Beyond the Relationship Calculator:
Using a Weighted Relationship Distance Metric to Prioritize, Categorize and Visualize Relatives

Ben Baker
FamilySearch
bakerb@familysearch.org

ABSTRACT

When people undertake the task to research their genealogy, most begin with themselves and work upward generationally to their parents, grandparents, great-grandparents and so forth. After gaining more experience, people generally then branch outward to look at collateral and spousal lines to more fully represent and understand their family.

As the number of people in the user's family tree grows, it becomes more difficult to know where to work next and to retain the context of where all of their relatives fit in. Whether a user is trying to create a completely new family tree or working in a tree already containing thousands of relatives, tools do not currently exist to help users prioritize where to work next on a particular task, categorize relatives into groups and easily visualize relationships beyond immediate family and direct lines.

Most genealogy products include relationship calculators to aid users in determining relationships between two persons, even visualizing the relatives connecting the two persons for added context. However, it would be much more valuable to have a method to prioritize, categorize and visualize all of the relatives throughout a connected family tree, regardless of how they are related.

This paper introduces a unified metric called *Weighted Relationship Distance (WRD)* that computes a single distance value between any two persons connected through direct, collateral and spousal lines based on closeness of relation. This metric can be used for many applications including:

- Providing a method to prioritize which relatives across direct, collateral and spousal lines a user could systematically work on to accomplish a particular task
- Categorizing relatives across all lines into easily identifiable buckets. (Ex. List all 1st cousins twice removed)
- Visualizing relatives a person in a user's tree likely knew personally beyond immediate family

1. INTRODUCTION

In the 2005 Family History Technology Workshop keynote address, Ransom Love stated that over 60% of the adult population wants to be involved in family history. [1] Some of the key characteristics of a majority of these people include:

- Have little time
- Need to have quick success in a format they understand
- Want to find an ancestor and information about them

Similarly, in a study performed in 2008, 95.9% of 584 LDS church member respondents agreed or strongly agreed with the statement "I think that doing genealogy is very important". However, 84.6% of the same respondents reported that they spent less than one hour per month or no time at all doing genealogical research. [2]

Reasons why these respondents didn't do genealogical research included not knowing where to start and it taking too much time to do something useful. To help address these concerns, the same group later launched a 20-minute genealogist website which has now been redirected to http://kinpoint.com/. As stated on the main page, a main purpose of this site is "to help people spend less time figuring out what you need to do, and more time doing genealogy work."

Current genealogical software and systems have made it relatively easy to create and maintain databases of connected individuals. Tools exist to show relationships between persons in these databases as pedigree, fan, ring, family group and descendancy charts or reports.

However, two things missing from current tools include:

- Something that helps users easily know where to work next and what they can do to be productive.
- A simple way to understand the context of data within the system not possible in existing charts and reports.

Follow up research on 20-minute genealogist efforts found that "Technology for graph-traversal and finding who needs research was the primary need of respondents. There was also a significant need for an improvement in context-preservation software." [3] This is particularly true for people who are working in family trees where an extensive amount of research has already been done.

This paper proposes a metric that traverses a connected family tree and enables users to have a systematic priority order for relatives, so they can easily know where to work next. Context is also improved by providing a means for more easily understandable categorization and visualization.

2. PROPOSED METRIC

In order to provide a single distance value between any two persons and enable such applications, a metric is needed that can provide a distance value that unifies the concept of distance across direct, collateral and spousal lines.

The weighted relationship distance (WRD) metric is defined as a function of three distances (g, c, m) from a base person in a connected family tree graph, each with an associated weighting factor (α, β, γ) applied as follows:

$$WRD(g,c,m) = \alpha(|g| + 1)e^{\beta c}e^{\gamma m} \quad (1)$$

2.1 Distances

Generational Distance g is defined as the number of generations from the base person to another person, starting with the base person's generational distance of 0. Values may be positive or negative integers, indicating the generations preceding the base person as positive values and generations after the base person as negative. For example, the father of the base person would have a value of g = 1, a 2nd great grandmother would have a value of g = 4 and a child would have a value of g = -1.

