GeneaQuilts: A System for Exploring Large Genealogies

Jean-Daniel Fekete, *Member, IEEE*, Anastasia Bezerianos, Pierre Dragicevic, Juhee Lee, Ben Watson, *Member, IEEE*

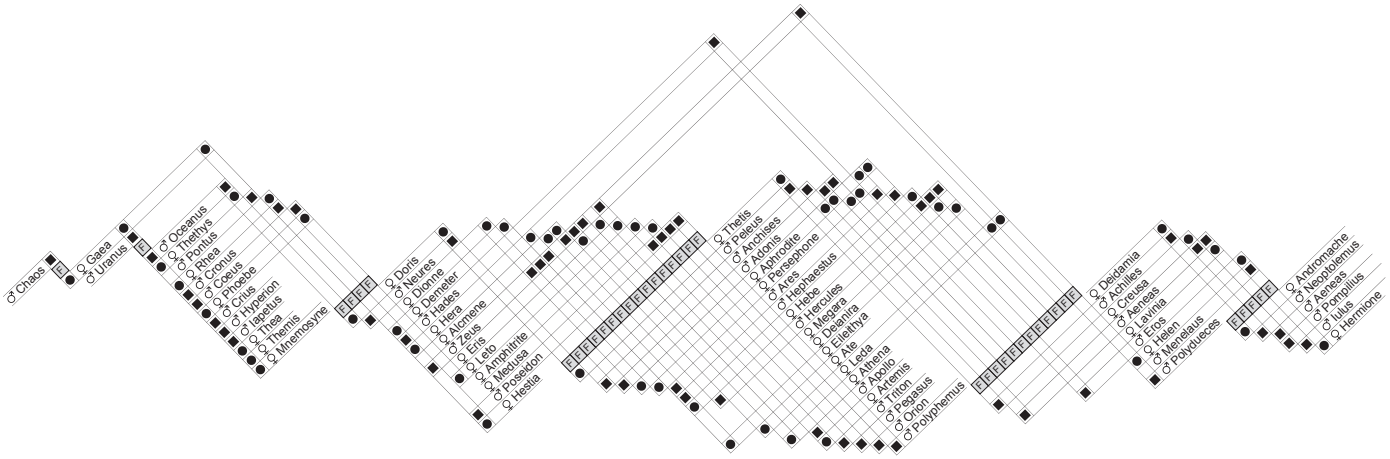


Fig. 1. The genealogy of Greek Gods depicted by GeneaQuilts. Each F icon represents a nuclear family composed of parents (black dots above the icon) and children (black dots below).

Abstract—GeneaQuilts is a new visualization technique for representing large genealogies of up to several thousand individuals. The visualization takes the form of a diagonally-filled matrix, where rows are individuals and columns are nuclear families. After identifying the major tasks performed in genealogical research and the limits of current software, we present an interactive genealogy exploration system based on GeneaQuilts. The system includes an overview, a timeline, search and filtering components, and a new interaction technique called Bring & Slide that allows fluid navigation in very large genealogies. We report on preliminary feedback from domain experts and show how our system supports a number of their tasks.

Index Terms—Genealogy visualization, interaction.

1 Introduction

Genealogy, i.e., the study of family relationships, is an increasingly popular activity pursued by millions of people, ranging from hobbyists to professional researchers [22]. This is reflected in the large number of commercial and free genealogical software packages available [12, 28, 24]. While most of these packages offer excellent support for building and maintaining genealogical databases, their support for visualizing these databases is weak. The most widespread visualizations are based on node-link diagrams, which have been shown to quickly become unreadable as graph size grows [13]. Considering that genealogical databases built by individuals can easily reach thousands of nodes, and those built by organizations tens of thousands, the need for a more scalable solution is clear.

We propose a solution based on a matrix representation, inspired from the Quilts visualization for layered graphs [33]. Quilts (see Figure 2) eliminate confusing link crossings of node-link diagrams, and display layered graphs in a more compact manner than traditional matrix representations. Our GeneaQuilts technique (Figure 1) maps rows to individuals and columns to nuclear families, effectively mapping groups of individuals from the same generation to alternating graph layers. We show how this approach allows us to benefit from all advantages of the original Quilt technique while avoiding its drawbacks.

This article's contribution is threefold: 1) we provide a list of basic genealogical tasks that can be of value to builders of genealogical visualization systems; 2) we introduce a novel visualization technique that can handle large genealogies that could not be depicted earlier and 3) we introduce a novel topology-aware graph navigation technique called Bring & Slide that allows to quickly follow paths in large genealogies. We have integrated our visualization and navigation technique to a genealogy exploration system that can handle complex datasets and has been very positively received by domain experts.

- *Jean-Daniel Fekete is with INRIA in Paris, France, E-mail: jean-daniel.fekete@inria.fr*
- *Anastasia Bezerianos is with École Centrale Paris, E-mail: anastasia.bezerianos@ecp.fr*
- *Pierre Dragicevic is with INRIA in Paris, France, E-mail: dragice@lri.fr*
- *Juhee Bae is with North Carolina State University*
- *Ben Watson is with North Carolina State University, E-mail: bwatson@ncsu.edu*

Manuscript received 31 March 2009; accepted 27 July 2009; posted online 11 October 2009; mailed on 5 October 2009.
For information on obtaining reprints of this article, please send email to: tvcg@computer.org .

2 Background

In this section we explain the data structure of genealogical graphs, discuss genealogical tasks, survey existing genealogy systems and briefly describe the Quilts visualization technique we build upon.

2.1 Data Structures

Genealogies are directed graphs, usually acyclic. There are two standard data structures for genealogical graphs: Ore-graphs and bipartite graphs. Ore-graphs have individual as vertices while directed edges (arcs) represent parent-to-child relationships and undirected edges represent marriages. Bipartite graphs have two types of vertices: individuals and nuclear families – referring to a wife, husband and their biological children. Directed edges map families to their parents, and children to their family.

Graph analysis systems such as Pajek [5] support both types of graphs. The GEDCOM standard format for genealogical data [32] uses bipartite graphs and can associate a large number of attributes to both individuals and families (e.g. birth/death date/place, events such as marriage, divorce, nobility titles, etc.). It is always possible to transform a bipartite graph into an Ore-graph or vice-versa but attributes associated with a family in a bipartite graph are usually lost or duplicated in an Ore-graph.

For representation purposes, most systems also view genealogical graphs as layered graphs, where layers are generations. Strictly speaking, a layered graph is a graph whose nodes are partitioned into sets (layers), and whose edges only run between successive layers. In practice, however, a genealogical graph often contains edges connecting non-successive generations. An approximate partitioning into layers can still be obtained by topologically sorting the genealogical graph.

