiple trees.

2.1 Multi-generational Studies in Social Science

The call for empirical studies at the individual level to examine socio-economic and demographic processes from a multi-generational perspective is relatively new. Most existing studies on the intergenerational transmission of individual traits and socio-economic status focus on two generations, i.e., parents and children. Only a few recent exceptions have started to study three generations, which try to identify whether grandparental characteristics have a direct influence on grandchild outcomes [31].

This is not surprising since individual-level data that cover more than three generations have only become available in recent decades [15]. A handful of empirical studies moving beyond three generational analysis is starting to take place [25, 42]. That being said, the current analytical framework of multi-generational research relies heavily on existing knowledge and methods of two- and three-generational studies, which are the simplest cases of multi-generational processes. It may ignore the complexity of dynamic and confounding effects through various mechanisms given large kin networks between many kin as well as multiple generations. In other words, existing knowledge of how to conduct multi-generational research is very limited.

Our work, therefore, provides new insights on social science and evolutionary research by investigating the association between life history traits, socio-economic status of the ancestor, and the structure of the family tree. TreeEvo and its resulting new findings help experts generate, verify, and reflect assumptions and methodology to study multi-generational socio-economic and demographic processes in human populations.

2.2 Genealogy Visualization

Various visualization techniques have been proposed to demonstrate the topology and attributes of family trees. We categorize these techniques into three groups according to their representation; node-based, line-based, and matrix-based. Note that the terms (e.g., genealogy, lineage, and pedigree) employed in different studies are often inconsistent but essentially have the same meaning as family tree in our study.

For node-based visualizations, each individual is presented as a node [13, 20, 32, 38, 45]. Using similar representations, existing genealogy software solutions [2, 44] extend the ability of showing individuals with multiple attributes and complex relationships among individuals. Some works use a stacked layout, such as Fan charts [16], to present family trees compactly. Though intuitive, node-based approaches do not scale well to a large number of individuals [7].

Contrary to node-based methods, line-based approaches present individuals as lines, which are usually used to convey a sense of time [7]. For example, Priestley [35] presents individuals as horizontal lines, the length of each line indicates the life span of the corresponding person. However, relationships between individuals are neglected. To address this problem, Genelines [3] adds indents for each line segment to represent descendant relationships. Kim et al. [24] propose TimeNets that indicates marriage and divorce using converging and diverging lines. Similar to node-based approaches, the scalability of line-based approaches is still limited.

Some genealogy visualizations support explorations of large datasets using matrices, where rows are observations and columns are variables [2]. However, this representation is not adequate to show an overview of relationships. More recently, Bezerianos et al. propose GeneaQuilts [7], which displays layered graphs in a more compact manner than traditional matrix representations, and has better scalability compared with node-based and line-based methods.

The above visualizations primarily focus on showing a single family tree, which cannot be used for the domain problems that we focus on. Unlike these techniques, TreeEvo provides an overview of over 54,000 family trees with an extended Sankey diagram, which enables experts to obtain an overall understanding of the growth and continuity of a collection of family trees.

2.3 Multiple Tree Visualization

Inspired by Graham and Kennedy's work [22], we categorize prior arts in multiple tree visualization according to visual presentation, including small multiples, animation, 3-D representations, agglomeration, and atomic representations.

Many works use small multiples, sub-dividing the available screen space into areas and depicting each hierarchical instance with Treemap [39], icicle plot [27], and node-link diagram [10], etc. in each area. For example, to visualize the evolution of hierarchical knowledge domains over time, Kutz et al. [28] present a patent collection with a sequence of Treemaps against a time line. Chi et al. [10] use a collection of Disk Trees to present the changing hierarchical structure of websites over a long period of time. Zhao et al. [48] employ a tabular layout for tree comparison and encode tree similarity as the background of a node-link digram.

Animation and 3-D representations are also popular techniques for demonstrating multiple hierarchies. For example, Card et al. [9] introduce Timetree, which visualizes changes in a tree structure between different time points through animation. From a different perspective, 3-D representations of multiple hierarchies generally present multiple and distinct tree representations in parallel planes [12, 22]. Relationships between the trees are shown by drawing edges [18, 43] or by coloring [46].

However, small multiples, animation and 3-D based approaches do not scale well because each tree presentation requires significant display space [22]. To use screen space effectively, some works use agglomeration [8, 21], which is visual aggregation of multiple trees so that correlating nodes in different trees overlay each other [21]. Others employ a river metaphor. For example, Cui et al. [11] design Roseriver to illustrate the evolution of hierarchical topics across a number of timestamps. Although slightly improved in scalability, the aforementioned approaches may still fail at displaying hundreds of trees or more. To address this issue, some work chose not to show all the trees at sensible level of details when the number of trees grows extremely large. For example, Amenta and Klingner [6] use a scatter plot to visualize a set of trees, where each point represents an individual tree, and distances between points indicate similarity between the associated trees. A detailed view of an individual tree is illustrated after the tree is selected. Unlike the scatter plot approach that displays each tree separately, TreeEvo, while maintaining scalability, aggregates the entire collection of family trees with an extended Sankey diagram to illustrate the general topology of thousands of trees.

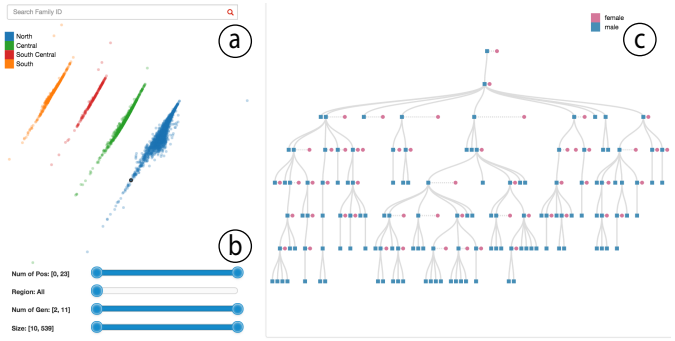

Figure 2. An early prototype that focuses on exploring individual family trees. (a) The collection panel visualizes all family trees as a scatter plot with color mapped to geographic information (blue, green, red, and orange represent north, central, south central, and south, respectively). We then use MDS algorithm [26] to locate each point. Two points close to each other means the two families are similar in three aspects, i.e., location, population and number of generations. (b) Experts can filter trees by different attributes. (c) When a point is selected, detailed structure of the corresponding family tree is presented in the tree panel.

3 Designing TreeEvo

We follow a typical user-centered iterative design framework [37] to develop TreeEvo. In this section, we describe the background of this project and the entire design process.

3.1 Working with Experts

The goal of this study is to support social scientists in conducting multi-generational analysis. We closely collaborate with six domain experts through the design, development, and evaluation of TreeEvo. One expert is a demographer and our internal expert (a co-author of this paper). He has extensive experience working on multi-generational datasets of historical populations from East Asia. From the beginning of our collaboration, he has been actively engaged in this project and wanted to employ visual analytics to address problems and challenges in multi-generational research. The other five are external experts (not co-authors), including one professor and four postgraduate students. The professor is well-established in the field of historical demography and sociology, specializing in multi-generational data and analysis with several major publications. The postgraduate students are all knowledgeable on social science theory, data, and methodology.

The entire design process includes three phases, each consisting of at least one formal interview session with all the experts and several informal discussions with the internal expert. The first phase (Section 3.3 to 3.5) aims to identify important research questions in multi-generational analysis and derive analytical tasks for the visualization. It first involves the development of an early prototype to just "see" the data. This is because multi-generational analysis is a relatively new area and our experts need to flesh out research questions after broadly exploring the dataset. We then work with the experts to identify opportunities and user tasks for the visualization by analyzing their traditional workflow of tackling the problem. The second phase includes the iterative development of TreeEvo based on the experts' feedback. Details of the resulting system are discussed in Section 4. In the third phase, we organize several interview sessions with different experts to evaluate the effectiveness and usefulness of TreeEvo. The results are reported in Section 5.

