

Know Your Ancestors Better: Demographic Visualization for Large Genealogical Data Sets

Tae Woo Kim and Geoffrey M. Draper

*Department of Computer & Information Sciences
Brigham Young University—Hawaii*

Abstract

Most genealogy software is designed to make it easy to view and edit data about individuals and their relationships to others. While this is very useful, sometimes it is desirable to view information and discover trends about hundreds or thousands of individuals all at once. We present a graphical visualization system for formulating queries and viewing their results for large populations rather than for individuals. Our method works with the GEDCOM standard, and is designed for ease of use and interactivity.

1. Introduction

There are many genealogy software packages that show individuals' information and relationships among individuals. Most genealogy software uses tree-based pedigree charts for displaying genealogical graphs. Tree-based pedigree charts are very convenient and effective for reading and understanding each individual's information and relationships. However, as it focuses only on individuals, it is hard to "zoom out" and examine the data set as a whole. Thus, there is a growing interest in the area of "macro-genealogy". We propose a novel visualization system for macro-genealogy that will help us to understand whole family tree by using a graphical tool for performing queries and visualizing results interactively.

2. Related Work

This research applies aspects of information visualization, namely, visual query languages and radial layouts, to genealogy.

2.1 Visual Query Language

There have been many efforts to make databases easier to use. Many techniques were introduced; one was and is still being used is 'Query By Example' [11]. It provides a way for a user to perform queries without any knowledge of a query language such as SQL. More recently, there is a web-based visualization of historical trends in baby naming which is called 'The Name Voyager'. As a

name is being typed, stacked graphs show the name's popularity in different years [10]. Similarly, our software uses a graphical interface to iteratively work with users. While “The Name Voyager” uses stacked graphs, our software uses a radial layout for both querying and visualization. Focusing on ease of use and interactivity, it generates a query string simply by dragging attributes around a graphical canvas.

2.2 Radial Layouts

Another area of related work is radial user interfaces. Laying out the components of a user interface in a radial, or circular, pattern is an increasingly popular motif in the field of human-computer interaction [4]. The most common example is that of a marking menu [6], although other applications have been built for the domains of network security [7, 5], hierarchical tree browsing [9], and time-series data [1]. The fan chart is an early example of a radial visualization for genealogy [3].

2.3 Genealogy

Most genealogical software can use and manage large data sets, but it is challenging to display all depiction of relationships in a screen. McGuffin and Balakrishnan use a dual-tree scheme so that they can display as much as possible in a screen to understand the family tree better [8]. We are approaching a different way to utilize and understand whole data set better by not focusing individuals but on the whole data set. The Name Voyager queries the whole data set but it has only one attribute. Our system has many attributes and it is possible to add as many attributes as desired. We are not suggesting a replacement of other genealogical software, but simply offering a complementary technique for visualizing genealogical data.