Collateral Distance c is defined as the "horizontal distance" from the base person to another person. Collateral distance can also be defined as the number of generations to closest common ancestor. Values may be positive integers, starting with a value of 0 for the base person and direct ancestors and descendants. Examples include a sibling where c = 1 and a 1st cousin where c = 2. As values of g and c are combined, all blood relatives may be represented. For example, a 4th cousin 3 times removed has g = 3 and c = 5.

Marriage Distance m is defined as the number of marriages between the base person and another person. Values may be positive multiple of 0.5, where non-integer values indicate persons related through a non-blood spouse of a direct ancestor. Marriage distance serves to add all connected non-blood relatives such as in-laws and additional spouses of ancestors who are not related by blood. Examples include a wife where m = 1, a brother-in-law where m = 1 and c = 1 and an additional wife of the base person's great grandfather m = 0.5 and g = 3.

2.2 Distance Weights and Weighting Factors

After distance values g, c and m have been calculated between two persons, weighting factors are applied to create three distance weights, which multiplied together form the WRD metric.

Generational Distance Weight G is computed using the generational weighting factor α as follows. The result increases linearly.

$$G = \alpha(|g| + 1) \quad (2)$$

Collateral Distance Weight C is computed using the generational weighting factor β as follows. The result increases exponentially, resulting in higher values more quickly as collateral distance increases than for generational distance.

$$C = e^{\beta c} \quad (3)$$

Similarly, a *Marriage Distance Weight M* is computed using the marriage weighting factor γ as follows.

$$M = e^{\gamma m} \quad (4)$$

As the three distance weights are multiplied together to for the WRD metric, relatives throughout a base person's family tree receive values that can be sorted to provide a continuous range of values indicating the closeness of relation of any person connected by a set of relationships to the base person.

Different weighting factors may be used for different purposes, depending on the desired application. Table 1 shows example values for distances, distance weights and resulting WRD values for several of the closest relatives of a base person, sorted by ascending WRD value.

	g	c	m	G	C	M	WRD
Base Person	0	0	0	1	1.0	1.0	1.00
Parents	1	0	0	2	1.0	1.0	2.00
Siblings	0	1	0	1	2.7	1.0	2.72
Grandparents	2	0	0	3	1.0	1.0	3.00
Children	-1	0	0	3.52	1.0	1.0	3.52
Great Grandparents	3	0	0	4	1.0	1.0	4.00
Spouse	0	0	1	1	1.0	4.1	4.14
2G Grandparents	4	0	0	5	1.0	1.0	5.00
Grandchildren	-2	0	0	5.28	1.0	1.0	5.28
Aunts/Uncles (children of grandparents)	1	1	0	2	2.7	1.0	5.44
3G Grandparents	5	0	0	6	1.0	1.0	6.00
4G Grandparents	6	0	0	7	1.0	1.0	7.00
Great Grandchildren	-4	0	0	8.8	1.0	1.0	8.80
1st cousins	0	2	0	1	7.4	1.0	7.39
5G Grandparents	7	0	0	8	1.0	1.0	8.00
Great aunts/uncles (siblings of grandparents)	2	1	0	3	2.7	1.0	8.15
Parents-in-law	1	0	1	2	1.0	4.1	8.27
6G Grandparents	8	0	0	9	1.0	1.0	9.00
Nieces/Nephews	-1	1	0	3.52	2.7	1.0	9.57
7G Grandparents	9	0	0	10	1.0	1.0	10.00

Table 1 – Example values for distances, weights and WRD

In the table above, the default value used for $\alpha = 1$ if $g >= 0$, otherwise $\alpha = 1.76$. The effect of this is to increase the distance of descendants over ancestors because earlier generations are of more relevance for genealogical purposes. The different values also help differentiate between positive and negative values of g in resultant WRD calculations.

The default value used for $\beta = 1.0$ and the default value of $\gamma = 1.42$. The different values of β and γ help produce unique WRD values for persons with the same c and m distances and more heavily weight relatives through marriage as more distant than collateral relatives.