3.2 Dataset

Owing to the efforts of social scientists and historians in recent decades, large-scale household- and individual-level data that cover populations for many generations have increasingly become available worldwide [15]. The real-world dataset that we use to ground our study is CMGPD-LN [29]. It is transcribed from triennial household registers compiled by the Qing Dynasty (1644-1912) government in Liaoning Province, Northeast China between 1749 and 1909. The dataset includes more than 1.5 million records of over 260,000 individuals.

3.3 Identifying Domain Research Questions

At the beginning of the study, our internal expert commented: *"Although I have been working on the dataset for three years, I do not know how each family tree looks like because no straightforward solution to plot kin networks or family trees is available with STATA [1]."*. This indicated a need for visually exploring the dataset. Therefore, we follow Shneiderman's mantra [40] to visualize the tree collection with a simple visualization shown in Figure 2. Taking the scalability issue into consideration [6], the family tree collection is displayed with an MDS layout [26] in the collection panel (Figure 2(a)). Each point represents a tree, and two points close to each other means that the two families are similar in three aspects, i.e., location, population, and number of generations. With the help of filtering (Figure 2(b)), experts are able to select family trees of interest. Detailed information of a family tree is displayed after the family is selected (Figure 2(c))

Our experts appreciated the early prototype because it gave them the ability to interactively and visually study the dataset. More specifically, they liked the node-link representation of family trees, because it aligns with the convention in social science. However, merely presenting an individual tree is not enough. During the exploration, the experts also triggered, and then raised, questions that cannot be answered by the prototype, such as *"How many family trees grow, or at least continue, at each generation?"* *"How many kinds of family trees exist in the dataset?"* and *"How is the tree structure associated with characteristics of the male founders?"* One expert further explained that existing studies focus on transmission of individual traits and socioeconomic status mostly between parents and children and rarely beyond three generations. However, prior efforts have not taken a large kin network into account, which is *"definitely not an easy task."* Therefore, our experts determined to study associations between the whole family tree structures and characteristics of male founders, aiming at a fundamental understanding of the shaping of multi-generational kin networks.

3.4 Understanding Analysis Workflow

After solidifying the research questions, we carry out discussions with our experts to understand their conventional approaches of solving the new problem in order to discover the challenges and opportunities for visualization. During the discussions, we apply both "talking" and "fly-on-the-wall" protocols [37] to understand how they work in a real-world context. We characterize the following four stages in their statistics-based approaches.

Data Cleaning. Our experts choose to focus on patrilineages—family trees that only consist of a male ancestor and his male decedents. This is because family reproductive strategies in patriarchal societies, such as historical China, have focused on the growth and continuity of male descendants [42]. Thus, the experts first clean and preprocess the original dataset, with the resulting analytical dataset containing 126,169 males. Each male can be considered a founder of one family if he has at least one male offspring. There are 54,128 family trees that consist of at least two generations. In a family tree, each node represents one male family member and each link means a father-son relationship. Some family trees last for 8 generations during the 160 years under observation. The size of family trees varies from 2 to 327, with an average of 6.91.

Hypothesis Generation. Based on the tree structure shown in the prototype, combined with research experience and intuition, the experts find that inclination is a valuable structural feature worth investigating, in addition to common structural features like size and depth. This is because inclination reflects the tendency of unbalanced development across generations. It is semantically meaningful for social scientists because such structural patterns may reveal different reproductive strategies regarding differential parental and kin investment to offspring. Our experts hypothesize associations between the inclination of a family tree and personal traits of the male founder as a first step to study this important yet unanswered question.

Variable Definition. Before statistical analysis, our experts identify and select dependent and independent variables. For illustration, our study currently includes three dependent variables indicating

```
Adjusted predictions                              Number of obs   =      14978
Model VCE     : OIM

Expression    : Pr(group==1), predict(outcome(1))
1._at         : f_bir_age       =          10
2._at         : f_bir_age       =          20
3._at         : f_bir_age       =          30
4._at         : f_bir_age       =          40
5._at         : f_bir_age       =          50
6._at         : f_bir_age       =          60
7._at         : f_bir_age       =          70
```

	Delta-method Margin	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_at						
1	.4284951	.0107412	39.89	0.000	.4074427	.4495475
2	.3595438	.0053892	66.72	0.000	.3489813	.3701064
3	.2831903	.0040682	69.61	0.000	.2752169	.2911638
4	.2083096	.0061568	33.83	0.000	.1962425	.2203767
5	.1436173	.0073902	19.43	0.000	.1291328	.1581018
6	.0937977	.0071908	13.04	0.000	.0797041	.1078914
7	.0588182	.006084	9.67	0.000	.0468937	.0707426

Figure 3. Plain text analytical results of predicted probabilities generated by STATA. More specifically, it shows the relationship between one personal trait (age at first birth, abbreviated as 'f_bir_age') and the predicted probabilities of one kind of family tree (inclination to the left, shown as 'group==1').

structure of family trees: number of generations, number of male members, and inclination. Unlike the first two straightforward features, inclination is newly recognized by experts with the assistance of our early visualization prototype. By aligning offspring in each generation by the birth order from left to right, the inclination of a family tree indicates how unbalanced its branches (i.e., descendant lines) grow. It reflects the cumulative consequence of survival and reproductive advantages enjoyed by first-borns over younger siblings, which is especially true in East Asia [14]. Because no previous empirical research studies the inclination of family trees, our experts define and operationalize the measurement as illustrated in Figure 4. Our experts selected several life history traits and the socio-economic status of male founders as independent variables based on existing literature [23, 42], including 1) age at first marriage, 2) age at first birth, 3) age at last birth, 4) number of sons, 5) life span, and 6) socio-economic status measured by whether they had a salaried official position.

Data Analysis. To find the association between dependent and independent variables, our experts employ a Multinomial Logit Model (MLM), which is used to predict the probabilities of different possible outcomes of a categorically distributed dependent variable, given a set of independent variables [19]. For example, at one point our internal expert expressed interest in *"how personal traits (e.g., age at first birth) of male founders affect the tree structure (e.g., inclination) with three generations,"* so he picked all the qualified family trees (number of generations ≥ 3) and calculated the value of inclination of the first three generations. To meet the discrete input requirement of MLM, he labeled each family tree as "left", "balanced" and "right" based on inclination. Then, he ran MLM with STATA [1]. The outcome was a number of tables showing model statistics and estimated coefficients, which can be transformed to predicted probabilities (Figure 3) and marginal effects. The predicted probabilities represent the relationship between a selected independent variable, e.g., age at first birth, and the probabilities of different family tree groups, e.g., "left", "balanced" and "right". The marginal effects are defined as the slope of the prediction function at a given value of the independent variable [47]. To have a quick and informative understanding of the analysis results, statistical diagrams are usually generated. However, since the STATA graphing option is not efficient, interactive, or convenient, the experts must make significant efforts to draw such diagrams.

3.5 Analytical Tasks

From the above study of our experts' workflow, we consolidate a set of key analytical tasks that are further classified into two categories: *structure identification* and *association analysis*.

