Automatically Recreating Probabilistic History through Genealogy

Robert Ball and Patrick Beck
School of Computing
Weber State University

Abstract—Genealogy, also known as family history, is one of the few things that all humanity has in common. Regardless of culture, race, or language, all mankind is connected through the murky past through family ties to each other. Visualization of one's ancestors is a common practice. However, visual analytics tools to understand ones past are surprisingly lacking. We present a specific analytical visualization paradigm for visualizing where one's ancestors most likely lived in the past based on family connections. We go beyond the basics of simply showing where ancestors were born and died and show possible locations of the movement of people in the past. Instead of showing such information as fact we allow genealogical analysts the ability to easily choose and configure what events a person likely attended. Of particular note is the ability to see overviews of where thousands of people lived over time. The genealogical analyst can quickly see where families were clumped in generations and the possible individual movements of people in time. In addition, the visualization is linked to details that allows further investigation of individuals and families. This paper introduces a new way to visualize genealogy where people try to understand not just how people are related, but where they lived and what events in their lives took place.

Index Terms— Family history, genealogy, mass movement, uncertainty

1 Introduction

People form identities and cultures based on their genealogy – their own personal past – and society forms an identity based on the narratives and stories that we tell ourselves. This leads to Arnon Herskovitz's suggestion that genealogy as an academic discipline is focused on the following: people, families, communities, representations, and data [7].

"We are all storytellers, and we are the stories we tell. … adolescents and young adults in modern societies are challenged to formulate meaningful answers to the twin identity questions: Who am I? How do I fit into the adult world?" [11]. In other words, "finding oneself" is nothing more than creating a narrative.

The two research questions that guided the work behind this paper are the following:

1. Where did an individual ancestor most likely live during their life given limited genealogical data?
2. Where did ALL of a person's ancestors live and how were they interconnected as families and communities?

Discovering all the locations and time spent at different locations during a deceased person's life is impossible. Making the job harder is that most genealogical records are based simply on vital records (e.g. birth, marriage, and death certificates).

For example, let us look at an example entry of a person in the most common genealogical file format – a GEDCOM file. A GEDCOM file, officially Genealogical Data Communication file, is a program-independent file that is used to transfer genealogical data between different programs [9].

The file is ASCII-based and there are hundreds of possible tags for information from longitude and latitude for locations to information about multimedia. The following is an example entry of from a gedcom file:

```
0 @I4691@ INDI
1 NAME Henry /Pitt/
1 SEX M
1 BIRT
2 DATE 7 OCT 1613
2 PLAC Bristol, Gloucestershire, England, United Kingdom
1 DEAT
2 DATE 22 AUG 1666
2 PLAC Will, Isle of Wight, Virginia, United States
```

Based on the above entry, all the can be easily gleaned is that Henry Pitt was male, he was born October 7, 1613 in Bristol, Gloucestershire, England, United Kingdom and died August 22, 1666 in Will, Isle of Wight, Virginia, United States. With minimal analysis one can also conclude that Henry Pitt died at the age of 52, about a month and half before his 53rd birthday.

2 Related Work

While new academic genealogical visualizations are fairly rare, there are several of particular note.

TimeNets is a temporal visualization that shows people as lines in terms of dates. Lines that get closer together indicate marriage while the same lines that diverge indicate divorce. Dashed lines that come from other lines indicate children. Its biggest weakness is that when viewed with large amounts of data the high quantity of lines can be confusing [10].

Following the idea of temporal visualizations, Ball introduces a temporal family-centric genealogy visualization paradigm. Following along the same idea as TimeNets, he introduces a paradigm of showing genealogical data grouped together by families in terms of time [2].

One particular recent development in genealogical visualizations is called GeneaQuilts. GeneaQuilts is a table where rows are individuals and columns are families. It excels in showing thousands of people at a time and allowing the user to interactively see how different people in a dataset are related. GeneaQuilt's main weakness is that it is inherently generation based and the user does not have any insight into when different generations lived [3].