3. POTENTIAL APPLICATIONS

3.1 Prioritization

Perhaps the most important application of the WRD metric is to provide a method for family history software systems to create tools to help users prioritize relatives in a genealogical database in a systematic way. This prioritization could be applied in a variety of ways to a variety of tasks, including:

- Persons with few or no sources attached. Could also be paired with persons who should appear in a specific source (i.e. which of my relatives should be in the 1910 US Census, but I haven't found them yet?)
- Likely areas where more children exist but are not in the database yet (Ex. Couples with one or no children)
- System-suggested historical records to review (Ex. Ancestry record hints and MyHeritage record matches)
- End of line relatives – People without known parents, including those outside direct lines
- "Doneness" of a task up to a particular WRD threshold
- Persons with few or no photos/stories attached
- Possible duplicates that may require merging
- Facilitate LDS temple work for closest relatives first and sharing more distantly related people with others
- Detection of data anomalies where corrections are likely needed (Ex. Unlikely dates, looping pedigrees, many sets of parents, etc.)
- Find people with missing conclusions (Ex. who doesn't have a death/burial?)
- Sort a to-do list by closeness of relation
- Applications across a set of users (Ex. closeness of relation of all LDS temple submissions on FamilySearch)

To begin to show the potential of some of these applications, a prototype system was developed to interface with a RootsMagic database of nearly 24,000 connected persons, including the author. Output is simply dumped in csv format and loaded into Excel to demo preliminary results. Initial applications included information about areas where additional children may exist, parents are missing and where LDS temple work may be required.

Even in this minimal implementation, the author was able to indentify several families who were 1st cousins of his grandmother who were not in FamilySearch Family Tree and were good candidates for additional research, sourcing, and LDS temple work.

3.2 Categorization

Categorizing relatives across generational, collateral and spousal lines into more easily identifiable buckets is another potential application of the WRD metric. Relationship calculators exist in most genealogy software to show the relationship between two persons in a human-understandable fashion. (Ex. First cousin twice removed)

However, because all persons with the same relation to a base person have the same WRD value, all persons with that relationship may be shown as a group with the same descriptive text. A minimal implementation on the author's RootsMagic database enables seeing all relatives in a particular category group, but more user-friendly implementations in family history software could use the same method.

A potential simplification of WRD metric values would be to further group them into ranges and present them in a "star ranking" format from 5 stars down to 0. Table 2 shows a proposed grouping of WRD ranges that further simplifies groupings to help users as they gain experience to concentrate on more distant relatives.

WRD Range	Star Ranking User Description	Example Relatives Within Range
[0-5]	[5.0-4.5] Novice	Direct Ancestors – 5 gen Siblings, Spouse, Children
(5-10]	(4.5-3.5] Beginning Intermediate	Direct Ancestors – 10 gen Siblings of directs – 3 gen 1st cousins, nieces/nephews, parents-in-law, great grandchildren
(10-25]	(3.5-2.5] Intermediate	Direct Ancestors – 25 gen Siblings of directs – 9 gen 1st cousins of directs – 3 gen 2nd cousins Spouse's directs – 6 gen Siblings-in-law
(25-50]	(2.5-1.5] Advanced Intermediate	Direct Ancestors – 50 gen Siblings of directs – 18 gen 1st cousins of directs – 6 gen 2nd cousins of directs – 2 gen Spouse's directs – 12 gen
(50-100]	(1.5-0.5] Advanced	Direct Ancestors – biblical Siblings of directs – 36 gen 1st cousins of directs – 12 gen 2nd cousins of directs – 4 gen 3rd cousins Spouse's directs – 24 gen Spouse's directs' siblings – 8 gen Spouse's directs' 1st cousins – 3 gen
(100-200]	(0.5-0] Expert	Direct Ancestors – all Siblings of directs – biblical

		1st cousins of directs – 27 gen 2nd cousins of directs – 9 gen 3rd cousins of directs – 2 gen 4th cousins Spouse's directs – 48 gen Spouse's directs' siblings – 17 gen Spouse's directs' 1st cousins – 6 gen
>200	0	All connected relatives

Table 2 – Proposed WRD Ranges and Example Relatives

It may seem like limiting WRD values of concern to 200 would not include very many relatives. However, about 70% (16,830) of the nearly 24,000 persons in the author's personal database had WRD values less than or equal to 200. There are also likely many more persons yet to be discovered with WRD values under 200 that are not yet recorded in this database. In addition, many of the persons with higher WRD values were in the database to show distant relationships to famous people.