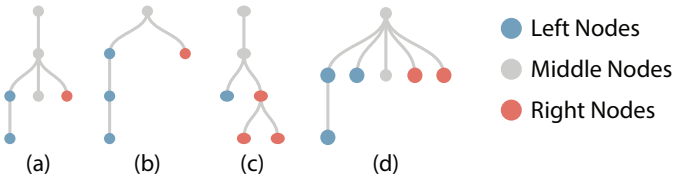

Figure 4. Examples of inclination, which indicates the tendency of unbalanced development of a family tree over generations. After ordering members in each generation by their birth order, the inclination of a root is defined as $(N_{left} - N_{right})/(N_{left} + N_{right})$, where N_{left} and N_{right} are the numbers of left and right nodes of the corresponding tree, respectively. The middle (gray) nodes, including a root node, are split equally to the left and the right. For example, the inclination values for (a), (b), (c) and (d) are $(3.5-2.5)/(3.5+2.5) = 0.167$, $(3.5-1.5)/(3.5+1.5) = 0.4$, $(2-4)/(2+4) = -0.333$, and $(4-3)/(4+3) = 0.143$, respectively.

3.5.1 Structure Identification

In addition to recognizing the structure of each single family tree, experts hope to understand general patterns of family trees in the entire dataset. To meet the needs, we specify our analytical tasks as follows:

T1: *Organizing the entire collection of family trees by depth.* Depth, or the number of generations, indicates the continuity or growth of family trees along with time and has important evolutionary implications to the founder. A macro-level overview of all the tress based on depth helps experts answer questions like *"in this population, how many families continue for at least n generations"*, *"how many families disappear after n generations"*, and *"Among those families lasting n generations, how many of them further continue to $n+1$ generations?"*

T2: *Aggregating family trees by structural features.* As discussed before, the structure of family trees can be measured by inclination, the size of offspring population, etc. Revealing distributions of these features can help experts understand whether (and to what extent) family trees grow in a balanced structure, and the proportion of family trees with different feature values in the whole collection.

T3: *Presenting the structures of family trees in details.* Our experts demand a familiar visualization of trees at lower-level, such as the node-link representation. This helps them better understand the specific structures of family trees with different feature values and examine their distributions across the whole dataset, such as trees sharing the same depth and inclination, or even identical structure.

3.5.2 Association Analysis

To discover association patterns, experts first seek to partition family trees into categorical groups according to specific measures of tree structure. Then, they apply MLM to conduct a multivariate analysis on the association effects. To achieve these goals, we distill the following analytical tasks:

T4: *Flexible partition of family trees based on structural features.* Some measurements have a straightforward definition. For example, when measuring the tendency of unbalanced development of family trees, the inclination to the left, middle, and right is clearly defined. However, partition according to other structural features, such as population, is subject to the decisions of users. For example, experts may like to split family trees into three groups, i.e., small population, median, and large population. Since the specific categorization and definition may vary, the flexibility in partition is an important design requirement from the experts.

T5: *Integration of statistical multivariate analysis with MLM.* As discussed earlier, MLM is one of the most important statistical methods that our experts rely on. To ease the analysis, visual presentation of predicted probabilities and marginal effects, that are usually difficult to generate without advanced skills in STATA, are needed. Based on the results of MLM, experts are able to answer questions like: *"Which life history traits and socio-economic status of the founder influence the chance of his family tree to continue or have certain structural features? And how?"*

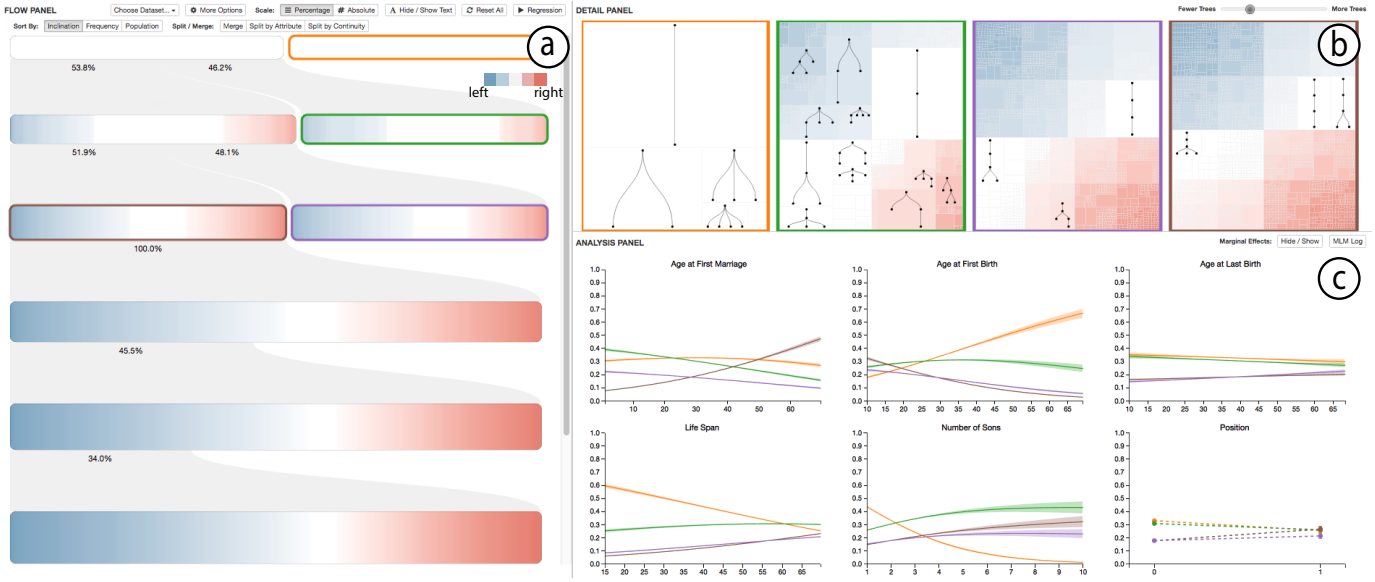

Figure 5. TreeEvo interface: (a) The Flow Panel extends the Sankey diagram [36] to organize the entire collection of family trees. (b) The Detail Panel shows the composition of the selected Sankey nodes using a space-filling visualization. (c) The Analysis Panel reports the results of the Multinomial Logit Model [19] estimations for all trees in the selected Sankey nodes, allowing experts to quantitatively analyze the statistical associations between specific ancestral traits and tree structural patterns.

4 TreeEvo Interface

Guided by aforementioned analytical tasks, we design TreeEvo consisting of three interactively coordinated views, i.e., a Flow Panel, a Detail Panel, and an Analysis Panel (Figure 5).

4.1 Summarizing Entire Tree Collection

As illustrated in Figure 5(a), the Flow Panel is designed to provide an overview of the entire dataset. Here, we discuss the visual design, the corresponding interactions, and design considerations in this panel.

4.1.1 Visual Representation

To support the exploration of the dataset targeting growth and continuity (**T1**), we employ a Sankey diagram [36] to group and align all family trees based on depths. For example, the topmost Sankey node contains all family trees with at least two generations. Sankey nodes at the second level contain trees with at least three generations, in which some trees in the first-level Sankey node are repeated here. For example, suppose that we have four different trees as the input, shown in Figure 6(a). The trees included in each Sankey node are represented in Figure 6(b). In some sense, this approach equates to breaking each tree into sub-trees, which are rooted by the male founder, up to certain depths and then grouping them by depth values. In addition, as shown in Figure 5(a), for each Sankey node, its width encodes the number of trees it contains, and its height is mapped to the depth of those trees. Intuitively, lower and taller Sankey nodes represent deeper trees. Gray flows between Sankey nodes indicate shared family trees, of which the amount is mapped to flow width. The detailed percentages of shared family trees are displayed on each side of the flow.

For our analytical tasks, a traditional Sankey diagram design is limited in two ways. First, elements represented by Sankey nodes are in aggregation. Therefore, many details, such as structural distribution of family trees, are lost (**T2**). Second, a traditional Sankey diagram requires that each Sankey node be defined in advance, and these nodes cannot be re-defined according to different criteria. This poses difficulties for our experts to partition and select tree collections with various structural features with flexibility, as required in **T4**.

