Solving Cycling Pedigrees or "Loops" by Analyzing Birth Ranges and Parent-Child Relationships

By: Alan B. Cannaday II
FamilySearch, 50 E. North Temple, Salt Lake City, UT

Abstract

When a person in a pedigree is discovered to be their own ancestor (and consequently their own descendent), an obvious error has been made. In turn, each person in the parental route, to relate that person to itself, can also trace themselves as their own ancestor, thus forming a cycling pedigree or "loop". Within the FamilySearch pedigree, over 750,000 persons are part of loops. Accordingly, in an accurate pedigree, loops would not exist. Though a large portion of the loops are of a size fewer than ten (approximately 10,000), and could be un-tangled by hand, the majority of the nodes are part of loops larger than ten and are difficult to detangle by hand. This includes the largest loop which is almost 700,000 strong.

The objective of this research (in progress) is to find a solution to detangling loops or help expedite detangling loops by hand while preserving as much viable information within the FamilySearch pedigree, giving patrons a more accurate, less confusing pedigree to work with. We do this by analyzing a combination of birth information, relationships, and inner persons to find candidates for splitting and un-realistic relationship deletion.

1. Forming a Cycling Pedigree

1.1 Inner Persons and Outer Persons

As explained above, a loop is detected when an individual in the pedigree is found to be their own ancestor. To explain the formation of loops we will look at the FamilySearch common pedigree (CP). In CP each "Person" starts out as an independent node with allocated information and relationships called assertions. Information that can be added as an assertion includes: birth, christening, death, burial, marriage, spouse, children, parents, etc…

When multiple independent persons in the database appears to be the same individual CP allows users to be able to merge theses persons into a unified person. This is accomplished using an inner/outer person data structure. The "inner" persons being the independent persons that the user merges together while the "outer" person serves as the representative to a set of individual persons that have been merged together. The outer person is then viewed by patrons as a summary of the information found in the inner persons.

Note: All information is stored in reference to inner persons. Even after a merge.

1.2 In the Beginning…

When the CP database was first established there were multiple sets of genealogical information that FamilySearch used as sources. These sources needed to be merged into one unified data set [1]. This was done by an early team in three steps:

1. Each individual in each source data set was first established as a new independent person, i.e., one "inner person" in its own "outer person".
2. Then, through a process called BULK_MERGE, all persons that were at least 99.5% likely to be the same person were then merged together. This was done by creating a new outer person to represent all persons being merged.
3. There was then a sequence of merges performed called RECURSIVE_FAMILY_MERGE, which looked at the relatives of those who had been merged by BULK_MERGE for cases where high confidence merges could be determined.

The CP database was then made available for patron use through new.familysearch.org. Checks were made during these process and guarantee that no loops where caused by BULK_MERGE or RECURSIVE_FAMILY_MERGE. Any existing loops at the time for public release were ingested by the initial sources.

1.3 Public Release and Causing Loops

Upon public release member patrons were able to create new persons, add assertions, and merge existing persons. Although, loops were speculated before the launch of new.familysearch.org, the first studies were complete in late 2009 and periodic studies have been done since (Figure 1). These studies show a growing trend of persons or records involved in loops.

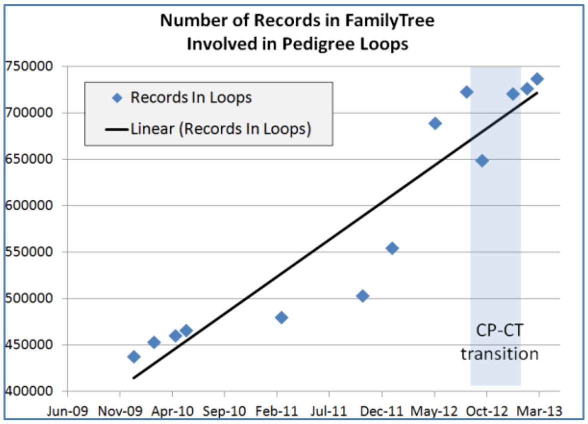

Figure 1: Shows the number of persons in the FamilySearch database involved in loops from late 2009 to March 2013. Also note the jump in records recorded in loops when this data shifted from being extracted from FamilySearch Common Pedigree (CP) to FamilySearch Common/public Tree (CT)