One of goals of this paper is to explore mass movement of families. Dodge, Weibel, and Lautenschütz systematically classify different ways to show movement patterns [5]. Similarly, Andrienko and Andrienko's work on designing visual analytics methods for movement data has been useful in designing flow movement and overview of locations [1].

The most commonly used visualization tool is the classic pedigree chart. It is a binary tree that shows the relationship of parents to children starting with the focus person as the root. Another common visualization is the fan chart that shows the focus person in the center of a circle and each proceeding generation in circular layers around that circle. For a more comprehensive list of modern genealogical visualizations see [2].

Looking at our example of Henry Pitt, more information can be understood about his past. For example, one can also show context for Henry Pitt's life. For example, websites exist that automatically

report who was the monarch in the UK at the time, what historical figures were alive, and other such related information [8].

Fig. 1. An example visualization from a genealogy website [12].

A number of map visualizations exist that show the overview of genealogical migrations of a focus person. Many of the commercial map visualizations do a good job highlighting relevant countries where a person lived. In Fig. 1 Henry Pitt is shown that he was born in the UK with a blue circle. It shows where he was born and where he died with a blue line in between to show movement.

The numbers on the circles in Fig. 1 indicate how many generations back the person is from the focus person. The focus person's genealogical data (Henry Pitt's in this example) goes back nine generations to Northern Sweden, UK, and parts of Germany.

These types of visualizations are useful in showing the country origins of the focus person. However, they lack in several important ways. First, they inherently act as overviews – there is rarely a way to get additional information. However, to be fair, what additional information could be shown? There is very little *certain* information that is usually known about the deceased.

Second, the generations of circles are literally on top of each other and make it difficult to understand and to differentiate different generations. As there are many of these types of commercial visualizations many allow the user to traverse generations with different degrees of interactivity – usually by using a slider to show different generations or different time frames. In essence, these types of visualizations are good for overview of data only and often do not support additional analysis.

3 Uncertainty in Genealogical Data

Commercial geospatial genealogical visualizations stop with an overview because that is where the novice genealogist usually stops their research. With no additional solid information - no additional vital records found - what else is there to visualize?

Professional genealogists do not stop with vital records. They look at land records, journals, court records – in short, any clue that will help them understand where a person lived.

However, such investigations are rife with uncertainty and uncertainty is not something people are generally comfortable with. Psychological studies about uncertainty show that when most people are confronted with uncertainty they hesitate or make less than rational decisions that they might otherwise make [13].

One interesting thing about genealogical data is that the analyst can never be *completely* certain about anything. Even official government vital records are often wrong – especially in the past before the use of computers. Errors in genealogical data is extensive. For example, children being born before their parents were born, husbands marrying wives before their wives were born, people being born in countries that did not exist at the time, etc. are all too common.

Let us return to Henry Pitt. He was not alone in the world; he had a family. He had a mother, a father, siblings, a wife, and children. What information can be gleaned when taking his family into consideration?

One of the more certain events that can be automatically gleaned is that Henry Pitt's mother was present when Henry was born. By knowing that Henry's mother was present at Henry's birth, we now know that in addition to Henry's mother's vital records she also was in Bristol on October 7, 1613 (when Henry was born).

Taking this one step further, but with less certainty, there is a chance that Henry's father was present. If he was there, then we know a little bit more about Henry's father. What about Henry's maternal grandmother? Historically, a woman's mother might be present to help with the birth of a new child.

Did Henry have siblings that were already born? Who else was also there at Henry's birth? If we ease the exact date, who else was *probably* there around Henry's birth? Who else *might* have come to visit the new born baby within a few weeks of the birth?

Even in today's world of constant travel and children moving far distances from their home, *The New York Times* recently reported that the typical American still lives on average only 18 miles from their mother [4]. Today's world reflects our past: for the most part people live close to their parents and tend to visit immediate family.