To address the aforementioned limitations, we construct each Sankey node in a new way inspired by transformation-based simplifications discussed by Monroe et al. [33]. First, given a collection of family trees with the same depth, we generalize each family tree as a pixel line, a slim line with pixel-level width, and align them side by side as shown in Figure 7(a). Second, we encode the color of each pixel line to structural feature of the family tree. For example, we map the inclination to a diverging color scheme, e.g., blue to red (Figure 7(b)). Finally, we sort these compactly aligned pixel lines according to structural features (Figure 7(c)). Using this coloring-sorting strategy, experts are able to (a) understand the distribution of a collection of family trees based on their structure (**T2**), and (b) partition tree collections according to various structural features with flexibility (**T4**).

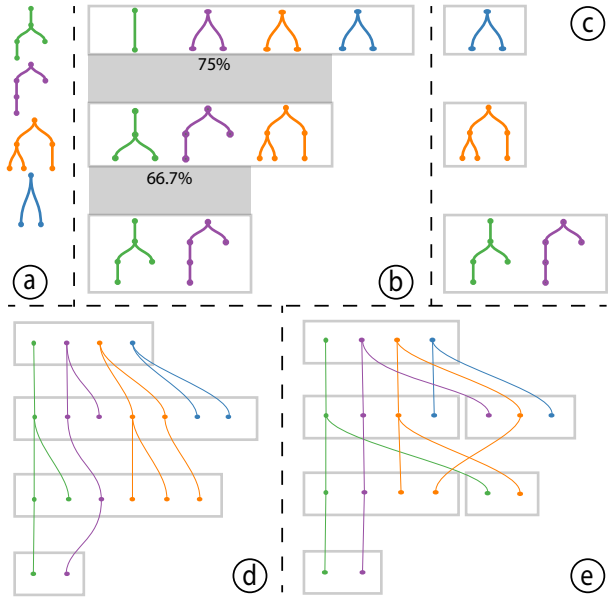

Figure 6. An illustration of different design alternatives for visual tree aggregation. (a) An input of four family trees marked with different colors. (b) A Sankey diagram-based design grouping sub-trees of the input by depth. (c) A bar chart design grouping trees by depth. (d) An alternative Sankey diagram design aligning and grouping tree nodes by depth. (e) An alternative design based on (d) but sorting tree nodes by birth order.

4.1.2 Interactions

The Flow Panel incorporates a number of interactions to facilitate flexible partition in different exploration scenarios.

Selecting. Selection happens when a user wants to 1) examine detailed family trees of a Sankey node, or 2) define family trees of interest before running MLM. When a user clicks a Sankey node, detailed information of the node is displayed in the Detail Panel. Further, multiple selection and undo-selection are supported.

Locating. When a user hovers over a Sankey node, the hovered family tree, represented as a pixel line, becomes gray and the user is able to investigate node-link structure of the tree with a tooltip pane. This allows users to preview, or review, family trees contained in one Sankey node, and recall the definition of the Sankey node.

Splitting. Splitting is the key operation to group family trees according to structural features. With the help of locating, a user is able to split Sankey nodes with flexibility. In addition to free-form splitting, a user can split a Sankey node by "Continuity" (i.e., family trees stopping at some generations) or "Attribute" (i.e., structural features of trees, such as inclination). Locating works closely with splitting to partition a Sankey node into multiple parts. For example, by pressing "Alt" when clicking a Sankey node, a user can set up multiple cutoff lines on a Sankey node. Then, the user is able to split the Sankey node by clicking "Split by Attribute", as illustrated in Figure 8.

Merging. TreeEvo allows users to merge multiple Sankey nodes on the same row. During the interview sessions with our experts, Merging is often used to undo splitting if the partition is not desirable. Further, to provide a short-cut to undo partitioning on all rows, a "Reset All" button is enabled on the top of Flow Panel. In practice, the button is useful for starting a new analytical process after finishing the old one.

Scaling. A user can change the scale of the width of a Sankey diagram to "Absolute" or "Percentage". "Absolute" means that the width of each Sankey node encodes the number of family trees it contains (Figure 11(a)), while "Percentage" unifies the total width of all Sankey nodes in the same row, as shown in Figure 11(b). Thus, we care more about how many family trees contained in one Sankey node account for all trees in the same row.

4.1.3 Discussion on Visual Aggregation of Trees

In multi-generational analysis of family trees, tracing structural changes across multiple generations (i.e., tree depths) is essential, which is enabled by the aforementioned analytical tasks (**T1** and **T2**). That is, each tree should be simplified and organized based on some criteria, such as depth and inclination, in an abstract visual summarization of the multi-generational changes. Moreover, at each generation, the overall distribution of the structure traits of all trees may differ, raising questions such as *"how personal traits of male founders affect the tree structure with five generations?"* To answer these questions and track the changes through generations, a user needs to filter and collect all trees at each generation based on the corresponding criteria (**T4**).

The above considerations lead us to choose the Sankey diagram which is further empowered with flexible partition of Sankey nodes (Figure 6(b)), because its "flow" metaphor naturally reveals the trends of tree structural traits across generations (depths). As described earlier, this design allows our experts to interactively select sub-trees of various depths with ease, and obtain an effective overview of the multi-generational structural changes in the tree collection. Our experts initially found it difficult to comprehend the design because they were unfamiliar with Sankey diagrams. However, they were later able to understand them with the help of an illustration similar to Figure 6(b). In the end, they found it easy to use for defining groups of family trees with different structural features.

Before our final design, we have explored several alternatives in the study. To begin with, we design a traditional bar chart as shown in Figure 6(c). Each bar groups all family trees with certain numbers of maximum generations (depth). For example, the first bar contains trees with two generations, and the second bar, three generations, etc. While the design is easy to understand, it is infeasible to select sub-trees in each bar because the design does not focus on sub-trees of each generation.

Inspired by directed acyclic graph visualization [21], we design another two alternatives (Figure 6(d) and (e)) for tree aggregation. These two approaches align all family trees, and group all male members of the same generation into the one Sankey node and use lines to indicate father-son relationships. The design in Figure 6(e) further sorts tree nodes based on birth order. Both methods have an advantage in selecting family members of interest. However, each Sankey node contains family members instead of family trees, which may cause confusion and hinder experts from selecting desired trees groups. In addition, the scalability is limited if a series of links representing father-son relationships are drawn.

During the design study, we proposed the above design alternatives to our experts with sketches and low-fidelity prototypes. An in-depth user study is needed to further confirm our observations of scalability, learning, and facilitated tasks for each design.

Figure 7. (a) Each family tree is represented by a pixel line, and all pixel lines are aligned adjacently. (b) Pixel lines are colored according to structural features (e.g., inclination). (c) Pixel lines are sorted based on structural feature to help users understand the distribution of the feature.

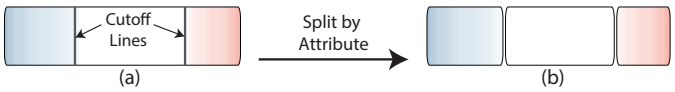

Figure 8. Example of splitting interaction. (a) Users first specify two cutoff lines on a Sankey node. (b) Then, they click the "Split by Attribute" button to generate three separate smaller Sankey nodes.

4.1.4 Design Process of Pixel Lines

We explore the design of displaying distributions of structural features on a Sankey node through an iterative process by working with our experts. Initially, we presented an area chart. Taking inclination as an example (Figure 9(a)), the x-axis represents the value of inclination, from -1 to 1, while the y-axis is the number of family trees. This design is able to show the distribution of structural features in a familiar way. However, it fails to provide enough details. For example, our experts cannot answer questions like *"what portion of family trees have inclination to the left or to the right in this Sankey node"* (**T2**).