2.2 Genealogy Tasks

Different tasks are performed by genealogy researchers and enthusiasts, and involve data collecting and recording, source documentation and analysis, and presentation [24]. Although all these tasks are important, we focus on the initial analysis phase, where genealogists attempt to explore their datasets and quickly form or verify hypotheses.

We compiled a list of basic analysis tasks, collected from three extensive interviews with 8 users involved in genealogy research: 3 historians investigating transmission of land ownership and office titles across multiple families in France, 4 anthropologists interested in inter-marriage strategies within small populations/tribes worldwide, and a semi-professional genealogist who investigates family ancestry of individuals or families. We also included analysis tasks supported by commercial genealogy tools and research prototypes. Although our system does not currently support all the described tasks, their enumeration serves as a guide for genealogy visualization systems and for identifying future extensions to our work.

Since a genealogical graph is a special kind of graph, we build our taxonomy on the "Task Taxonomy for Graph Visualization" [18]. The graph objects become *genealogical entities*: individuals, nuclear families, and paths (e.g., ancestors or descendants) in the genealogy graph or sub-graphs. The tasks are categorized as:

Topology-Based (T) tasks where users need to identify global structures or patterns of interest in their data or between specific entities:

T1: Identify one's ancestors (pedigree) and/or descendants [24].

T2: Examine a nuclear family (identify parents, children).

T3: Identify one's extended family (aunts, uncles, nephews, cousins)

T4: Examine the nature of family relations between two or more people in a genealogy. e.g., find if they are connected, if they have common ancestors, find all paths linking two individuals, examine if they are consanguine (by blood) or conjugal (by marriage) relatives, determine the nature of their connection (siblings, n-th degree cousins or uncles, etc.) [20].

T5: Find cases of inter-marriages between family members (both consanguine and conjugal relatives). These connections are often referred to as "rings" [14] within families and may result in pedigree collapse (cases where married couples have common blood ancestors [29]). The types of such inter-marriages are also important (e.g., between parental uncles and nieces, between maternal cousins, etc.) as well as the degree (how many generations are included in the ring).

T6: Identify complex family events, such as divorce, serial and non-serial polygamy, or marriages across generations (generation skipping or merging) [21].

T7: Identify the important individuals in the genealogy (e.g., founder of a dynasty or of the largest lineage, or individual with largest number of children or marriages).

Attribute-Based (A) tasks that involve the exploration of relations and attributes outside of blood and marriage connections:

A1: View detailed information on an entity's attribute, e.g., an individual's birth date, their location or the date of a marriage.

A2: Organize important events for a family (e.g., births, deaths, marriages, etc.) in chronological order [24]. This requires dealing with ambiguous or missing dates that is very common in genealogy data.

A3: Compare attributes of different individuals such as gender, status, etc. Of special interest is how attributes propagate within a family (e.g., inheritance, physical characteristics and genetic diseases) or across families. Commonly found examples include the succession of the title of patriarch within a family [35] or the succession of a political office across families.

A4: Examine the evolution of numerical attributes across time and families. For example investigate how the dowry amount has evolved within a family or compare the division/distribution of inherited land between families across generations.

A5: Explore relationships outside blood and marriage, such as trading partners between families, foster children, family neighbors and friends, etc. [16]. For professional genealogists these relationships can also be crucial links for further research to expand their datasets.

Browsing and Filtering (S) tasks such as:

S1: Search data by person or family name and/or by specific attribute (e.g., date, location, title) [24].

S2: Get details on specific people (e.g., their location, birth-date) or nuclear families (e.g., if the union is legitimate, date of marriage, etc.).

S3: Filter the datasets based on entities of interest (e.g., individuals, bloodlines, etc.) or identified structures (e.g., rings, common ancestors of two or more individuals, etc.).

Overview (O) tasks, related to the users understanding the purpose and possible limitations of a new dataset:

O1: Get a grasp of the focus of the dataset. For example, it is often useful for researchers to understand if the focus of a dataset is a specific individual (e.g., Christ in the Bible genealogy), a family and inter-marriages, or the tracking of someone's paternal lineage (in which case only paternal ancestors will be shown).

O2: Always keep an eye on as much of the dataset as possible for context, without sacrificing the readability of the data.

O3: Identify possible limitations, like missing data from the dataset or uncertainty as to the validity of sources.

2.3 Genealogy systems

For centuries, genealogical relationships have been illustrated in books with hand-crafted charts of a few dozen individuals. Genealogy software can now technically accommodate datasets of hundreds of thousands of individuals[1]. Nevertheless, no software can visualize a large dataset in a legible way. So far, three types of approaches have been used for visualizing genealogies: node-based representations, line-based representations and tabular representations.

2.3.1 Node-Based Representations

For each person in a genealogy there is one tree of descendants and one of ancestors (pedigree). Most commercial software visualize these tree structures using traditional tree diagrams, or offer alternatives such as Fan charts [6] (radial space filling diagrams) or hourglass charts [9] (also called "centrifugal views"), drawing both descendant and ancestor trees. Hourglass charts have many similarities to the Zoomtree interface [34]. All of these visualizations break down quickly as the number of individuals grows, and fail even sooner when they depict

[1]Online databases such as "Roglo" `http://roglo.eu/roglo` reference more than 3 million individuals.

not just consanguine trees (descendants and ancestors) but also the lattice formed by conjugal relationships (marriages).

Visualizing a genealogical graph using a node-link diagram – either from an Ore-graph or from a bipartite graph – usually involves assigning a layer (i.e., a generation) to each individual and trying to minimize the crossings between layers, as in Sugyiama et al.'s algorithm [31]. But even with improved versions of this algorithm [10, 5], large genealogies exhibit too many crossings to be suitable for exploration or presentation. Genealogy systems seldom implement these algorithms and usually resort to unpublished heuristics to layout the graphs, all of which break on special cases (e.g. cycles or multiple marriages on several generations). To solve the problem, they rely on hand-editing the layout, which is impractical for large genealogies.

Dual Trees [19], which are similar to Multi-Trees [9], extend the hourglass chart by offsetting and connecting roots of ancestor and descendant trees, with each root having an hourglass chart. This technique minimizes edge crossings but does not eliminate them, and it still only shows a limited number of nodes on screen. To address this, the authors proposed interaction techniques for expanding or collapsing nodes and transitioning between subsets of the dual trees.

In short, all node-based approaches have serious drawbacks: they do not scale well to large numbers of individuals, they cannot represent large family lattices, they fail to highlight complex relationships (such as polygamy), they do not show temporal attributes (like birth dates) and finally they fail to convey larger context and distant relationships.