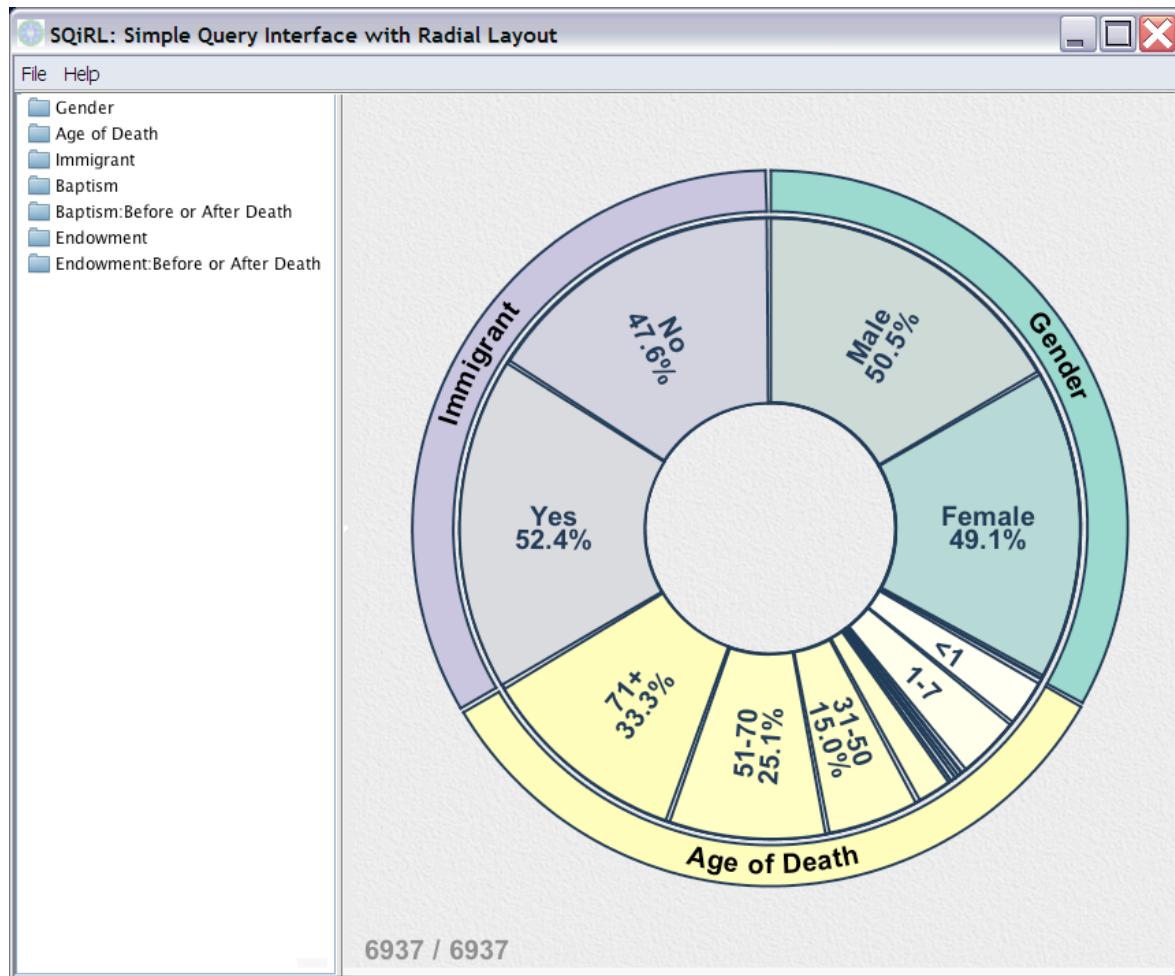
Genealogical Data Communication, GEDCOM, is one of the most common formats used for genealogy information exchange. The GEDCOM specification was originally developed in 1985 and is owned and managed by the Family History Department of the Church of Jesus Christ of Latter-day Saints. The GEDCOM specification uses a set of tags to describe the information in a family file, such as INDI for individual, FAM for family, BIRT for birth and DATE for a date.

3. Visualization and Interaction

Our work is based on SQiRL [2], a prototype data visualization system currently developed and maintained by faculty and students at BYU-Hawaii. While originally created for querying the results of

election polls, we will show in this paper that it may also be used to support genealogical research -- in particular, finding global trends in a family tree. In this paper, we will limit our discussion to those features of SQiRL which demonstrate its utility for querying GEDCOM-based genealogical data sets. For a fuller treatment of the visualization techniques employed by SQiRL, we refer the reader to [2].

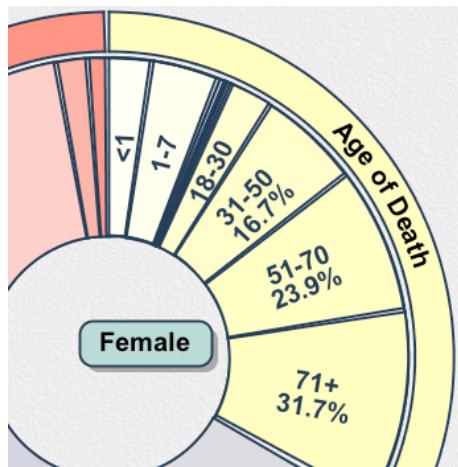
The distinguishing feature of SQiRL's user interface is its graphical canvas. The canvas itself is dominated by a large “donut chart” (essentially a pie chart with the center removed). To the left of the canvas is a sidebar containing a two-level tree widget containing “attributes” (such as *Gender*) and their corresponding “values” (such as *Female*, *Male*, and *Unknown*). Figure 1 shows the SQiRL system at startup.