To support more details, we employ pixel-based techniques [34]. We have tried a pixel-map based method. As illustrated in Figure 9(b), each pixel, or small rectangle, represents a family tree and all pixels are sorted from top to bottom, from left to right by inclination. This design is able to provide more details compared with the area chart. However, partition may be sometimes undesirable in practice. For example, when partitioning a Sankey node into two parts, i.e., inclination to the right and others, our experts find many errors in the partition. As shown in Figure 9(d), white rectangles represent balanced family trees. However, those with a black border are categorized into the "inclination to the right" group. Compared with the pixel map design, pixel lines (Figure 9(c)) allow experts to split Sankey nodes more precisely (Figure 9(e)). During our interview in the second phase, the experts were able to generate desired partitions with the pixel-line design, and they appreciated the understanding of the distribution of structural features in each Sankey node.

4.2 Displaying Details

When a user selects one, or multiple Sankey nodes, detailed composition of each Sankey node is displayed in a space-filling representation in the Detail Panel (**T3**), as shown in Figure 5(b). When designing the Detail Panel, we find it challenging to provide enough details while avoiding information overload. Based on an observation that many family trees share the same node-link structure, we first group trees with the same structure, and then employ a Treemap algorithm [17] to visualize each group of trees in a compact layout. As shown in Figure 5(b), each rectangle in a Treemap represents a group of family trees with the same structure. The area of each rectangle encodes the number of family trees, and the color is mapped to the value of

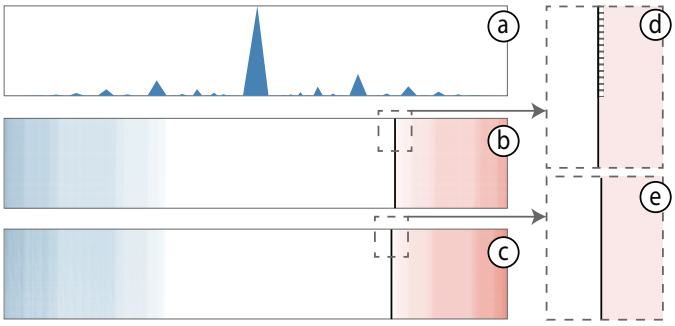

Figure 9. Design alternatives to visualize family trees in a Sankey node. Although area chart (a) may present the distribution of structural features, it does not provide enough details. Partition in pixel maps (b) may cause a number of errors (highlighted with black border in (d)). On the other hand, partition in pixel lines (c) is more precise (no mis-partitioned pixel lines in (e)). The black vertical lines in (d) and (e) indicate cutoff lines.

structural features, which is consistent with the color encoding of pixel lines in the Flow Panel. If a rectangle is large enough and has an adequate aspect-ratio, we overlay the node-link tree structure onto it for users to explore high-frequency trees with ease.

To help users examine infrequent tree structures, which are represented as small rectangles in the Detail Panel, semantic zooming is supported. Specifically, when a user clicks a small rectangle, the rectangle expands to occupy more screen space while other rectangles shrink to a smaller area. The tree structure is then displayed. An animation is played to ease the transition between different visual states of the rectangle. Furthermore, a user can click the expanded rectangle to restore the layout.

We choose Nmap [17] to calculate the space-filling layout because it can generate rectangles with a higher aspect-ratio, which is important for providing adequate space to display the tree structure. To feed a number of groups of trees as the input of Nmap algorithm, we first assign an initial placement to each tree group. That is, we place all groups from left to right in the Detail Panel after sorting them according to structural features, which is consistent with the sorting strategy of pixel lines in the Flow Panel.

4.3 Analyzing with MLM

The Analysis Panel (Figure 5(c)) illustrates analytical results calculated by MLM, i.e., predicted probabilities and marginal effects (**T5**). It coordinates closely with two other panels. After a user selects the Sankey nodes of interest and presses "Regression" in the Flow Panel, six prediction diagrams for six personal traits of male founders (e.g., age at first birth and number of sons) are displayed in the Analysis Panel. To save space, marginal effect diagrams are not displayed by default. A user can choose to show, or hide, marginal effect diagrams by clicking a button in this panel. The color of each line in prediction diagrams and marginal effect diagrams is the same as the highlighting color of the Sankey diagram in the Flow Panel, as well as the border color of rectangle groups in the Detail Panel.

The prediction diagram (Figure 10(a)) presents the relationship between a selected predictor (x-axis) and the predicted probabilities of the different categories (y-axis) [47]. The marginal effect is defined as the slope of the prediction function at a given value of the independent variable. Therefore, marginal effect diagrams (Figure 10(b)) inform us about the change in predicted probabilities due to a change in a particular independent variable [47].

These diagrams help the experts reveal deeper and quantitative insights. For example, Figure 10(a) shows the relationship between a predictor, i.e., the number of sons, and four categories of trees that are encoded in different colors. During an interview of the second phase, our experts found that when a founder has four sons, the probability that the family tree stops at the second generation is about 17.2%, while the probability of stopping at the third generation is about 38.2%.

Independent variables can be either continuous or discrete when running MLM. For example, age is a continuous variable, including

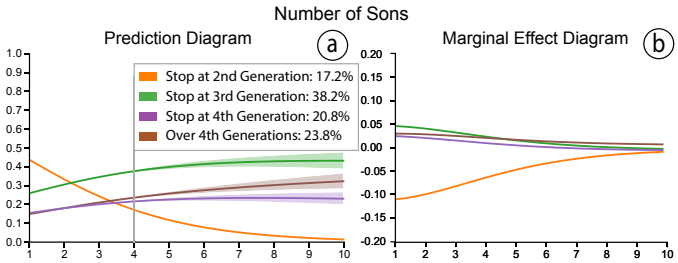

Figure 10. Outcome of analyzing with MLM, i.e., (a) prediction diagram that is the same as the fifth diagram in Figure 5(c), and (b) the corresponding marginal effect diagram.

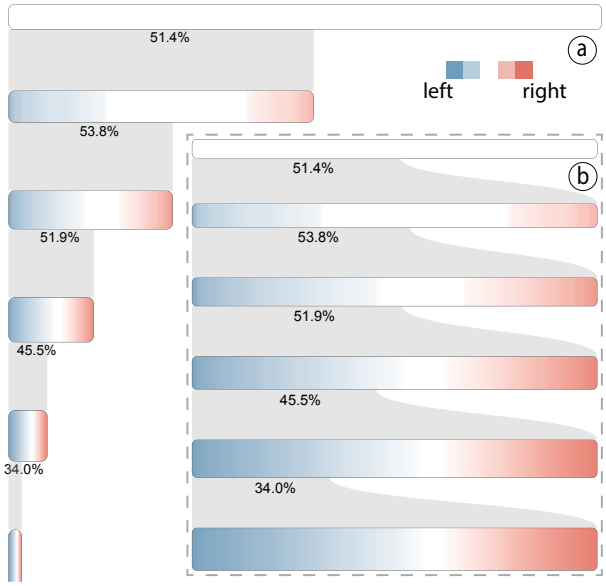

Figure 11. The Flow Panel shows the inclination patterns of all family trees in (a) "absolute" and (b) "percentage" scales, respectively. We can see that white area in both scales decreases when the number of generation increases.

age at first marriage, age at first birth, and age when last observed alive, among others. But the indicator of whether the male founder held a salaried official position is a discrete variable. As a convention, we employ line charts to present continuous variables, and use scatter plots for discrete variables in both diagrams. The confidence interval is important interpreting and understanding results in prediction diagrams, and is displayed by area charts for continuous variable and error bars for discrete variable.