2.3.2 Line-Based Representations

Another approach to genealogy visualization represents individuals as lines rather than nodes. For example Bertin [4] presents individuals as line segments and families as points. Each segment has two points, one connecting the individual to her parents, and the other to her children. But multiple marriages are difficult to depict: they require duplicating the lines representing the person for each marriage. P-graphs [35] use a similar representation, with the person's gender indicated by the line orientation (vertical or tilted) and additional notations on the line-segments indicate gender and patriarchal succession. P-graphs are often used for genealogy charts in anthropological literature, as the directions of the lines form interesting patterns when examining intermarriages within a family or clan. Héran uses a similar representation for aggregated populations [15] and argues that gender is the most important feature to show in genealogical representations to clearly distinguish between matriarchal and patriarchal strategies.

Depictions representing individuals as lines are often used to convey a sense of time. For example in [27] individuals are presented as horizontal lines of the life spans of famous people from 1200 B.C to 1750 A.D. However, relationships between individuals are not shown. Genelines [28] extend the timeline visualization by adding connector lines between married individuals and "hang" children from these lines. R. Munroe [25] recently hand-crafted timelines of interactions among movie characters, shown as lines of different color that converge while characters are together. Finally, Genograms in GenoPro [12] extend the number of relations visualized: lines depicting a marriage represent ordinal time; more complex relationships like divorce and re-marriage are depicted through special symbols that overload the visual representation.

2.3.3 Tabular Representations

Finally, most Genealogy systems provide extensive ways to navigate in large datasets by the means of tables: tables of individuals, tables of marriages, tables of places, etc. However, tables alone are poor at showing an overview of the relations between people and at supporting navigation and exploration. They need to be coupled with clear and scalable visualizations.

2.3.4 Structural Analysis

Some genealogy systems provide analysis tools, especially for the purposes of kinship analysis. Ethnographers study the strategies adopted by groups and build models of stable societies based on different kinship systems. Therefore, they develop tools to check their models in specific populations. The main characteristic of these models is based on marriage patterns and ring structures. A ring structure is a cycle in the non-oriented genealogy graph, closed by a marriage. For example, the Bible genealogy shows that Mary and Joseph have a common ancestor: King David. Therefore, there is a closed cycle starting at King David, splitting in two descendant lines, one reaching Joseph, the other one reaching Mary and closing at their marriage.

Several tools have been designed to perform this analysis; the most recent, Puck [1] counts the different types of rings in the genealogy database, and provides algorithms to categorize individuals by attribute or measures and build simplified structural graphs. Puck provides simple graphing capabilities to show distributions or evolutions but relies on Pajek [5] for its graph visualization and analysis capabilities. Puck and Pajek are similar in the style of their interface: complex and feature rich. The communication between them requires good expertise, which makes these tools challenging for historians and less computer-educated users. Furthermore, these tools reveal structural properties but the link to the actual individuals is usually lost.

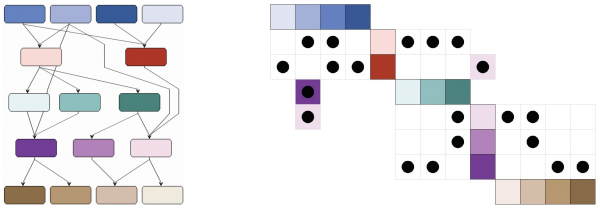


Fig. 2. A node-link depiction of a small layered graph (left) and its quilts depiction (right).

2.4 Quilts

Researchers have advocated matrix-based representations as a scalable alternative to node-link representations [13]. Recently, a variant called Quilts was introduced [33], that can represent layered graphs – as well as "quasi" layered graphs – in a more compact way.

Figure 2 illustrates the original Quilts visualization. The left image shows a node-link diagram of a directed graph where nodes have been assigned a layer (a row). Most edges run between successive layers. The right image shows the corresponding Quilts: nodes are laid out in a zigzagging pattern across the matrix diagonal, as opposed to being on the matrix's borders like in classical matrix representations. The nodes from the top layer (in blue) are laid out horizontally and the nodes from the second layer (in red) are laid out vertically. Links between the two layers are depicted as black dots, forming a *sub-matrix*. To the right of the second (red) layer is the sub-matrix depicting relationships between the second and the third layer (in green).

Problems arise when there are links between two non-successive layers, i.e., *skip links*. For example, it can be seen from the left image that two links go from the 1st to the 4th layer, and one from the 2nd to the 4th layer. Since not all skipped links can be displayed positionally on the Quilts (e.g. 2nd to 4th layer), Quilts appends skip-links to submatrices and uses a color-coding scheme to refer to distant nodes. In Figure 2 for example, two colored dots have been added to the first (blue/red) submatrix to show links from the 1st (blue) to the 4th (purple) layer. Another colored dot has been added to the right to depict the link from the 2nd (red) to the 4th (purple) layer. However, this solution is seriously limited, as it can be difficult or impossible to find the matching color of the destination node, especially in large graphs.

GeneaQuilts builds on Quilts, adapting them to bipartite layered graphs: vertical layers all contain the same type of node (individuals in a genealogy) and horizontal layers contain nodes of a different type (nuclear families in a genealogy). Since links between horizontal and vertical layers are not possible, we re-enable the positional coding of skip links, overcoming Quilts' weak point.

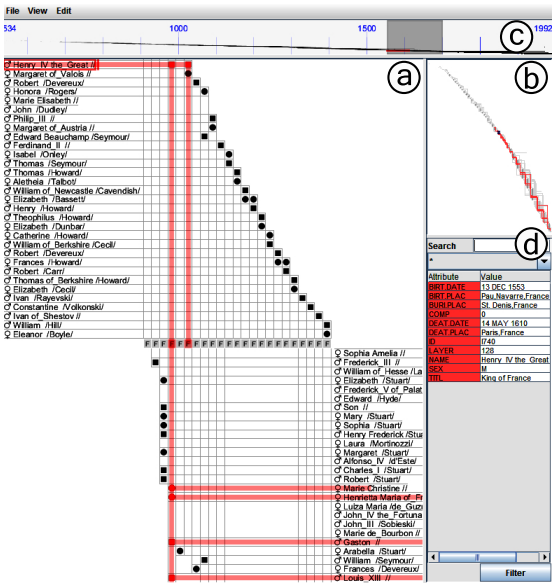


Fig. 3. The GeneaQuilts System showing part of the Royal families.

3 The GeneaQuilts System

The GeneaQuilts system supports genealogy dataset files in several formats, including GEDCOM [32] and has four main visual components (Fig. 3): a) a main visualization window, b) an overview window, c) a timeline and d) a query panel. The interactions have been designed to allow rapid navigation and exploration, while avoiding extensive interface components and menus.

