Date Range Propagation in Genealogical Databases

Randy Wilson
FamilySearch.org

Abstract. *Genealogical data is rarely complete on a given individual in a particular source. A birth certificate, for example, will have a birth date for the child, but not a marriage or death date; nor, typically, will it have any dates for the father or mother. However, even though there is no explicit date listed in a record for an individual's birth or death events, reasonable ranges can often be calculated from the available information. These estimated ranges can be very useful in improving accuracy when calculating match scores between a pair of individuals; in ruling out unreasonable search results; and in calculating which individuals in a database might reasonably still be alive. This paper discusses how to calculate estimated birth, marriage and death ranges on individuals based on available information. It also shows how to propagate this knowledge across generations when necessary. It concludes with examples of how these techniques have been used in practice and with a brief introduction to how propagation of probability distributions could improve accuracy further.*

1. Introduction

Genealogical data can be found in many sources—birth, marriage and death certificates; census records; wills and probates; tombstones; family bibles; journals; books of remembrance; web pages; compiled databases; and so on. Each appearance of a person in a source can contain a different set of information about that person. One appearance in a census might provide information about a person's parents and siblings. Another appearance in a marriage record might provide information about the person's spouse and marriage date.

It is rare, however, for any source to paint a complete picture of an individual. In such cases, it can often be useful to estimate reasonable boundaries for the missing data. A person with an unknown birth place, for example, who is known to have been married in Tennessee, is sometimes listed with a birth place give as "of Tennessee." This provides a hint that the event place is not known, but has a reasonable chance of being in or near this location.

The same thing can happen with dates as well, where one record might have a birth date and another a death date. Or one might have a child's birth date and another a spouse's death date. While there is no pair of corresponding dates available for direct comparison in such cases, there is often enough information to make useful inferences.

This paper focuses on the problem of estimating date ranges from the dates available in a record. While it is not usually possible to fill in such data with certainty, estimated date ranges can be useful in a variety of situations.

Matching. When matching two individuals (i.e., using a *record linkage* algorithm), it is common to compare birth dates, marriage dates and death dates between individuals. However, when one record has a birth date and the other has a death date the algorithm cannot directly compare corresponding dates. What it can do, however, is determine whether one date is reasonable, given the other one. For example, if one individual was born in 1903, and the other died in 1850, then they cannot be the same person, since a person cannot die before they were born.

Searching. Related to doing careful scoring of two individuals is the problem of finding relevant matches in a large set of individuals given a search query. When dates are included as search terms, date propagation can be used to filter out "unreasonable" individuals. For example, if a search is done for John Abbott born in 1903, then it would be unreasonable to include individuals who died before that time.

Determining living status. It is often necessary to distinguish between data about living individuals, whose information must be kept private, and deceased individuals, whose data can typically be shared. Many individuals in genealogical databases do not have a death date listed, even though most individuals in these databases are in fact deceased. By using the known dates in a database, we can determine which individuals could reasonably still be alive (and thus protect their information) while allowing the information about clearly deceased individuals to be shared.

This paper explains how to use *date range propagation* to estimate date ranges that can be used in each of these situations to improve accuracy. Section 2 describes how date range propagation works and how data was gathered to make it possible. Section 3 gives some examples of how date propagation has been used by FamilySearch in practice. Section 4 proposes some further work involving the propagation of probability distributions instead of flat ranges to

improve accuracy when ranges are propagated across many generations.

2. Date range propagation

Date range propagation is the process of estimating a reasonable *birth*, *marriage* and *death*year range for each individual in a lineage-linked genealogical database from known dates and relationships in the data. Consider a relationship graph G with n persons $p_1..p_n$. Each person has a *gender* (*male, female* or *unknown*), and can have *relatives* of type *father*, *mother*, *spouse* and *child*, each of which is another person in G. Each person also has events of type *birth*, *christening*, *marriage*, *death*, and *other*. (Burial events are treated as death events, though given a lower priority when a death event is listed; and *other* events include residence, census, or any other events that indicate that a person was alive at a given time).

Each event can be *specific*, meaning that the day, month and year are specified (e.g., "3 Feb 1820"); or *year-only*, meaning that there is only a year (or approximate year) given for the date (e.g., "1852" or "about 1875"). In general, specific dates tend to be more reliable, while year-only or estimated dates are often off by several years, and so result in less precise estimates.

Each *estimated range* is calculated from that person's own event(s) of that type; that person's other events; and the events of that person's direct relatives. This is done by applying a *delta* (which contains a *min* and a *max*) to a known year or previously estimated range in order to produce another estimated range. Both deltas and year ranges are notated in this paper using the convention *min..max*.

When more than one piece of evidence is available to estimate a range, the estimated ranges derived from the various available events are intersected, resulting in a narrower range that is consistent with all of the available data.