(Figure 1: System at startup. The dominant feature of the user interface is an interactive "donut chart" to which sectors can be added and removed. Querying specific subpopulations is done by dragging attributes from the sidebar into the donut's interior.)

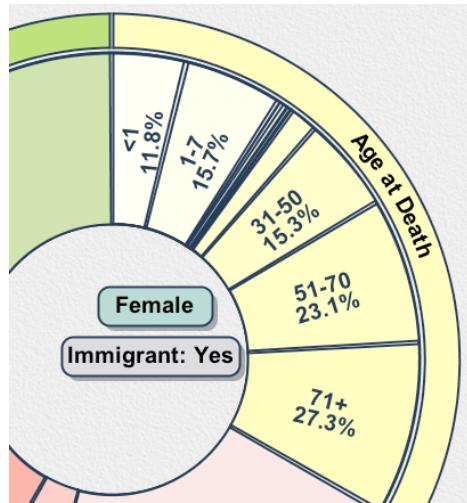
By default, the donut chart displays information about the entire GEDCOM file. For example, in Figure 1, we see that 52.4% percent of the people in this GEDCOM file emigrated during their lifetime (meaning their country of residence at the time of death is different from their country of birth). It also shows that one third of the people died over the age of 71. The number of individuals in the current GEDCOM (6937 in this file) is displayed in the lower-left corner of the canvas.

While viewing this information for the data set as a whole is useful, it is more often the case that users wish to “drill down” into a specific sub-population and view their characteristics. For example, to limit the query only to females, one drags the value labeled *Female* from the side panel into the center of the donut-shaped widget. The sectors along the donut then resize to show the percentages that correspond to the females in the data set. In Figure 2, we see that slightly fewer women survived to old age, as compared to the population as a whole. The resizing of the sectors occurs in a smooth animation, helping the user to maintain context. Furthermore, the number in the corner of the canvas updates to show that there are 3408 individuals in the data set that match the attribute *Female*.



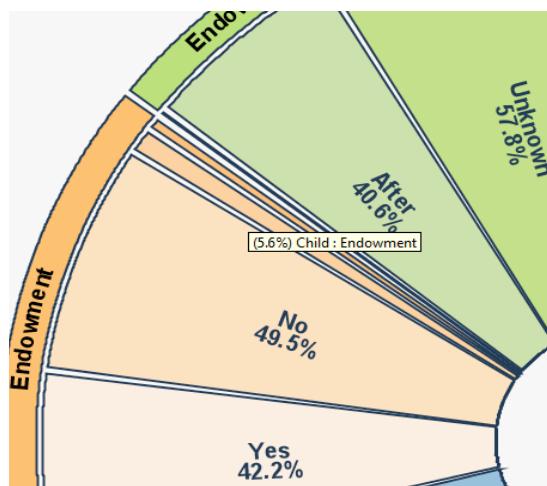
(Figure 2: Placing the "Female" icon in the donut's interior restricts the current query to females only.)

More complex queries can be constructed by placing multiple attributes into the donut's interior: disparate attributes are *ANDed* together. For example, placing the icons for *Female* and *Immigrant: Yes* would cause the system to limit the query to only those women who also moved away from their place of birth. As shown in Figure 3, women who emigrated were much more likely to die at a young age than women in general.



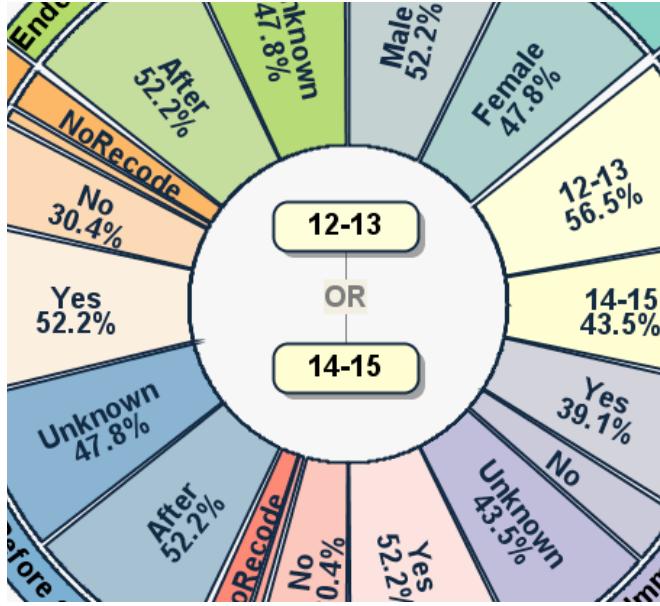
(Figure 3: Placing multiple icons in the donut's interior performs a logical AND on the attributes. The donut chart now shows results only for females who emigrated from their birthplace.)