In order to analyze all of these possibilities, and many more, we built a customizable prototype to visualize where a person *may* have been during his or her life based on the fact that people live in families and often attend important family events of their immediate family.

4 Prototype

Our prototype is highly customizable on what is considered the immediate and extended families. Fig. 3 is a screenshot of the configuration panel of the tool. All of these configurable assumptions help the genealogical analyst understand where an individual most likely was during their lifetime.

Fig. 3. Screenshot of the defaults for the customizable panel. This panel enables the user to configure what is considered the immediate and extended families. It also allows the user to select what events that the immediate and extended family would participate in.

Based on the configurable options of who would attend what events, our visualization prototype shows dozens of additional possible places of where Henry Pitt likely was during his life as opposed to only three events (including his marriage) when only vital records are taken into account.

Even though it is impossible to know where people lived and called home during their lives, these uncertain, but probabilistic assumptions, help us form stories of our ancestors. The key of moving past the basic vital records is that in order to understand the past the genealogical analyst must make assumptions.

If an analyst visualizes all the events for all the known data for a generation, then we can answer the more important question of not where an individual called home, but where families called home.

For example, given our scenario that Henry Pitt was born in Bristol, UK and died in Isle of Wight County, Virginia, US, what additional information can be gleaned? Visualizing Henry Pitt's family (his grandparents, parents, himself, his spouse, and his children), one can see that the majority of his life was *probably* spent in Virginia with the rest of his family.

Of course, it is possible that he lived far away from Virginia and constantly traveled there for his immediate family events. It is also possible that he was completely estranged from his wife and his 10 children and was never at any of their births. This, however, is highly unlikely. Many possibilities arise, which are left for the genealogical analyst to ascertain.

However, the first law of geography states, "Everything is related to everything else, but near things are more related than distant things" [14] Applying this to genealogy, one could easily state that families in the same immediate families in a given generation usually lived closer together.

4.1 Overview

Before beginning, the prototype preprocesses the genealogical file in order to extract all latitude and longitude information about the locations that will be visualized.

When the prototype begins, after the one time preprocessing step, it reads in a GEDCOM file. The user then chooses a focus person from an alphabetical list.

After the user chooses the focus person, the algorithm adds all the selected people that are in the current configuration. For example, if siblings are part of the immediate family, the algorithm adds all children of the focus person's parents. (This can be more advanced based on divorce and half/step sibling configuration). If grandchildren are part of the immediate family, then all of the focus person's children's children are added and so on throughout all the selected parts of the configuration panel.

After the immediate family of the focus person is determined, the algorithm goes through all of the events of the immediate family and the focus person and adds the events to the visualization. The events that are added are based on the current configuration.

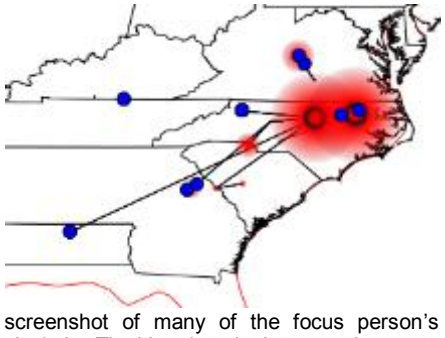

Fig. 4. A screenshot of many of the focus person's ancestors' movements in 1795. The blue dots depict several ancestors with lines trailing them to show their movement.

We have found that showing the movement of people as they move from event to event is required so that the user can make sense of the time sequence. When the blue dot (movement of the focus person) is moving from event to event, it does not represent where the person actually was. For example, a user might say, "The person was born here, went there with his mother to this other event, then went here …" The animation of people as dots represents the *likelihood* of actual position, which helps the user's cognitive mental model understand their movement in time.

Fig. 4 shows an example of many of the focus person's ancestors movement in 1795. One of the weaknesses of the visualization is that a blue dot could represent one person or whole families moving together. In order to see exactly who is being depicted the user would have to click for additional information.