2.1 Deriving date deltas from known data

Some work has been done previously on date range propagation (Vox & Randell, 2000; Despain, 2001), but the ranges used have typically been hand-picked from heuristics. In this work, the *deltas* were derived from observed data as follows. A database of 15 million individuals was examined for cases where pairs of events had dates present in the data. Whenever one of the event types being estimated (i.e., an individuals birth, marriage or death) had a date, and one of the events used as evidence (e.g., father's birth) also had a date, then the difference between the years was calculated and added to a set of delta values for that case.

There were several dimensions across which the cases were segregated, resulting in a 5-dimensional array of cases. These dimensions are as follows.

1. ***Target event***: *birth, marriage, death (if single), death (if married).*This is the event whose year is being estimated. Birth, marriage and death are the three event types being estimated by date range propagation. However, the reasonable death range is different for those who have lived to adulthood than those who have not. Therefore, if a person has a spouse, children or a marriage event, they are said to "look married", and the stats were collected separately from those who "look single".
2. ***Relative type****: individual, father, mother, spouse, child.* Note that "individual" (i.e., "self") is one of the "relative types" that can have event dates that contribute to a date range calculation. This allows the algorithm to deal with a person's own events in the same way as relatives' events. For example, and individual's birth year can be used to estimate their death year.
3. ***Source event****: birth, christening, marriage, death, other*. This is the set of events used to estimate date ranges for the target event. Burial events are included under death events, though a burial event is ignored if a death event is present for an individual. "Other" events include residence, census, or any other event that indicates that a person was alive on a given date.
4. ***Gender****: male, female, unknown.* The gender of the "target" individual whose ranges are being estimated often makes a difference as to what is reasonable, due to differences in life expectancy, common marriage age and child-bearing years. Deltas for the "unknown" case were gathered by including data for male, female and unknown gender cases, but are used only when the target individual's gender is unknown.
5. ***Exactness****: specific, year-only.* A date is said to be *specific* only if it has a day, month and year. For the *source event*, specific and year-only events were put into two separate groups, and it was indeed found that year-only events had a wider distribution than specific events, because they were more likely to be estimated and thus incorrect. (Only target events with specific dates were used in gathering statistics, because the specific dates were more representative of the "truth" that the statistics were trying to model.)

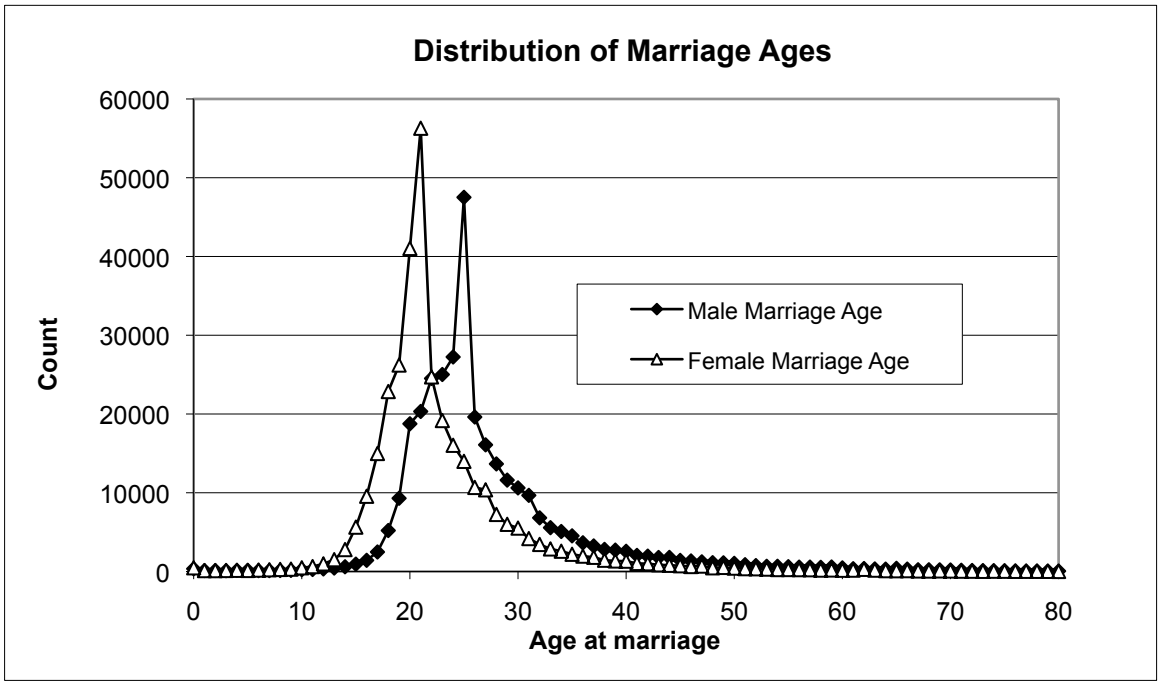

Figure 1. Distribution of age at marriage for male and female.