For those sectors that are too narrow to contain a printed label, “mousing over” the region will reveal a small popup tool-tip with the sector’s name on it (Figure 4).



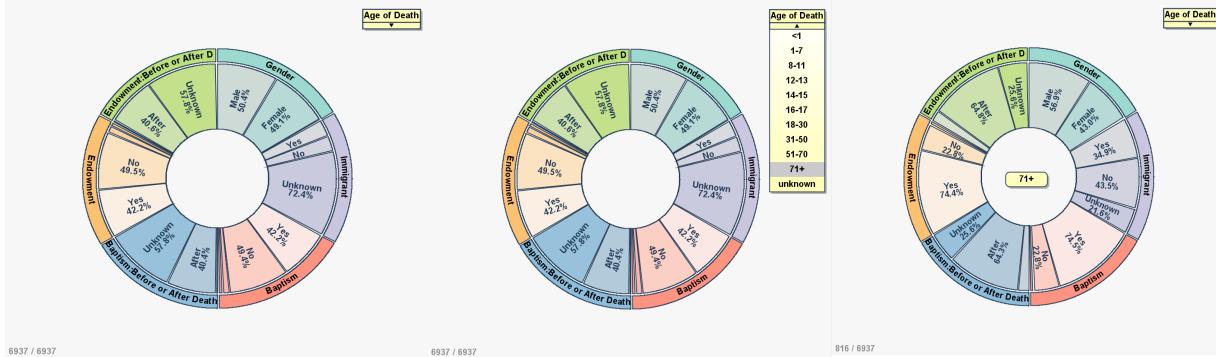
(Figure 4: Popup tool tips reveal information for sectors that are too narrow to see otherwise.)

When two values are drawn from the same attribute, they are *ORed* together to form the query. In the visualization, this is indicated by rendering a thin labeled edge between the icons (Figure 5).



(Figure 5: Placing multiple icons from the same "category" into the donut's interior results in a logical OR. Here, we see the results for all persons who died between the ages of 12 and 15.)

In addition to being on or inside the donut, icons may also be dragged outside it. After dragging out an attribute, it will appear as drop-down box. If we expand the drop-down box, it will show all the values of the attribute. Again, each value can be dragged into the middle circle to focus the query on a specific group (Figure 6).



(Figure 6: Dragging an attribute to the space outside the ring causes a "combo box" to appear, from which specific values may be selected.)

In addition, SQiRL also supports the direct comparison of two separate queries; however, this feature is currently in the experimental phase and is beyond the scope of this paper.

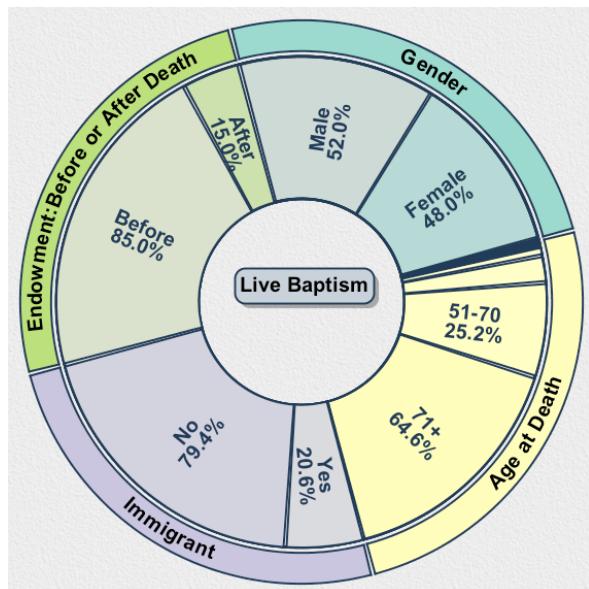
4. Data Set

The GEDCOM file format is widely recognized as the standard format for genealogical data interchange. We examined a number of existing GEDCOM parsers online, but found them to be ill-suited for the requirements of our database schema. Thus, we wrote a custom GEDCOM parser which converts the data into the format required by the SQiRL system. Parsing the GEDCOM is a one-time pre-processing step for a given data set; thereafter the system reads the data directly from the database.