5 Case Study

To study the effectiveness of TreeEvo, we conduct several interview sessions with the experts whom we work with through the design process. Each interview lasted for one hour. We first demonstrated the system for 15 minutes by introducing the design and interaction. We provided a use case sample to our experts to allow them to learn by example. The following 45 minutes were used for free exploration of the CMGPD-LN dataset. Experts were encouraged to think aloud, and speak out about whatever they were thinking and doing during their exploration. We took notes about their feedback at the same time.

In this section, we describe how the experts used TreeEvo to explore and gain insights into the dataset, concluding several cases found by our experts and formulating them into a case study. We denote the internal expert as E0, and the five external experts as E1-5.

5.1 Insights Discovery

5.1.1 Getting the Gist

After loading the data into TreeEvo, E1 first set the scale to "Absolute", and sorted family trees by "Inclination" in the Flow Panel (Figure 11(a)). He immediately observed that the white area in each Sankey node, which refers to the frequency of balanced family trees, decreases each generation. Thus, he wondered whether the proportion of balanced trees also decreases across generations (**T2**). To answer this question, he clicked "Percentage", to standardize the width of Sankey nodes of each generation, as illustrated in Figure 11(b). He observed that the white area decreases when the number of generation increases. This implies that, in order to make the family last many generations, it is probably hard to keep the entire tree structure — or, more specifically, each generation — developed in balance without strategies of differential investment. In each generation, individuals may have different survival and reproduction chances so that not all have an equal number of offspring in the next generation.

To further understand how inclination affects the growth of family trees (**T1**), E1 partitioned each Sankey node into three groups, i.e., inclination to the left, middle, and right, as shown in Figure 1(a). Then, he observed that family trees with inclination to the left and right are more likely to keep the inclination starting with the third generation, as indicated by the gray flow connecting two generations (Figure 1(a)). This tendency suggests that unequal growth in the earlier generations may in fact shape the structure of the family tree and therefore have multi-generational implications for later generations. Further, there are more family trees with inclination to the left than to the right. E1 commented, *"These findings provide empirical evidence in line with first-/early-born favoritism, consistent with Confucian familial values."*

5.1.2 Examining Details

To get an intuitive understanding of how family trees look like in each Sankey node (**T3**), E1 clicked the Sankey node filled with blue gradient color on the second row (Figure 1(a)), where all family trees have inclination to the left. Then, detailed structures of these trees are illustrated in the Detail Panel (Figure 1(b)). To check an extreme case of inclination to the left, he selected the top-left pixel in this panel which has the darkest background color. Then, the rectangle expands to show more details of the tree structure. Similarly, the expert explored the family trees with an extreme inclination to the right. *"I like the smooth animation and interactiveness, which make the exploration easier and more effective."*, E2 added, *"Structure is an abstract term for me, but the system provides a straightforward way of understanding the structure of family trees. It is awesome to see various left- and right-inclined trees with different inclination values."*

5.1.3 Referring to Continuity and Growth

The key problem E0 wanted to know was *"how and to what extent is the structure of the family tree associated with the personal traits of its root"* (**T5**). The expert started by examining the association referring to continuity and growth of family trees. He clicked "Reset All" to clear all partitions set for previous tasks, and split the first three Sankey nodes according to their continuity by generation. After that, he selected four tree sets (Figure 5(a)). Specifically, he selected trees stopping at the second generation (orange border), stopping at the third generation (green border), stopping at the fourth generation (purple border) and growing over four generations (brown border) Then, he pressed "Regression" for MLM estimations of the association between ancestral life history traits and the probability of tree growth outcomes, i.e. the four selected groups.

The results are illustrated in the Analysis Panel. As shown in Figure 5(c), the second diagram shows the influence of the age at first birth (AFB). The orange line represents the probability of family trees stopping at the second generation, which is positively associated with AFB. Lines of other outcomes, on the other hand, have an opposite trend, especially when AFB is greater than 30. The expert explained that, if a male had his first son too late, he may have less chance for many sons and less time to raise sons. He further commented, *"Given high mortality rates in historical populations, his family tree is more likely to end up with two generations — if his son dies — than continue to last for more generations."* The fourth diagram shows the estimated effect of life span of the founder, measured by the age at the last alive observation, the orange line decreased when the last age increased. *"If a male person can live longer, he of course has more chances to reproduce and to take care of the sons. But it seems that such factor extends beyond even two generations (green, purple, and brown lines)."* The expert added, *"Such results exceed my expectations, and I need to conduct further research to understand why that is."*

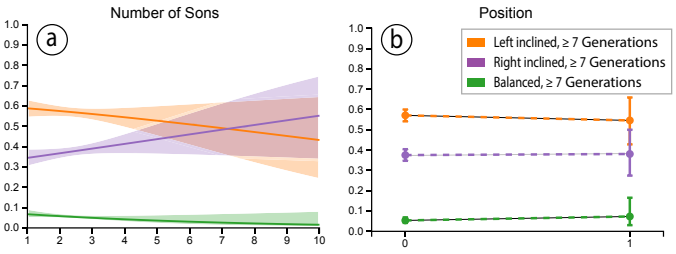

Figure 12. Two predicted probability charts show the influence of two personal traits, i.e., (a) the number of sons and (b) position status, have on the inclination of family trees.

According to the fifth diagram in Figure 5(c) (or Figure 10(a) for clearer illustration) for the effect of number of sons, all green, purple, and brown lines show an increasing trend while the orange line shows a decreasing trend, indicating that as the number of sons grew, the chances of family trees extending beyond two generations increased. *"It indicates that when a male had more than three sons, the probability of his family tree lasting for three or more generations is at least 80%. It is interesting to observe that this is not just driven by three generation family trees. The probability of having a family tree of four or more generations is also non-trivial"*, commented by E0.

5.1.4 Referring to Inclination

Next, with an interest in the influence of birth order preference, E3 decided to investigate the association between life history traits of the founder and the inclination of family trees. To look for long-term influence across multiple generations, he cleared all partitions previously done by E0, and selected the Sankey node at the six level, which contains all family trees with at least seven generations. He sorted Sankey nodes based on inclination and partitioned each Sankey node into three sets, i.e., a set containing all trees with inclination to the left — those early borns (orange border), a set with all balanced trees (green border) and a set containing all trees with inclination to the right — those later borns (purple border) (**T4**).

From a diagram in the Analysis Panel with the title "Number of Sons" (Figure 12(a)), he observed that the predicted probability of balanced trees (green line) was statistically significantly lower than unbalanced trees (orange and purple lines). He inferred that it was hard to keep the tree balanced when it had many (seven in this case) generations. In addition, he observed that when the founder had more than four sons, the orange curve, including the confidence interval area, overlapped with the purple curve and its confidence interval. E3 commented, *"when the initial family is big and probably rich, they may have different strategies and/or easily diversify growth in later generations. A big family is often an indication of rich conditions in the historical context."* Further, E3 added, *"On the contrary, the long-term influence of the birth order preference is stronger when the family is small at the beginning (the founder has no more than four sons)"*.

Then, the diagram titled "Position", as illustrated in Figure 12(b), caught his attention. It showed how a founder's high or low social status, measured by whether the founder held a salaried official position, had an influence on the inclination. One (value of x-axis) means position holding and zero means no holding. The predicted probability of left inclination, indicated by the orange point, is about 58% and greater than others (purple and green points). This difference is especially evident among family trees with the founder of no position holding.