3.1 The GeneaQuilts visualization

The main genealogy window (Fig. 3a) shows a detailed view of the GeneaQuilts visualization. We explain the visualization on a simple example, discuss its benefits, and provide key implementation details.

3.1.1 How to read GeneaQuilts

Figure 4 illustrates the visualization on a simple example. It shows three lists of people, each of which is a generation of individuals. The top-left generation is the oldest and the bottom-right one the youngest. In front of each name is an icon indicating the person's gender. The three icons with an "F" are nuclear families. They are also organized in generations and are laid out horizontally.

Black dots indicate relationships within families. Dots above a family icon point to the parents and dots below point to the children (round dots point to females and square dots point to males). Consider, for example, the rightmost "F" icon in Figure 4: the round dot above indicates Marge is the mother in this nuclear family and the square dot indicates Homer is the father. The three dots below indicate they have two daughters and a son. It is hence easy to focus on a nuclear family/column (**T2**) and identify parents and siblings.

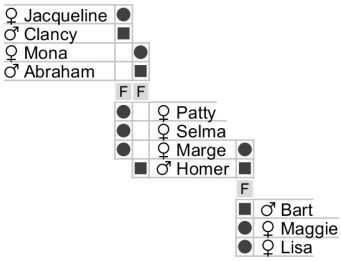


Fig. 4. GeneaQuilts Visualization of the Simpson Family.

It is also possible to focus on individuals (**T1**). Dots to the right of an individual reveal spouses and children, whereas dots to the left show parents and siblings. In Figure 4, the dot to the left of Homer points to a path to his two parents, and the dot to the left of Marge points to a path to her two parents and her two sisters.

Figure 1 contains a more comprehensive example and shows how this representation is more compact than a full relational matrix, as most of the content is close to the diagonal and each individual appears only once. Taking Zeus as an example (4th generation), it can be also seen that multiple marriages are easily visualized, including cross-generational ones (Zeus with Leda). Cross generational births (skip links from families to individuals) are represented the same way. As discussed before, the ability to represent skip-links positionally is a significant improvement over the original Quilts technique.

3.1.2 Layer Assignment

GeneaQuilts relies a lot on a correct assignment of layers. For example, it is not possible to display links from parents to children in the same or a previous generation, so the layer assignment algorithm must ensure children are at lower layers than their parents. We use the algorithm described by Gansner et al. [10] to assign layers and order individuals and families. This algorithm is used by the program dot [11] to layout layered graphs and its result is used directly in GeneaQuilts as it exactly optimizes our layout.

First, it allocates layers guaranteeing generational consistency if the graph is acyclic, and practically minimizes the number of reversed arcs when the graph has cycles. It also minimizes the sum of links length so that the vertical layout is as packed as possible.

Ordering the individuals and families is done by trying to optimize the bandwidth of the matrix so that connected items remain close from each other and close to the diagonal. There are several heuristics to perform this optimization; the most popular published by Siirtola & Mäkinen [30] uses the "barycenter heuristic", a well-known method that is implemented by dot in a slightly better way.

Therefore, our layout algorithm translates a genealogy graph in the dot format, runs the dot program to compute the layout as (X,Y) vertex positions, and assigns generations according to the Y positions and orders in each generation according to the X positions.

Using this method, it is easy to identify interesting families with many children. This basic visualization can display without any additional embellishment complex relationships (**T6**) such as polygamy (more than two parents in a nuclear family/column) or serial polygamy (membership in more than one nuclear families/columns). Moreover, it is easy to identify cross-generation marriages, as the resulting nuclear families tend to extend beyond the bounds of the diagonal representation (e.g. Zeus in Figure 1).

Finally, the current layout does not directly take into account dates (e.g. birth, marriage, death), but when dealing with genealogies made of several disconnected components, we try to align the layers according to dates after they have been assigned by dot. However, very few genealogies have dates at all so, in the worst case, the undated components are positioned to the right end of the GeneaQuilts visualization.

3.2 Selection

Because focusing on an entity or a set of entities is important (**T1-T3**), GeneaQuilts provides rich click & drag selection capabilities. All blood paths related to selected entities (ancestors and descendants) are highlighted. Moreover, details for the entity are shown on the Query panel in the corresponding color (**A1**) (Fig. 3). As genealogists are often interested in one of the ancestor/descendant trees, clicking on a selection toggles between four modes: highlighting the whole blood line, highlighting only ancestors, only descendants, or neither.

Since comparing the influence of individuals is often of interest to genealogists (**T7**), we implemented *selection dragging*: dragging a selection up/down over other individuals changes the current selection and updates the associated trees. Thus with a simple drag gesture, the user can quickly view and compare the trees of different individuals.

Finally, users can also perform multiple selections. Each new selection and its trees get assigned a new color, and details on all the

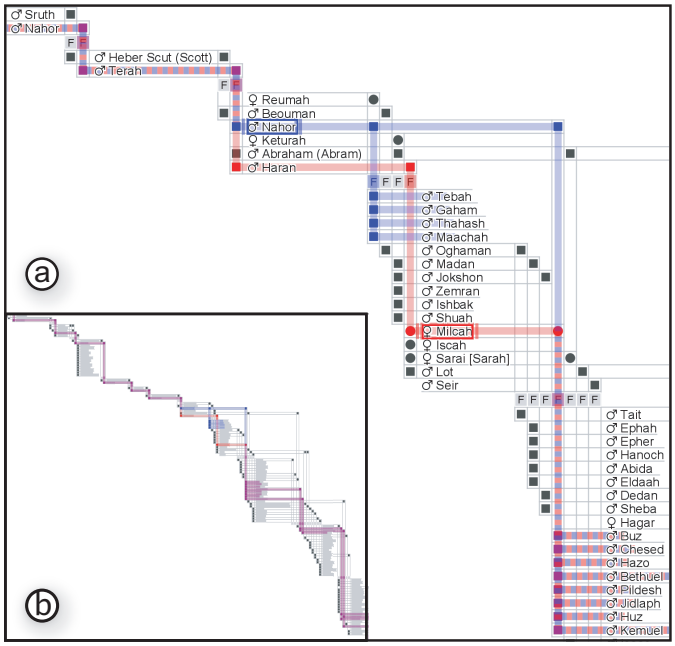


Fig. 5. Multiple selections in the Bible. (a) Selecting the Milcah's bloodline in red and her husband's in blue reveals their common descendants, but also a close common ancestor (Terah) shown by the two lines blended. (b) The bloodlines blending is also visible in the overview.