Exactly what relatives of the focus person should be shown? Following genealogical visualization precedent, the default is to only show the direct line of the focus person. For example, for the most recent years of the dataset, only the focus person and parents of the focus person would be shown as blue dots. The focus person's sibling, cousins, aunts, uncles, etc. would not be shown as separate dots.

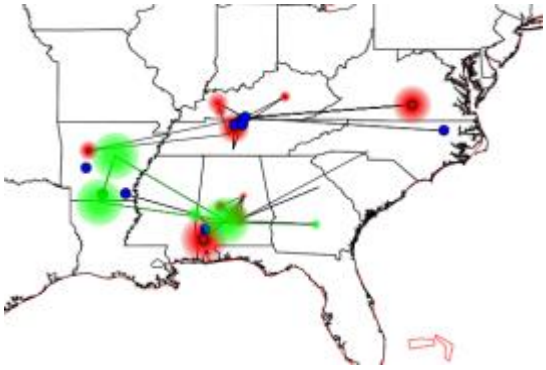

Fig. 5. Example of a single person's locations highlighted in green.

Visualizing all relatives does not help people form their personal narratives. People are generally interested in understanding who their direct line is – e.g. parents, grandparents, grand-parents, etc. – and not as much on other people like aunts/uncles, great-aunts/uncles [6,11].

Fig. 5 shows a selected person with the fade option to hide old events. When selected, all of the events that the selected person probably attended and the selected person's movement lines are shown green.

On the left side of the prototype are additional details for each person including how they are related to other people in the dataset – see Fig. 4. The addition shows textually all the events that a person likely attended.

4.2 Time – Overview and Specific Years (Fade)

There are two main user interface components that control time. First, there is a slider that controls what the currently depicted time is. For example, in the example dataset being shown, the slider ranges from 1613 (Henry Pitt's birth) to 2006 (the last recorded birth of a grandchild of the focus person in the dataset).

Second, there is a "fade" option. The fade option allows the user to stop the prototype from visualizing all events. For example, if the fade option is selected and is set to 30 years and the time slider is set to 1750, then only events that occurred between 1720 – 1750 are shown. If the fade option were not selected and the time slider is set to 1750, then all events from the beginning of the dataset to 1750 would be shown.

Regardless if the fade option is selected or not, the direct line of ancestors of the focus person are shown as blue dots at the most likely places for the specific time. People that were not alive, either not born yet or dead, during the depicted year set by the time slider are not shown.

4.3 Location of Living Relatives Based on Depicted Year (Blue Dots)

We depict people that are currently alive during the depicted time period as blue dots. When a person moves we depict a line between the event they came from and their current location.

Where exactly was a person at exact times? The only high probable locations that can be ascertained are based on actual events.

If we only show blue dots at high probable locations, then the user interface becomes highly confusing for the user. A blue dot that is on one side of the map then suddenly at another is essentially meaningless when dozens of ancestors are being depicted at once.

The algorithm we use for showing location of living people during a designated time is based on discontinuous path events. The algorithm links events based on time and is a generalization. It is *only* truly probable when the animation time corresponds to the exact event.

For example, if all that is known about an individual is a birth of 1800 at location X and a death of 1850 at location Y, then between 1800 and 1850 the individual's blue dot will slowly be moved between the two points with a line from location X to the current location of the blue dot. When time == 1825 the blue dot will be half-way between location X and location Y. When time >= 1850 then the circle location of location Y will exist and the line between X and Y will be complete.

We have found that showing the movement of people as they move from event to event is required so that the user can make sense of the time sequence. However, the reader should note that when the blue dot is moving from event to event that it does not represent where the person actually was. For example, a user might say, "The person was born here, went there with his mother to this other event, went here …" The animation of people as blue dots represents the *least likelihood of actual position*, but helps the user's cognitive mental model understand their movement in time.