Among family trees of a high-status founder, the difference between left and right inclination is however not statistically significant since their confidence intervals overlap with each other. E3 noted, *"This finding confirms the previous explanation, that is, poor (without official position) families have high probabilities of favoring first-/early-born, while rich (with official position) families care less about it or have more diversity. Maybe poor families tend to concentrate their limited resources to their first-/early-born to ensure the continuity. But rich families could provide enough resources to all sons to maximize the overall chances of lineage continuity."*

5.2 Qualitative Feedback

All experts appreciated the insights found with TreeEvo. E1 mentioned that all these insights are new and have not been discovered before. He pointed out that current system inspired them to pursue two new research directions in multi-generational analysis. First, in addition to tree roots, one could include personal traits of family members into analysis process. Second, tree structure could be considered an independent variable in the MLM. For example, given trees of the first three generations, experts want to know how the structure of ancestral lines have influenced the following tree structure. More encouragingly, E1 particularly valued the visual analytics component and would like to cooperate on a project that he is actively working on. He commented, *"visualization helps us generate hypothesis, and provides an intuitive way of understanding analytical results as well as the dataset."*

Since E0 had tried both a statistics approach (Section 3.4) and a visual analytics approach for association analysis, he compared both approaches and noted, *"I prefer TreeEvo to STATA [1] or R [4] in the analysis. Since structure is an abstract concept, it is hard to understand the statistics results without visually spotting the tree structure. TreeEvo provides a visual way of interpreting the analysis results."* He further added, *"Although we can draw the same (node-link) family tree using R, we will not do it because it is time-consuming and we do not know how effectively it can help the analysis. TreeEvo is really a convenient tool since it not only shows family trees intuitively, but also embeds analytical modules to show association results."*

During the interview, experts also commented on the usability issues of TreeEvo. For example, E4 was curious about the result of merging two Sankey nodes on different rows. She tried but nothing happened because this operation is not allowed. *"I hoped to see a dialog saying that the operation is invalid,"* she commented. She also pointed out that TreeEvo did not show the number of family trees in each generation. We plan to improve these usability issues in the future.

6 Discussion

In this section, we discuss the limitations of TreeEvo and how the visualization design can be applied in other application domains.

6.1 Limitations

Although the case study has demonstrated the effectiveness of TreeEvo in multigenerational analysis of family trees, it still has limitations.

First, the design and visual encoding of the extended Sankey diagram has a steep learning curve. The experts found it time- and attention-consuming to comprehend the visualization at first. However, after getting used to the diagram, our experts spoke highly of the design, and they could partition and select various groups of family trees intuitively and naturally. In future research, we plan to investigate intuitive presentations to 1) lower the learning curve and 2) keep the flexibility and expressiveness as Sankey diagram provides.

Second, although TreeEvo can handle a large number of trees, it may not scale well when the depth of trees increases, even when there is only one family tree with a large number of generations. This may result in a large Sankey diagram with too many levels. Allowing for interactively merging and splitting of multiple generations in the Sankey might solve the problem.

Third, although the design of pixel lines is suitable for presenting single structural feature of family trees, e.g., inclination or population, it is not able to depict multiple structural features at the same time. For example, our experts may want to group family trees based on both inclination and population as well as other features. Employing dimension reduction techniques, such as MDS [26] and t-SNE [30], may resolve this issue by projecting multi-dimensional features onto 1D and visualized by pixel lines. However, loss of information occurs during the dimension reduction.

Fourth, we selectively choose personal traits of male founders to drive our study. However, personal traits of other ancestors, e.g., all individuals in the first two generations, even though they are not the focus of this study, are worth investigating as well. Rich interactions are needed for enabling such investigation. For example, the system could allow users to select personal traits of ancestors from the family tree structure.

6.2 Generalizability of the Design

TreeEvo extends Sankey diagram to organize a tree collection and provide an overview of tree statistics. In addition, trees with complex structures are simplified by pixel lines to reveal structure-level details in each Sankey node. This idea can be widely applied to other datasets (e.g., the history of organismal lineages as they change through time) with large quantity of trees. To be specific, we can employ aggregation methods, e.g., Sankey diagram, to reduces the visual complexity of initially overwhelming phylogenetic trees. At the same time, the pixel-based techniques are introduced to provide fine-grained details about the evolution.

It also worth noting that although TreeEvo is designed for tackling multi-generational analysis in social science, the entire system can be applied for evolutionary studies to examine the transmission of genetic and behavioral traits across generations, as well as for comparison analysis in a large tree collection. For example, a user can select two subsets of trees with different criteria and browse their structural changes across levels to identify differences.

7 Conclusion and Future Work

We have presented a design study exploring the association between life history traits, socio-economic status of male founders and the structures of family trees in the following generations. The results of our study are twofold. First, we characterize tasks in the domain of demography. We help experts identify an unknown structural feature, i.e., inclination, that indicates different reproductive strategies regarding differential parental and kin investments in offspring. Second, we design and develop TreeEvo, a visual analytics system for hypotheses generation and verification about the association. TreeEvo is featured with an enhanced Sankey diagram, which organizes thousands of family trees by growth and continuity, and provides detailed information of each family tree on the Sankey node. Also, it breaks the limit of traditional Sankey diagram, and allows a flexible partition for custom-defined Sankey nodes. We validate our design through one in-depth case study that reveals multi-generational implications of reproductive strategies, which has never been studied before in relevant domains.

There are a number of promising future directions. First, to obtain a deeper understanding of the associations between ancestral traits and tree structures, we plan to a) include personal traits of family members (in addition to tree roots), in Multinomial Logit Models, and b) combine the analysis of actual timeline of family trees and environmental factors. Second, we want to pursue comparisons across multiple datasets. For example, it would be very beneficial to compare how individual traits of ancestors have influence across generations in different countries, such as China, Japan, and the United States. Finally, we wish to evaluate TreeEvo with more experts from demography or related domains to further improve our system.

Acknowledgments

We thank all the domain experts involved in the studies. We also thank Miss Du Rao for helping us recording the video and the anonymous reviewers for their insightful comments and suggestions.

References

[1] Data analysis and statistical software. http://www.stata.com/, 2016.

[2] Genealogy software - genopro. http://www.genopro.com/, 2016.

[3] Genelines timeline software: Create amazing charts. http://progenygenealogy.com/products/timeline-charts.aspx, 2016.

[4] The r project for statistical computing. https://www.r-project.org/, 2016.

[5] Spss, data mining, statistical analysis software, predictive analysis, predictive analytics, decision support systems. www.spss.com, 2016.

[6] N. Amenta and J. Klingner. Case study: visualizing sets of evolutionary trees. In *IEEE Symposium on Information Visualization, 2002. INFOVIS 2002.*, pp. 71–74. IEEE, 2002. doi: 10.1109/infvis.2002.1173150

[7] A. Bezerianos, P. Dragicevic, J. D. Fekete, J. Bae, and B. Watson. Geneaquilts: A system for exploring large genealogies. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1073–1081, Nov. 2010. doi: 10.1109/tvcg.2010.159

[8] M. Burch and S. Diehl. Trees in a treemap: visualizing multiple hierarchies. In *Visualization and Data Analysis 2006*, pp. 600–606. SPIE-Intl Soc Optical Eng, Jan. 2006. doi: 10.1117/12.643272

[9] S. Card, B. Suh, B. Pendleton, J. Heer, and J. Bodnar. Time tree: Exploring time changing hierarchies. In *IEEE Symposium on Visual Analytics Science And Technology*, pp. 3–10. IEEE, Oct. 2006. doi: 10.1109/VAST.2006.261450

[10] E. H. Chi, J. Pitkow, J. Mackinlay, P. Pirolli, R. Gossweiler, and S. K. Card. Visualizing the evolution of web ecologies. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 400–407. Association for Computing Machinery (ACM), 1998. doi: 10.1145/274644.274699

[11] W. Cui, S. Liu, Z. Wu, and H. Wei. How Hierarchical Topics Evolve in Large Text Corpora. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2281–2290, Dec. 2014. doi: 10.1109/TVCG.2014.2346433

[12] A.-S. Dadzie and A. Burger. Providing visualisation support for the analysis of anatomy ontology data. *BMC bioinformatics*, 6(1):1, 2005.