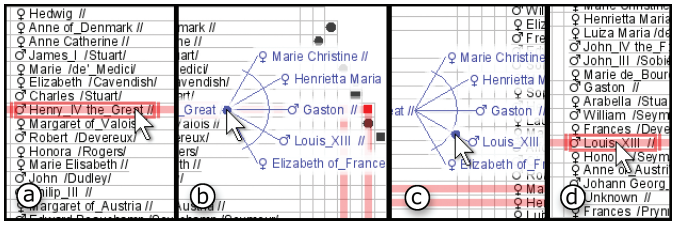


Fig. 6. Bring & Slide from Henry IV to his son Louis XIII in the European Royal Family genealogy.

selections appear in the Query panel with the corresponding color. In a zoomed-out view, the colors of multiple highlighted paths are alpha-blended where they overlap (Fig. 5b). In a zoomed-in view, we use a dashed pattern instead (Fig. 5a). This eliminates ambiguities (e.g., blue and red would give purple, which could be the color of another selection) and provides all the information to understand the provenance of the many paths that are crossing.

The color blending feature can help identify complex relationships (**T4**). For example by selecting an individual's mother and father in two different colors, we can immediately see paternal and maternal grand-parents, cousins, uncles and aunts. By selecting two arbitrary individuals we can immediately see if they have common ancestors or descendants by investigating locations where their tree paths cross and change color. Finally, color blending can be used to identify consanguineous marriages (**T5**) or membership to specific tribes.

3.3 Bring & Slide

Users can pan and zoom directly on the visualization. But since our system is built for large scale genealogies, panning across generations (layers) can be cumbersome, especially since researchers often want to only follow specific paths in the genealogy. To this end we have provided a novel navigation technique called *Bring & Slide* that enables fluid navigation through ancestors and descendants (**T1**). This technique is a fusion of the Bring & Go and Link Sliding navigation techniques for node-link diagrams [23].

If the user selects an individual and drags to the right, names of all descendants appear as proxies to the right of the mouse cursor (Fig. 6). As the user drags towards one of the proxies, the view animates and pans under the proxies so that the descendant of interest is eventually brought under the mouse cursor when the proxy is reached. This drag gesture requires a fixed distance to pan to the destination (50 pixels, more for very large families to avoid siblings from occluding each other), independent from its distance in the visualization.

Once the destination is reached, it becomes the new selection. If the user continues dragging to the right, the proxies of the new selection's descendants appear and the user can navigate further down the bloodline. Similarly, by selecting an individual and dragging left, the proxies of her parents appear to the left of the mouse cursor. So the user can in a single drag navigate quickly across large portions of the genealogy, possibly going back-and-forth in the bloodline.

3.4 Overview

In order for users to always maintain the context of their focus (**O2**), an overview window presents a zoomed-out view of the entire dataset (Fig. 3b). The region in focus is indicated in the overview by a "panner" in the form of a semi-translucent focus rectangle.

The user may drag the panner to quickly refocus the main window. Since the GeneaQuilts representation has an elongated shape, we compute a spline that approximates this shape and restrict the dragging of the focus rectangle along this spline. An automatic zoom-to-fit feature is provided: as the user drags, the focus rectangle adjusts in size to fit vertically all elements above and below the spline.

The overview clearly shows the shape of the current dataset and thus gives an indication of its coverage (**O1**). For example a fanning-out dataset indicates a better coverage on the descendants of a family (e.g. the European royal family dataset), a fanning-in may indicate coverage aimed at a specific person (e.g. Jesus in the Bible genealogy); a fairly constant width may represent a dataset of parental lineage or a title passed through generations covering only the title bearers. Moreover, the overview shape helps users to identify interesting patterns at a glance (**T6**), such as dense generations (long layers) or generation skips (matrix links that do not closely follow the main diagonal).

Selections and their feedback, as well as highlighted search results (Sec. 3.5.1) are also visible on the overview window with guaranteed visibility [26]. Thus selection dragging in the main window immediately shows the influence of different individuals on the whole dataset as their entire bloodlines get highlighted in quick succession (**T7**). In addition, viewing the bloodlines of individuals on the overview provides insights to the evolution of families (e.g. a bloodline that dies out) or can be used to identify potential missing data or errors in the dataset (**O3**) if the bloodlines are not complete (e.g. ancestor tree starts late or descendant tree stops early).

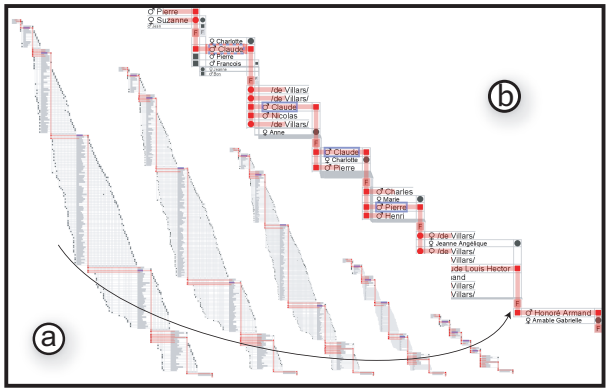
3.5 Queries

The primary purpose of the query window (Figure 3d) is to provide details on the attributes associated to selected entries, which are highlighted in the same color as in the main window. Clicking on the identifier of an individual pans the main window to reveal it, allowing easy association between attributes and the genealogy visualization, which is important when switching from browsing attributes to browsing structure.

3.5.1 Search

The query window also serves as a "search" and "filtering" panel. A text field allows users to add search terms on any entry or attribute in the dataset (**S1,S2**). Multi-term searches are possible. A drop down list for constraining search to specific attributes (e.g. entity name or birth place) is also present. As the user types her search terms, the entities that satisfy the terms are interactively highlighted on the main window and overview, and the number of matches is displayed. By pressing Enter, the user finalizes his/her search, and the results are selected using a color specific to that query.

Search in combination with visual selections can be used to examine attributes of different individuals of the genealogy (e.g. gender,



]

Fig. 7. Animation (a) of filtering using DOI, (b) focused on a selected bloodline (red) and search results (blue).

status, location), as well as how these attributes (such as land ownership, office, decease) propagate within a family (**A3**), as we will discuss in our use cases.

3.5.2 Filtering

Genealogists often want to focus on small parts of their dataset at a time, such as a specific bloodline or individuals in a particular government office. To support this feature, we provide a filtered mode. When turned on, filtering computes a degree of interest (DOI) [8] for all entries based on their distance to the user's current selection(s) and highlighted path(s), and shrinks entries with low DOI (Fig. 7). These changes are smoothly animated. Thus users can switch their focus to specific parts of the dataset without losing context.

3.6 Timeline

To support temporal tasks (**A2**), we provide a timeline where each individual is visualized as a horizontal line spanning its life range (Fig. 3c) . This range is computed as the union of all the dates related to the individual (usually birth, marriage, death, burial) mentioned in the GEDCOM file. The vertical position is computed according to the generation and position within the generation.