4.1 Helps for LDS Genealogists

As noted above, our system identifies several attributes that are of interest to genealogy hobbyists and professionals in general, such as age at death and emigration patterns. In addition, our system also maintains attributes and values relating to LDS ordinances, specifically regarding whether baptisms and endowments were performed during an individual's lifetime, or by vicarious proxy.

For example, Figure 7 shows that, in this particular data set, the vast majority of those individuals who were baptized during their own lifetime did not emigrate from their place of birth. They were also far more likely to have received the endowment during their lifetime.



(Figure 7: Interestingly, those persons who were baptized during their lifetime also tended to live a long time.)

5. Future Work

The GEDCOM format contains more information than our system currently supports. There is not only information about individuals but also about families. Our current software uses only individuals' information, but it may be good to parse more family-related information too. For example, we could show how many members of the entire family have been married and how many children each individual had. It would also be helpful to conduct a user study to evaluate how well this tool helps genealogists to understand trends in their data.

6. Conclusion

There are many genealogy software applications which let users see the relationships and information about an individual to have better understanding of an individual. This is essential to genealogy work, but we believe there is much value in being able to see trends relating to the entire family tree. We accomplished this by using a graphical tool for performing queries and visualizing results in radial chart. The graphical interface is based on a drag-and-drop metaphor that is easy to learn. We invite producers of existing genealogy software to consider incorporating the ideas in this paper into their own products.

References

- [1] Carlis, J.V., Konstan, J.A., "Interactive Visualization of Serial Periodic Data," *Proc. 11th Ann. ACM Symp. User Interface Software and Technology (UIST '98)*, pp. 29-38, 1998.
- [2] Draper, G.M. and Riesenfeld R.F., "Who Votes For What? A Visual Query Language for Opinion Data," *IEEE Trans. Visualization and Computer Graphics*, vol. 14, no. 6, pp. 1197-1204, Nov./Dec. 2008.
- [3] Draper, G.M., Riesenfeld, R.F., "Interactive Fan Charts: A Space-saving Technique for Genealogical Graph Exploration," *Proc. Eighth Ann. Workshop Technology for Family History and Genealogical Research*, 2008.
- [4] Draper, G.M., Livnat, Y., and Riesenfeld R.F., "A Survey of Radial Methods for Information Visualization," *IEEE Trans. Visualization and Computer Graphics*, vol. 15, no. 5, Sep./Oct. 2009.
- [5] Keim, D.A., Mansmann, F., Schneidewind, and T. Schreck, "Monitoring Network Traffic with Radial Traffic Analyzer," *Proc. IEEE Symp. Vivual Analytics Science and technology. (VAST '06)*, 3 pp. 123-128, 2006.

- [6] Kurtenbach and Buxton, W. "The Limits of Expert Performance Using Hierarchic Marking Menus," *Proc.Computer-Human Interacion(CHI '93)*, pp. 482-487, 1993.
- [7] Livnat, Y., Agutter, S., Moon, R.F. Erbacher, and Foresti, S., "A Visualization Paradigm for Network Intrusion Detection," *Proc.IEEE Workshop Information Assurance and Security*, pp. 30-37, 2005.
- [8] McGuffin, M. J and Balakrishnan, R., "Interactive Visualization of Genealogical Graphs," *Proc. IEEE Symposium on information Visualization*, 16-23, 2005.
- [9] Mohammadi-Aragh, M.J. and Jankun-Kelly, T., "Moiretrees: Visualization and Interaction for Multi-Hierarchical Data," *Proc. EuroGraffhics – IEEE VGTC Symp. Visualization*, K.W. Brodlie, D.J. Duke, and K.I. Joy, eds., pp. 59-66, 2005.
- [10] Name Voyager: <http://babynamewizard.com/namevoyager/inv0105.htm>
- [11] Zloof, M. M., "Query by Example: a data base language," *IBM Systems Journal*, 16(4):324-343, 1977.