We discovered that some of the main causations for loops are the following:

1. Incorrect merges that are clearly wrong given the existing data of the persons considered for merging.
    - Example: Persons merged that were born 500 years difference
2. Incorrect parent-child relationships added.
    - Example: Persons born with a 500 year difference indicated as one or the others parent.
3. Rarely, person "hijacking", which can lead to incorrect merges and parent-child relationships
    - Example: A parent of children born in 1600 AD and later the parent having a birth date assertion of 1754 AD.

Due to the complexity of solving the hijacking issue, which would be an independent paper, the rest of this paper will mainly address solutions regarding the first two issues: improper merges of individuals with clearly contradicting information and incorrect parent-child relationships.

2 Using Birth Ranges to Analyze Relationships

2.1 Birth Ranges

As explained, each inner persons contains a set of possible assertions with information appertains to that specific inner person. One of these possible assertions is birth information. Birth information, specifically birth year can tell us the probable generation position of an individual in a pedigree and the probable birth information of relative. For example, it is common practice in genealogy to approximate a parent's birth year to be 26 years before the child.

As persons are merged, a birth range can be established for that person using the earliest and latest birthdates asserted into the inner persons. A typical birth range will have a relatively narrow window of within five years, but if persons have been poorly merges the birth range as been observed to be as wide as 600 years (Example: Figure 2).

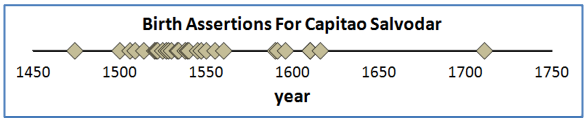

Figure 2: This is a real example of the birth assertions of a person in the common pedigree that has 158 inner persons merged together. The full range is about 150 years with a 60 year cluster between 1500 and 1560.

The birth-range can then be compared with the birth-ranges of parents and children to establish possible relationship legitimacy issues. The way the parent and child birth-ranges overlap can then be fit into six distinct types of overlap (Figure 3) which can establish further insight. For example, type 0 in Figure 3 shows the lack of overlap as the child birth-range entirely precedes the parent birth-range. This type indicates that the relationship has a high possibility of ill-legitimacy due to the fact that a parent's birth must precede a child's birth.

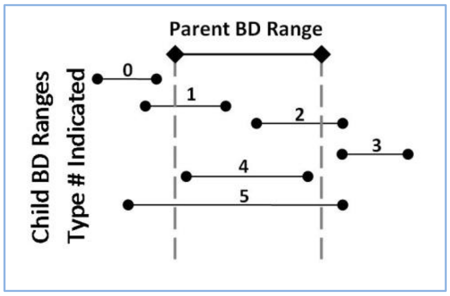

Figure 3: The types of birth-range overlap that can occur between parent and child. Child birth-ranges number designated to type is displayed above each child birth range.

2.2 Analysis of Relationships

The legitimacy for parent-child relationships can further be determined by using a probable age range for parenting children. Our range was established using some data from the FamilySearch database and an historical record data set. We determined that over 99.9% of all parenting ages fall between 15 to 65 years (Figure 4).

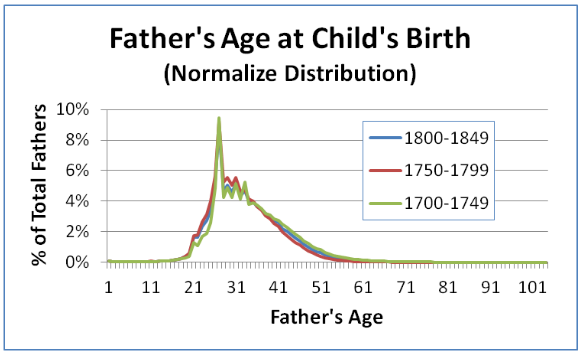

Figure 4: Shows graph of the normal percentage of a father's age at child’s birth in three 50 year time ranges. This allowed us to establish a range of probable legitimacy for parenting a child to be 15-65 years for men. Note the peak at 26 years due to estimated birth year difference for parents and children.