The time range of all individuals visible in the main viewport is shown as a semi-transparent focus rectangle on top of the timeline. Selections are also highlighted on the timeline in their own color, which allows comparing the life-spans of several individuals. The timeline is also useful at spotting erroneous dates that are frequent in genealogies since genealogy systems rarely perform any checking on dates.

3.7 Dataset Examples

We tested our system with several genealogical datasets of various sizes available to the community, such as the genealogy of the Greek pantheon (Fig. 1), the European royal families, and the Christian Bible. We were able to display and manipulate at interactive rates large datasets (e.g. the European royal family genealogy) of over 3000 individuals, with approximately 80 identified generations spanning the range of several hundred years, as well as all of the datasets available on the Anthropology Web site [2], including the "Ragusan" nobility data with 5999 individuals. Moreover, we were able to use directly the personal datasets of the expert users who took part in the user feedback sessions. We report here on interesting findings in two popular genealogies: the Bible genealogy and the European noble families.

3.7.1 The Bible

The Christian Bible is full of interesting genealogical patterns that are clearly visible using GeneaQuilts. The common ancestor to Mary and Joseph, King David, can be easily seen by selecting Mary and Joseph in different colors. Note that Mary's ancestry is not explicitly detailed in the Bible, but the dataset we used reflects a specific interpretation whereby Joseph's ancestry as reported by Luke is attributed to Mary. We also already mentioned a case of consanguineous marriage between one of Abraham's brothers Nahor and his niece Milcah (Fig. 5). Another noticeable event is reported in the Book of Genesis, chapter 19: the story of Lot, who had a son with each of his two daughters[2] Fig. 8 shows this story with Lot selected. The Greek pantheon also contains many occurrences of consanguinity (Fig. 1).

3.7.2 The European Noble Families

The European noble families also contain well-known individuals, such as Henry VIII who had 6 wives (and two mistresses not included in the dataset). In Fig. 9, the six filled squares to the right of Henry VIII depicting his marriages are clearly visible. Looking at the families, it is clear that he only had children with three wives (Jane Seymour, Catherine of Aragon and Anne Boleyn) but the children are not visible in the viewport. Triggering the Bring & Slide tool shows the children, in the order of marriages with a continuous arc to connect children from the same marriage. Therefore, without sliding or panning, the number and names of children are visible along with the family they belong to. Also, Fig. 9 reveals — maybe unexpectedly for non History-savvy readers — that at the same period, Catherine Parr married four times and several other men and women married twice.

3.8 Scalability

There are several interpretations of scalability. Technically, GeneaQuilts can load and visualize at interactive rates genealogies of up to 10,000 nodes, using only the standard Java 1.6 libraries with Piccolo2D for the rendering [3]. Using an OpenGL graphics pipeline would probably push the limit to several million nodes [7].

In terms of readability and usability, the bottleneck is screen size. When the number of individual grows, GeneaQuilts becomes less useful: each generation becomes a very long list of names, the families cannot be visible along with the individuals and extensive navigation is required to see patterns. The problem occurs when one generation contains more than a few hundred individuals: the viewport cannot show the previous or next generation while maintaining the legibility of name labels, and only the overall structure can be seen through the overview window. Still, using selection/search and filtering usually simplifies the visualization to a usable size.

For datasets above tens of thousands of individuals, we believe other visualization strategies should be provided. The initial quilts article proposes simple aggregation mechanisms [33] that can be easily implemented but we have not investigated its compatibility with the visualization tasks that still make sense at this level of scale.

4 User Feedback and use cases

The three anthropologists and four historians who took part in our initial extended interviews (sec. 2.2) were later presented with GeneaQuilts and provided us with preliminary feedback and use cases. They were briefly trained on the system using a toy dataset and then loaded their own research data to explore and provide us with comments.

[2]Full details can be found in the "Lot (Bible)" entry of Wikipedia.

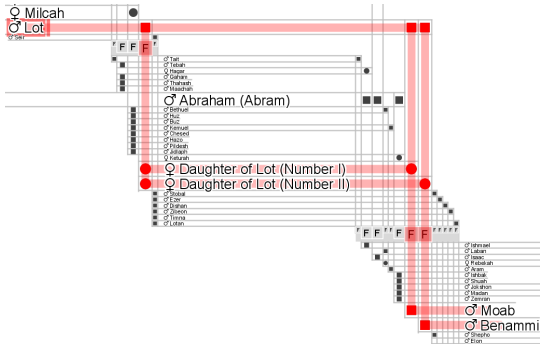


Fig. 8. Lot and his descendants from the Book of Genesis.

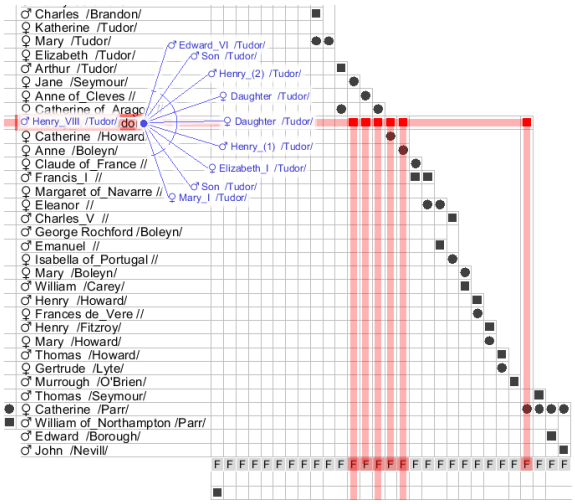


Fig. 9. Henry VIII and his Wives.

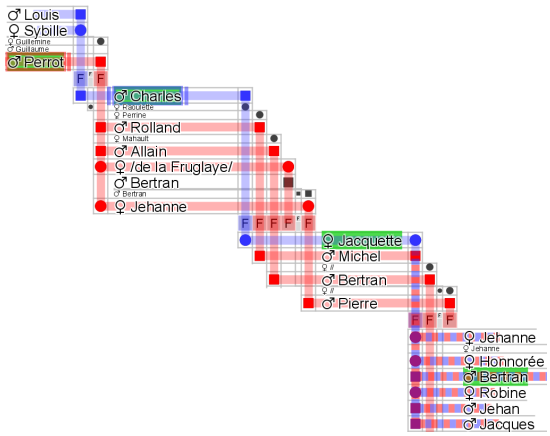


Fig. 10. Evolution of land ownership in a 15th century family.

4.1 General Feedback

Participants commented quickly on the fact that as this is a new visualization it requires some initial familiarization. They acknowledged they never thought of visually exploring large genealogies due to lack of a system to do it properly. Nevertheless they all agreed that this visualization has many benefits compared to traditional graphs or P-graphs as it can display a larger number of individuals without link crossing or overlap.