Fig. 4 shows an example of many of the focus person's ancestors movement in 1795. One of the weaknesses of the visualization is that a blue dot could represent one person or whole families. In order to see exactly who is being depicted the user would have to click for additional information.

4.4 Selecting an Individual Person on the Map

The ability to see an overview of a family movements through time is crucial to understanding the narrative of the family. Equally important is the ability to see how one person relates to the rest of the family.

Fig. 6 shows a selected person with the fade option to hide old events. When selected, all of the events that the selected person probably attended and the selected person's movement lines are shown green and drawn last (so that they can be seen without being covered).

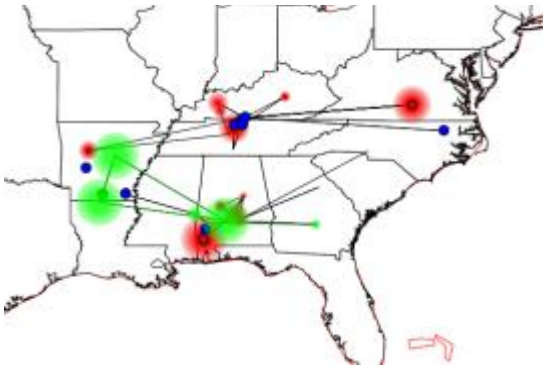

Fig. 6. Example of a single person's locations highlighted in green.

4.5 Additional Details

On the left side of the prototype are additional details for each person including how they are related to other people in the dataset – see Fig. 7. Besides the addition of showing all the events that a person likely attended, this side of the prototype is typical of most genealogical tools, so there is not much too comment on.

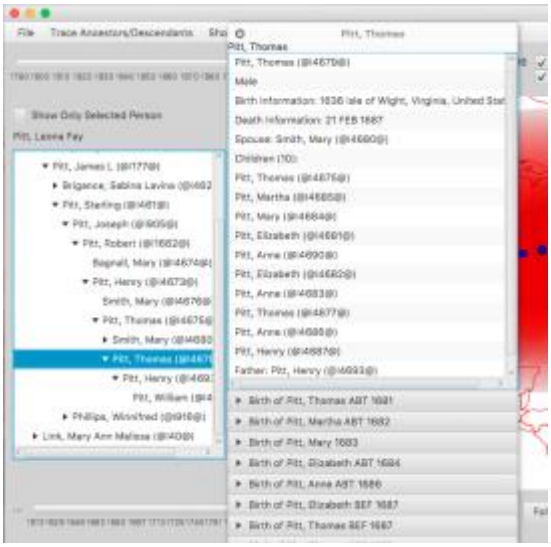

Fig. 7. Example screenshot showing all of the events of Thomas Pitt, Henry Pitt's son. Details include the information from the gedcom file, like gender, date of birth, spouse information, etc. In addition, all the likely events that he attended are included.

5 Challenges and Future Work

For our prototype, we shared some of the biggest challenges that currently exist in genealogical tools. First, automatically parsing dates. The following are some example of dates used by people: <1735>, Abt. 1735, Abt 1735, 1734-1735, around 1735, 1 May 1735, etc.

Second, automatically parsing locations that search engines, such as Google, do not even understand (e.g. Colony of North Carolina, British Colonial America). Since many places have changes names, or do not exist any more, parsing locations can be difficult. Locations that no longer exist is a reason why we advocate using historically-correct political maps.

Third, our visualization is only as good as the data. We share the challenge of getting accurate data. However, by using our prototype to visualize genealogical data we have already helped genealogical professionals find previously undetected errors in data. For example, we have helped professionals detect errors of deceased persons being born in countries that did not exist at the time.

The greatest weakness of this prototype is that if the date and the location of the individual's events are not known then it cannot be visualized. Heuristics can be used to determine approximate dates and locations, but at that level of uncertainty there is too much guesswork and is not shown.