Categories may also be used to perform tasks across a set of persons more quickly. For example, in FamilySearch Family Tree it is possible to place a watch on a person to receive e-mail notifications when changes are made by others. Instead of applying watches one by one, all persons with a particular relation or even star ranking could be watched in a single operation.

3.3 Visualization

Many genealogical software systems include visualizations of relationships between persons in relationship calculators. However, it is usually not possible to see relatives of a particular person beyond immediate family and/or direct ancestors and descendants without multiple views. Widening the relatives shown may be useful for genealogical research leads as well as to better understand the bigger picture a relative lived in to include people they likely knew.

While the initial WRD metric is well suited to prioritization for many genealogical purposes, the exponentiation of the collateral and marriage distances can make some relatives more distant than is perhaps desired. In civil law, degrees of kinship are used to determine "next of kin" in the event of the death of a person without a will. [4]

A simpler metric similar to degrees of kinship may be more effective to determine relatives a person likely knew. If a simple sum of the absolute value of the three distance values is used, a Simple Relationship Distance (SRD) may be obtained as follows:

$$SRD(g,c,m) = |g| + c + m \quad (5)$$

Table 3 shows a sample of some of the relatives that would be included in the first three SRD values. Visualization of these relatives could be as simple as showing them in tabular form in a report. A more visual method for visualizing pedigrees that includes collateral and spousal relations was presented by Mike Miller of Branches Genealogy at RootsTech 2012. [5] Using a visualization method such as this would make it much easier for users to see the context that an ancestor lived in.

SRD	1	2	3
Example Relatives	Parents Children Siblings Spouse	Grandparents Grandchildren 1st Cousins Aunts/Uncles[1] Nieces/Nephews Parents-in-law Siblings-in-law[3]	Great Grandparents Great Grandchildren 1st Cousins Once Rmvd Aunts/Uncles[2] 2nd Cousins Great Aunts/Uncles Grandparents-in-law Siblings-in-law[4]

Table 3 – Sample Relatives with SRD <= 3

1 – Children of grandparents
2 – Spouses of children of grandparents
3 – Siblings of spouse and spouses of siblings
4 – Spouses of siblings of spouse

4. CONCLUSIONS AND FUTURE WORK

A metric has been proposed that produces a single distance value for all persons in a connected family tree. An initial prototype has been developed to work with a RootsMagic database and shows a few of the promising applications of this metric.

Additional work is planned to further realize additional benefits of the various potential applications of the WRD metric. Due to performance limitations of the current FamilySearch developer API, it is anticipated that in order to make the WRD metric useful in FamilySearch applications, a move to a NoSQL database such as MongoDB as proposed in [6] is likely necessary. The author hopes to see this proposed work implemented at FamilySearch and intends to work with desktop software companies to help implement applications described in this paper.

5. REFERENCES

[1] *Technology: Shifting the Genealogical Paradigm and Realizing the Vision.* **Ransom Love.** 2005 Family History Technology Workshop Keynote http://fht.byu.edu/prev_workshops/workshop05/FHTCD/keynote-RansomLove.ppt [Accessed Mar 2013]

[2] *The 20-Minute Genealogist: A Context-Preservation Metaphor for Assisted Family History Research.* **Charles D. Knutson, Jonathan Krein** **http://sequoia.cs.byu.edu/lab/files/pubs/Knutson2009.pdf** [Accessed Mar 2013]

[3] *Difficulties of 20 Minute Genealogists.* **Mike Nelson** *http://www.academia.edu/529073/Difficulties_of_20_Minute_Genealogists* [Accessed Mar 2013]

[4] http://www.mystatewill.com/degrees_of_kinship.html

[5] *Visualization of Genealogy Data*. **Mike Miller**. RootsTech 2012

[6] *HumMONGOus Family Tree Application Running on a MongoDB Cluster – A Case Study.* **Randy Bliss, Judson Flamm, Randall Johnson, Tom Valletta.** SORT Conference 2012