Both sets of users found the overview window with the focus indicator an extremely useful feature that allows them to keep context during their exploration.

Both groups of users asked for the ability to edit and add content to the dataset using GeneaQuilts, as well as the possibility to export selected parts to use for private or scientific presentations.

Finally, all of our users pointed out that they are not aware of any other system that can help them visualize and explore large genealogy datasets without losing context and expressed a strong desire to be given the prototype as is to start using it with their datasets.

4.2 Anthropologists

The three anthropologists were particularly impressed with what they called "the cleanness" of the visualization. As they usually deal with large datasets of clans or tribes that intermarry across generations, link density is often an issue in other visualizations, but not in GeneaQuilts.

The anthropologists mentioned that they tend to share datasets more than historians and thus commented that when getting a new dataset, the overview combined with the timeline provides great insight into the focus of the dataset and the possibility that parts of the dataset may be incomplete.

Anthropologists commented on the usefulness of multiple selections and bloodline highlighting for identifying inter-marriages. For example, by highlighting a bride's bloodline when there is a skip-link, they could see a marriage between an uncle and a niece, a very common occurrence in specific tribes or groups. By highlighting the parents of the bride, they could also determine whether she got married to an uncle from the paternal or maternal line, another aspect of interest to them. They similarly were able to spot marriages between maternal and paternal cousins of first degree.

Anthropologists commented that their structural analysis system could count the different types of rings (sec. 2.3.4) but GeneaQuilts could tell if some patterns were spread evenly in a population or clustered around a specific family or time period. They insisted that keeping a tighter relation between statistics and visualization was very important to their studies.

Although they found selection useful for exploring specific individuals, given their interest in inter-marriages, anthropologists requested three additional features: 1) the ability to incorporate their own research software that computes rings and project the results on the visualization; 2) the option to show all possible connection paths between two selected individuals (through marriage or blood); and 3) an interactive topology search feature based on their notation (e.g. W=MBDS means "show me all the Wives married to their Mother's Brother's Daughter's Son's"). Anthropologists also mentioned that the representations they use clearly differentiate gender, and that GeneaQuilts could use colors for that purpose.

4.3 Historians

The historians provided us with two datasets: *the "Fruglaye" family* in the 15th and 16th century in the region of Brittany [17], containing 44 individual and land property data as attributes (Fig. 10); and *the "Broe" family* (Fig. 7) in the 16th and 17th century in the region of Paris, containing 223 individuals an a large number of attributes for some of them, describing events that occured during their life.

The historians found of great importance the search feature on all attributes, especially when combined with the filtering and timeline. This helped them explore how attributes change over time, a task that is currently not supported by other tools. Looking at the Fruglaye family tree on Fig. 10, built from the journal of a descendant [17], they were able to see the story of one of the lands. Typing the name of the land in the search box reveals all individuals owning the land (in green). The land initially belongs to Perrot de la Fruglaye selected in red. It is then passed (the journal explains it has been sold) to Charles Pelouaisel (in blue) and remains in the Pelouaisel family. On the third generation, Jacquette and Michel marry (dashed blue/red pattern) and the land returns to the "de la Fruglaye" family through a woman, an anomaly at a time when only men inherited lands.

Besides the evolution of lands and titles across time and families, the historians commented that they would have liked ways to observe how numerical attributes (such as dowry, inheritance or salary) changed across generations. They also commented that in their work they would like to see connections other than marriage and lineage between individuals, such as trading or friend relationships. However, in the lack of convenient software to enter and visualize this data, they never captured it in digital form.

5 Discussion and Future Work

GeneaQuilts aims at providing a scalable visualization technique for easy exploration and navigation through genealogy datasets. We have discussed how a large number of genealogy research tasks can be performed directly using this basic visualization augmented with simple interaction techniques. In this section we discuss how the basic GeneaQuilts visualization can be extended to support the remaining tasks identified in Sec.2.2 and our user feedback session.

GeneaQuilts visualizations are readable and stable maps for genealogies; they can then be overlaid and augmented with all the vi-

sual attributes routinely used in information visualization and linked to more related views. We provided a timeline to link structure with time but more types of linking have been asked for in our interviews.

First of all, we can augment the system with the calculation of structural metrics, such as inter-marriage rings of specific length. These ring paths can then be shown on the main visualization, the overview window and the query window in discrete colors, similarly to how different selections are currently displayed. Moreover, using Puck [1] we can calculate and display all paths (sanguine or not) between two individuals and overlay them on the visualization. The options to calculate such structures can be placed in the query window or on the system menu, that currently only serves to open dataset files. Such an integration would turn GeneaQuilts into a visual tool for genealogy analytics, bridging the gap between statistical/model-based and detailed/exploratory-based analysis.

We can also augment GeneaQuilts with additional computed metrics, such as the genetic distance from an individual or the percentage of genes shared between individuals. These distances or similarities could be conveyed with colors. This would allow anthropologists, biologists and doctors to identify specific blood relations of an individual (e.g. aunts/uncles and cousins of the n-th degree, etc.)

In its current form, the existence of a relationship on the visualization is depicted by a filled cell in the matrix, its shape indicates gender, and its relative position (left or right of the individual) indicates the type of relationship. By using color attributes on the matrix cells we could overlay types of relationships other than blood and marriage.

One request of our users was to present the evolution of numerical attributes across generations. We plan to extend our visualization to provide summaries of specific attributes on the timeline in the form of simple bar-charts that are filtered by and colored in a manner corresponding to any possible user selections. Numerical attributes could also be shown next to individuals in the form of histograms.

Finally, our support for filtering partly addresses the problem of visual clutter by allowing users to focus on specific parts of the visualization. We further plan to allow users to aggregate parts of the visualization. In a way similar to the original Quilts [33], the color intensity of matrix cells could indicate the density of aggregated area.

6 Conclusions

We have presented GeneaQuilts, a novel visualization technique that depicts genealogies in the form of a layered, diagonally-filled matrix. Our visualization eliminates crossings and accommodates very large datasets in the order of thousands of individuals. By depicting individuals as rows and families as columns, with parents always at a higher layer (generation) than their children, our visualization clearly exhibits marriage and parent/children relationships, as well as other interesting relationships such as cross-generational and consanguine marriages.

We implemented GeneaQuilts as a component of a larger prototype system aimed at genealogy exploration, which supports interaction techniques designed for rapid navigation in large datasets. Our system was very positively received by domain experts, and was shown to support a large number of genealogy research tasks identified through extended interviews. As GeneaQuilts is a novel visualization technique, we have also discussed how our current system can be extended to support an even larger range of identified tasks.