For future work, our greatest need is to add different file formats for additional information. For example, being able to add information from tax records, land records, etc. would increase the narrative told and decrease the amount of uncertainty about the individuals depicted.

In addition, in the future we would like to be able to to add culturally specific events, such as quinceañeras. A quinceañera is a Hispanic cultural event when a girl becomes a woman at the age of fifteen.

6 Conclusion

In this paper we presented a new way of visualizing our ancestors lives. We created a highly customizable prototype that shows both overviews of entire families over hundreds of years and individual movements of individual people.

The overviews provide emergent patterns that show the epicenter or epicenters of where families lived during a time period. By "lived" we mean where families spent the majority of their lives.

As soon as one leaves the certain world behind and explores uncertainty then a wealth of information can be gathered and new patterns and insights can be gleaned. Instead of simply knowing when a person might have been born and died, we look at many likely possibilities.

We explore simple and obvious probabilities like how mother's were at their children's birth. We also explore how families in the past usually lived close by each other and would visit each other, especially for significant events like children's births, marriages, and deaths.

The following summarizes how our prototype is different from current geospatial genealogy visualization:

- Shows historically-correct maps for easy-to-identify anachronisms. An example anachronism is a person being born in "Iowa, USA" before Iowa was a state.
- Based on customizable parameters showing likely places of where the person was during their lifetime even with limited information by using additional information about relatives.
- Shows overlapping places based on opacity creating a heat map of the most lived-in areas.
- Shows likely relative size of importance of places.
- Shows overview first with the ability to drill down for additional information.
- Interactive so that the person can further analyze the data. The user is given the ability to see the entire dataset at once or move through the data with varying degrees of configurability.
- Shows relatives that are normally ignored in traditional visualizations in order to understand entire families.

References

[1] Andrienko, N. and Andrienko, G. Designing visual analytics methods for massive collection of movement data. *Cartographica* 2007; 42(2): 117-128.

[2] Ball, R. Visualizing Genealogy Through a Family-Centric Perspective. *Information Visualization Journal*. Vol 16, Issue 1, 2017.

[3] Bezerianos A, Dragicevic P, Fekete J, Bae J, and Watson B. GeneaQuilts: A System for Exploring Large Genealogies. IEEE Transactions on Visualization and Computer Graphics 2010; 16(6), 1073-1081.

[4] Bui, Q. and Miller, C. The Typical American Lives Only 18 Miles From Mom. Dec. 23, 2015, *The New York Times*. http://www.nytimes.com/interactive/2015/12/24/upshot/24up-family.html

[5] Dodge, S., Weibel, R., and Lautenschütz, A. Towards a Taxonomy of Movement Patterns. *Information visualization* 7.3-4 (2008): 240-252.

[6] Hackstaff, K. Family Genealogy: A Sociological Imagination Reveals Intersectional Relations. Sociology Compass, 2010; 4(8): 658 – 672.

[7] Herskovitz, A. A Suggested Taxonomy of Genealogy as a Multidisciplinary Academic Research Field. Journal of Multidisciplinary Research 2012; 4(3), 5-21.

[8] HistoryLines. https://historylines.com/

[9] An introduction to GEDCOM, http://www.tamurajones.net/AGentleIntroductionToGEDCOM.xhtml

[10] Kim, N Card, S and Heer, J. Tracing Genealogical Data with TimeNets. In: Advanced Visual Interfaces 2010, pp. 241-248.

[11] McAdams, D.P. Personal narratives and the life story. In L. A. Pervin & O. P. John (Eds), Handbook of personality: Theory and research, 3rd ed., New York: Guilford Press, 1999, pp. 242-262.

[12] Rootsmapper. https://rootsmapper.com/

[13] Saar, M. Genealogy and Subjectivity. European Journal of Philosophy 2002; 10(2): 231 – 245.

[14] Tobler W. (1970) "A computer movie simulating urban growth in the Detroit region". Economic Geography, 46(2): 234-240.