Using the birth-range and probable age range for parenting children, we then compare the boundaries for each birth range to determine legitimacy: earliest child BD to earliest parent BD, earliest child BD to latest parent BD, latest child BD to earliest parent BD, and latest child BD to latest parent BD. Knowing the type of overlap and which of the four comparisons previously discussed are returned as legitimate.

Each relationship is then categorized as follows (colors indicated correspond with sample images):

1. FULL_HIT (Green): All four of the comparisons indicated legitimate relationships.
2. PARTIAL_HIT (Black): Only some of the comparisons indicated legitimate relationships.
3. NO_HIT (Red): None of the comparisons indicated legitimate relationships.
4. UNDETERMINEABLE (Dark Green): Either parent or child have no birth year asserted.

The PARTIAL_HIT and NO_HIT categorized (except of overlap type 0) can then be analyzed using the full set of birth information to determine possible legitimate birth information contained in the inner persons. This is done by comparing each parent birth year assertion (taken from all the inner persons) to each child birth year assertion (again, taken from all the inner persons). A new category can then be added:

5. INNER_PERSON_HIT (Magenta)

This gives us an overall prospective of the legitimacy of the loops and where possible invalidities might occur within the relationships of the loops. Birth-range can also give us some additional information about the person as a whole.

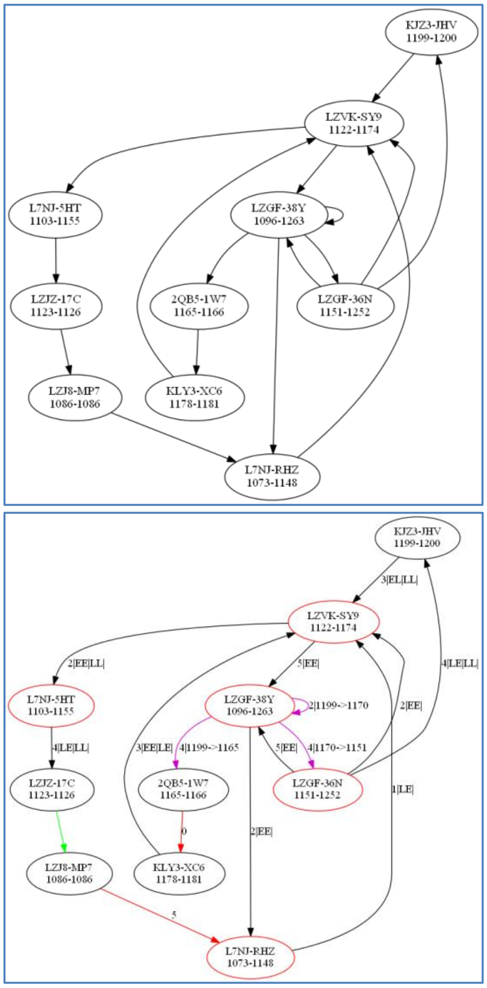

Figure 5: The top figure is real loop of size 10. Each node is an outer person with id and birth-range. Each edge represents a child-to-parent relationship. The bottom image is off the same loops after relationship analysis has been preformed. Edges are colored as indicated in list of categories in this section and labeled with the birth-range overlap type (Figure 3) with additional information. The indicator for PARTIAL_HITs represents child boundary then parent boundary that indicated legitimate relationship (EL means the Earliest child birth information and Latest parent birth information have legitimate). The indicator for INNER_PERSON_HITs indicates a possible legitimate relationship indicated by inner persons. Persons in red indicate a wide birth range.

3 Splitting Persons

3.1 Splitting Persons with Wide Birth Ranges

Due to bad merges, there are persons in the pedigree that have been merged that need to be split back apart into independent entities. Birth range can be used as an identifying tool to find good possible candidates for splitting (Note: It can be observed that individuals with wide birth ranges generally have high count of PARTIAL_HIT and INNER_PERSON_HIT relationships).