[13] W. de Nooy, A. Mrvar, and V. Batagelj. *Exploratory Social Network Analysis with Pajek*. Cambridge University Press (CUP), 2005. doi: 10.1017/cbo9780511806452

[14] H. Dong. *Patriarchy, Family System and Kin Effects on Individual Demographic Behavior Throughout the Life Course: East Asia*. PhD thesis, Hong Kong University of Science and Technology, 2016.

[15] H. Dong, C. Campbell, S. Kurosu, W. Yang, and J. Z. Lee. New sources for comparative social science: Historical population panel data from east asia. *Demography*, 52(3):1061–1088, May 2015. doi: 10.1007/s13524-015-0397-y

[16] G. M. Draper and R. F. Riesenfeld. Interactive fan charts: A space-saving technique for genealogical graph exploration. In *Proceedings of the 8th Annual Workshop on Technology for Family History and Genealogical Research*, 2008.

[17] F. S. L. G. Duarte, F. Sikansi, F. M. Fatore, S. G. Fadel, and F. V. Paulovich. Nmap: A novel neighborhood preservation space-filling algorithm. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2063–2071, Dec. 2014. doi: 10.1109/TVCG.2014.2346276

[18] T. Dwyer and F. Schreiber. Optimal leaf ordering for two and a half dimensional phylogenetic tree visualisation. In *Proceedings of the 8th International Conference on Information Visualisation*, vol. 35, pp. 109–115. Australian Computer Society, Inc., 2004.

[19] J. Engel. Polytomous logistic regression. *Statistica Neerlandica*, 42(4):233–252, 1988. doi: 10.1111/j.1467-9574.1988.tb01238.x

[20] G. W. Furnas and J. Zacks. Multitrees: enriching and reusing hierarchical structure. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 330–336. ACM, Association for Computing Machinery (ACM), 1994. doi: 10.1145/259963.260396

[21] M. Graham and J. Kennedy. Exploring Multiple Trees through DAG Representations. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1294–1301, Nov. 2007. doi: 10.1109/tvcg.2007.70556

[22] M. Graham and J. Kennedy. A survey of multiple tree visualisation. *Information Visualization*, 9(4):235–252, Dec. 2010. doi: 10.1057/ivs.2009.29

[23] K. Hill and H. Kaplan. Life history traits in humans: Theory and empirical studies. *Annual Review of Anthropology*, 28(1):397–430, Oct. 1999. doi: 10.1146/annurev.anthro.28.1.397

[24] N. W. Kim, S. K. Card, and J. Heer. Tracing genealogical data with timenets. In *Proceedings of the International Conference on Advanced Visual Interfaces*, pp. 241–248. Association for Computing Machinery (ACM), 2010. doi: 10.1145/1842993.1843035

[25] A. Knigge. Beyond the parental generation: The influence of grandfathers and great-grandfathers on status attainment. *Demography*, 53(4):1219–1244, July 2016. doi: 10.1007/s13524-016-0486-6

[26] J. Kruskal and M. Wish. Multidimensional scaling. *Quantitative Applications in the social Sciences Series*, 1978. doi: 10.4135/9781412985130

[27] J. B. Kruskal and J. M. Landwehr. Icicle plots: Better displays for hierarchical clustering. *The American Statistician*, 37(2):162, May 1983. doi: 10.2307/2685881

[28] D. Kutz. Examining the evolution and distribution of patent classifications. In *Proceedings of the 8th International Conference on Information Visualisation*, pp. 983–988. IEEE, 2004. doi: 10.1109/iv.2004.1320261

[29] J. Lee, C. D. Campbell, and S. Chen. China multi-generational panel dataset, liaoning (cmgpd-ln) 1749-1909: User guide. Inter-university Consortium for Political and Social Research, 2010.

[30] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

[31] R. D. Mare. A multigenerational view of inequality. *Demography*, 48(1):1–23, Jan. 2011. doi: 10.1007/s13524-011-0014-7

[32] M. J. McGuffin and R. Balakrishnan. Interactive visualization of genealogical graphs. In *IEEE Symposium on Information Visualization*, pp. 16–23. IEEE, IEEE, 2005. doi: 10.1109/infvis.2005.1532124

[33] M. Monroe, R. Lan, H. Lee, C. Plaisant, and B. Shneiderman. Temporal event sequence simplification. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2227–2236, Dec. 2013. doi: 10.1109/tvcg.2013.200

[34] D. Oelke, H. Janetzko, S. Simon, K. Neuhaus, and D. A. Keim. Visual boosting in pixel-based visualizations. In *Computer Graphics Forum*, vol. 30, pp. 871–880. Wiley-Blackwell, June 2011. doi: 10.1111/j.1467-8659.2011.01936.x

[35] J. Priestley. A chart of biography. London: J. Johnson, St. Paul's Church Yard, 1765.

[36] P. Riehmann, M. Hanfler, and B. Froehlich. Interactive sankey diagrams. In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.*, pp. 233–240. IEEE, Oct. 2005. doi: 10.1109/infvis.2005.1532152

[37] M. Sedlmair, M. Meyer, and T. Munzner. Design study methodology: Reflections from the trenches and the stacks. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2431–2440, Dec. 2012. doi: 10.1109/tvcg.2012.213

[38] P. D. Shaw, M. Graham, J. Kennedy, I. Milne, and D. F. Marshall. Helium: visualization of large scale plant pedigrees. *BMC Bioinformatics*, 15(1):259, 2014. doi: 10.1186/1471-2105-15-259

[39] B. Shneiderman. Tree visualization with tree-maps: 2-d space-filling approach. *ACM Transactions on Graphics*, 11(1):92–99, Jan. 1992. doi: 10.1145/102377.115768

[40] B. Shneiderman. The eyes have it: a task by data type taxonomy for information visualizations. In *Proceedings 1996 IEEE Symposium on Visual Languages*. IEEE, Sept. 1996. doi: 10.1109/VL.1996.545307

[41] X. Song and C. D. Campbell. Genealogical microdata and their significance for social science. *Annual Review of Sociology*, 43, 2017. doi: 10.1146/annurev-soc-073014-112157

[42] X. Song, C. D. Campbell, and J. Z. Lee. Ancestry matters patrilineage growth and extinction. *American Sociological Review*, 80(3):574–602, June 2015. doi: 10.1177/0003122415576516

[43] A. Telea and D. Auber. Code Flows: Visualizing Structural Evolution of Source Code. *Computer Graphics Forum*, 27(3):831–838, May 2008. doi: 10.1111/j.1467-8659.2008.01214.x

[44] R. E. Voorrips, M. C. Bink, and W. E. van de Weg. Pedimap: Software for the visualization of genetic and phenotypic data in pedigrees. *Journal of Heredity*, 103(6):903–907, Oct. 2012. doi: 10.1093/jhered/ess060

[45] J. Wesson, M. C. du Plessis, and C. Oosthuizen. A ZoomTree interface for searching genealogical information. p. 131. Association for Computing Machinery (ACM), New York, New York, USA, 2004. doi: 10.1145/1029949.1029974

[46] R. Wettel and M. Lanza. Visual Exploration of Large-Scale System Evolution. In *Working Conference on Reverse Engineering*, pp. 219–228. IEEE, Oct. 2008. doi: 10.1109/WCRE.2008.55

[47] J. N. Wulff. Interpreting results from the multinomial logit model. *Organizational Research Methods*, 18(2):300–325, Apr. 2015. doi: 10.1177/1094428114560024

[48] J. Zhao, F. Chevalier, C. Collins, and R. Balakrishnan. Facilitating discourse analysis with interactive visualization. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2639–2648, 2012. doi: 10.1109/TVCG.2012.226