We plan to extend the basic GeneaQuilts functionality in the manner discussed in the future work section and hope it will be quickly adopted by historians, ethnographers, anthropologists and hobbyists for their explorations and analyzes.

References

[1] Puck: Program for the use and computation of kinship data, http://www.kintip.net/content/view/55/21. Program.

[2] Kinsources: Kinship data repository, http://kinsource.net/kinsrc/bin/view/kinsources/, 2010.

[3] B. B. Bederson, J. Grosjean, and J. Meyer. Toolkit design for interactive structured graphics. *IEEE Trans. Softw. Eng.*, 30(8):535–546, 2004.

[4] J. Bertin. *Sémiologie graphique : Les diagrammes - Les réseaux - Les cartes*. Editions de l'Ecole des Hautes Etudes en Sciences, Paris, France, les réimpressions edition, 1967.

[5] W. de Nooy, A. Mrvar, and V. Batagelj. *Exploratory Social Network Analysis with Pajek*. Structural Analysis in the Social Sciences. Cambridge University Press, Mar. 2005.

[6] G. M. Draper and F. Riesenfeld. Interactive fan charts: A space-saving technique for genealogical graph exploration. In *8th Workshop on Technology for Family History and Genealogical Research*, 2009.

[7] J.-D. Fekete and C. Plaisant. Interactive information visualization of a million items. In *INFOVIS '02: Proceedings of the IEEE Symposium on Information Visualization (InfoVis'02)*, page 117, Washington, DC, USA, 2002. IEEE Computer Society.

[8] G. W. Furnas. Generalized fisheye views. In *Proceedings of the ACM CHI'86 Conference on Human Factors in Computer Systems*, pages 16–23, 1986.

[9] G. W. Furnas and J. Zacks. Multitrees: enriching and reusing hierarchical structure. In *CHI '94: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 330–336, New York, NY, USA, 1994. ACM.

[10] E. R. Gansner, E. Koutsofios, S. C. North, and K.-P. Vo. A technique for drawing directed graphs. *IEEE Trans. Softw. Eng.*, 19(3):214–230, 1993.

[11] E. R. Gansner and S. C. North. An open graph visualization system and its applications to software engineering. *Softw. Pract. Exper.*, 30(11):1203–1233, 2000.

[12] GenoPro. Genopro, inc. http://www.genopro.com/.

[13] M. Ghoniem, J.-D. Fekete, and P. Castagliola. On the readability of graphs using node-link and matrix-based representations: a controlled experiment and statistical analysis. *Information Visualization*, 4(2):114–135, 2005.

[14] K. Hamberger, M. Houseman, I. Daillant, L. Barry, and D. White. Matrimonial Ring Structures. *Mathématiques et Sciences Humaines*, (168):83–119, 2004.

[15] F. Héran. *Figures de la parenté*. Sociologies. Presses Universitaires de France, Paris, Mar. 2009.

[16] G. B. Hoffman. Genealogy in the new times, 1999. http://www.genealogy.com/genealogy/61 gary.html.

[17] R. Laigue. Le livre de raison de Jehan de la Fruglaye, seigneur de la Villaubaust. *Bulletin archéologique de l'association bretonne*, XX:108–132, 1901.

[18] B. Lee, C. Plaisant, C. Sims Parr, J.-D. Fekete, and N. Henry. Task taxonomy for graph visualization. In *BELIV '06: Proceedings of the 2006 AVI workshop on BEyond time and errors*, pages 1–5, New York, NY, USA, 2006. ACM.

[19] M. J. McGuffin and R. Balakrishnan. Interactive visualization of genealogical graphs. In *INFOVIS '05: Proceedings of the Proceedings of the 2005 IEEE Symposium on Information Visualization*, page 3, Washington, DC, USA, 2005. IEEE Computer Society.

[20] M. J. McGuffin and m. c. schraefel. A comparison of hyperstructures: zzstructures, mspaces, and polyarchies. In *HYPERTEXT '04: Proceedings of the fifteenth ACM conference on Hypertext and hypermedia*, pages 153–162, New York, NY, USA, 2004. ACM.

[21] E. S. Mills. Analyzing and reviewing published sources. *OnBoard: Newsletter of the Board for Certification of Genealogists*, 3(16), May 1997.

[22] E. S. Mills. Genealogy in the 'information age': History's new frontier? *National Genealogical Society Quarterly*, 91:260–77, Dec. 2003.

[23] T. Moscovich, F. Chevalier, N. Henry, E. Pietriga, and J.-D. Fekete. Topology-aware navigation in large networks. In *CHI '09: Proceedings of the 27th international conference on Human factors in computing systems*, pages 2319–2328, New York, NY, USA, 2009. ACM.

[24] S. W. Mumford. The genealogical software report card ©2000, http://www.mumford.ca/reportcard/basic.html, 2005.

[25] R. Munroe. Xkcd#657, http://xkcd.com/657/, 2010.

[26] T. Munzner, F. Guimbretière, S. Tasiran, L. Zhang, and Y. Zhou. TreeJuxtaposer: scalable tree comparison using focus+context with guaranteed visibility. In *Computer Graphics (ACM SIGGRAPH 2003 Proceedings)*, pages 453–462, 2003.

[27] J. Priestley. *A Chart of Biography*. London: J. Johnson, St. Paul's Church Yard, 1765.

[28] Progeny Genealogy Inc. Genelines, http://progenygenealogy.com/genelines.html, 2010.

[29] A. Shoumatoff. *The Mountain of Names: A History of the Human Family*.

Simon & Schuster, Inc., 1985.

[30] H. Siirtola and E. Mäkinen. Constructing and reconstructing the reorderable matrix. *Information Visualization*, 4(1):32–48, 2005.

[31] K. Sugiyama, S. Tagawa, and M. Toda. Methods for visual understanding of hierarchical system structures. *IEEE Trans. Systems, Man and Cybernetics*, 11(2):109–125, Feb. 1981.

[32] The Church of Jesus Christ of Latter-day Saints. The GEDCOM Standard Release 5.5, Jan. 1996.

[33] B. Watson, D. Brink, M. Stallmann, R. Devarajan, M. Rakow, T.-M. Rhyne, and H. Patel. Visualizing very large layered graphs with quilts. *IEEE Information Visualization Conference Poster*, 2007.

[34] J. Wesson, M. d. Plessis, and C. Oosthuizen. A zoomtree interface for searching genealogical information. In *AFRIGRAPH '04: Proceedings of the 3rd international conference on Computer graphics, virtual reality, visualisation and interaction in Africa*, pages 131–136, New York, NY, USA, 2004. ACM.

[35] D. R. White and P. Jorion. Representing and computing kinship: A new approach. *Current Anthropology*, 33(4):454–463, Aug. - Oct. 1992.