After splitting the nodes a disambiguation or clustering process can then be used to identify candidates to be remerged (Example in Figure 6). This can be done effectively by hand, using disambiguation software, or by reverting the inner persons clusters established before public release (Refer to section 2.1).

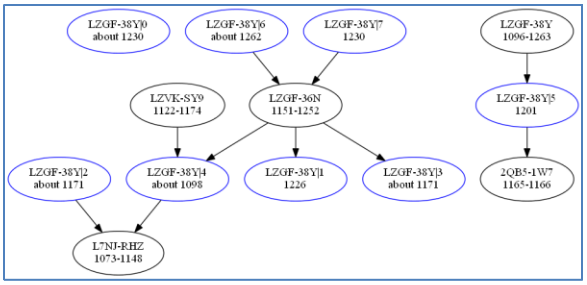

Figure 6: Split and re-clustering of, in blue, the inner persons of node LZGF-38Y (same as in Figure 5) with relationships corresponding to inner person assertions. All of our re-clustering is performed using the FamilySearch search.model.setFixer which was used in some of the original disambiguation during the BULK_MERGE process (Section 1.2).

If multiple persons in a loop can be identified as good candidates for splitting then the splitting and re-clustering can be performed for each person. In addition, by allocating relationships to the cluster which contains the appropriate inner persons followed up with the relationship analysis from section 2.2, loops will begin to untangle and the true legitimacy of the relationships become less obscure (Example in Figure 7).

Conclusion:

We believe that through further research, the loops in the pedigree can be mainly solved through means of birth-range and child-parent relationship analysis as described. This will provide a more accurate pedigree for patron while preserving as much accurate information as possible. Despite the fact that most of the loops are smaller and manageable to solve by hand, the majority of the nodes involved are in larger loops which is not a simple task to untangle.

By introducing some automated processes to analyze and tag problematic persons and relationships in the pedigree we will be able to expedite and decrease the work necessary to completely untangle loops.

Furthermore, forwarding research would include:

- Pooling the inner persons of split persons within a loop. It has been observed that within loops, set of re-clustered inner persons from different poorly merged persons could also be clustered.
- Agglomerative clustering of inner persons.

References:

[1] Wilson, Randy. “Bulk record linkage for Common Pedigree” (2009), <https://almtools.ldschurch.org/fhconfluence/display/Product/Bulk+record+linkage+for+Common+Pedigree>

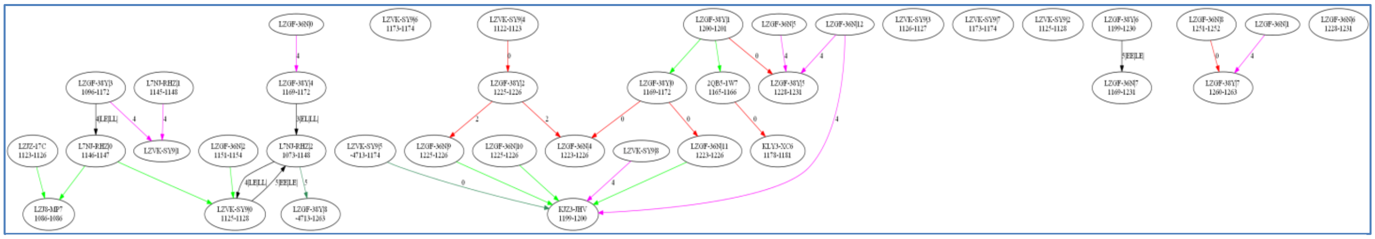

Figure 7: Split and re-clustering of all nodes with wide birth-ranges in Figure 5 with same system of labeling. Note that the only loops that remains consist of only two nodes, otherwise the loop is untangled. Also notable is that some of the relationships which we previously marked as INNER_PERONS_HITS are now marked as NO_HITS. *All clusters without relationships in this graph may have relationships unrelated to this loop.*