Initially, "about" dates were put into a third group, with the assumption that these would have even more variation than year-only dates. Surprisingly, however, "about" dates tended to have a slightly *tighter* distribution than year-only dates. Apparently, when someone takes the trouble to say "about", they are making less of a wild guess than when they only include the year. It is also possible that year-only dates were originally estimated by some earlier algorithm. Therefore, year-only and "about" dates were both grouped into the "year-only" category.

For each target event that had a specific date, the delta value was calculated between that date and any available source event, and the appropriate counter was incremented. Counters representing delta values from -100 to +100 were used for each case, along with a "less than -100" and "greater than +100" value.

Due to errors in the data (typographical errors, bad merges, incorrect conclusions, etc.), there were biologically impossible results in the data, such as individuals dying long before they were born. There were also probably some possible but extremely rare occurrences, such as a father having a child at age 85, and that child living to be 105 years old.

To make the results useful in practice, the top and bottom 1% of the data were dropped from each distribution in order to come up with a range that could be used in the date range propagation.

As an example, Figure 1 shows one of these distributions, namely, the distribution of delta values between marriage year and birth year for males and females (using specific source dates). In this case, the deltas correspond to ages. Note that the most common age of marriage is 25 for males and 21 for females. (As noted above, due to data errors there are delta values down to 0 and beyond; and off the right side of the chart as well).

Given this data from this distribution, and dropping the outliers, we get the following two delta ranges:

delta(*birth, individual, marriage, male, specific*)
= 17..63

delta(*birth, individual, marriage, female, specific*)
= 14..52

meaning that a male is usually (i.e., 98% of the time) between ages 17 and 63 (inclusive) when he is married, and a female is usually 14 to 52 years old when she is married.

Figure 2 shows another, related distribution, namely, the distribution of deltas between a female individual's birth and their spouse's (i.e., husband's) birth. Note that the most common difference is 4 (meaning the husband is 4 years older than the wife), which agrees with the difference seen in Figure 1 between the most common age at marriage for males (25) and females (21). Dropping the highest and lowest 1% of the data yields delta ranges of:

delta(*birth, spouse, birth, male, specific*)
= -10..24

delta(*birth, spouse, birth, female, specific*)
= -24..10

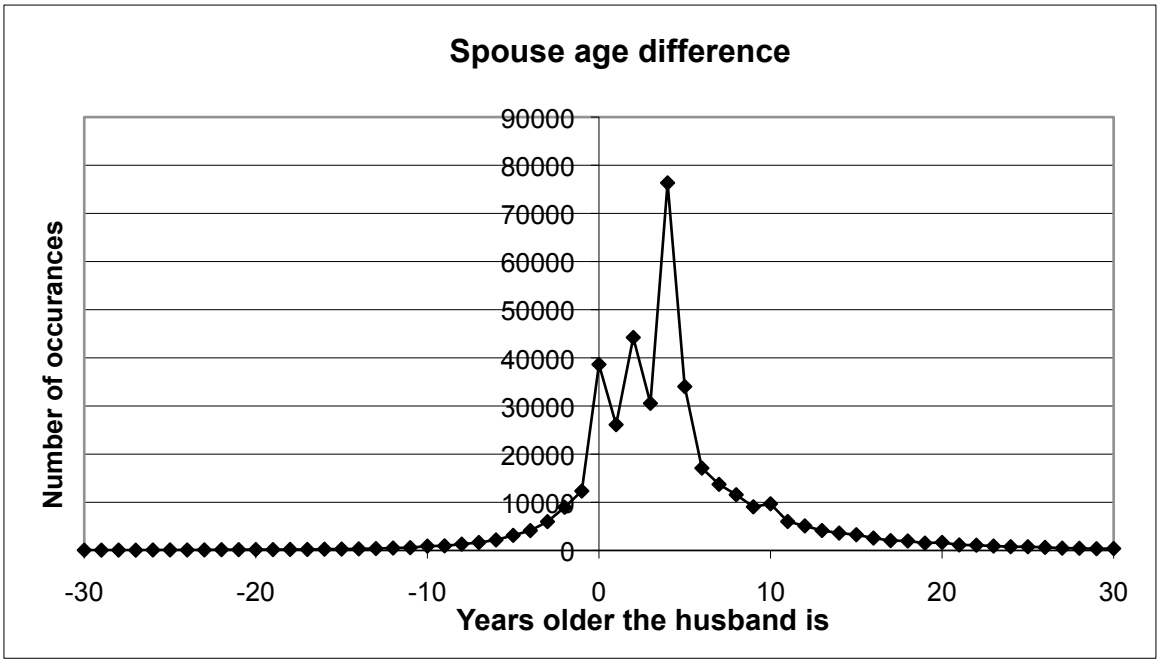

Figure 2. Distribution of deltas between a female individual's birth and her husband's birth, *i.e.*, distribution of age difference between a wife and her husband at the time of marriage.

meaning that the husband is usually between 10 years younger and 24 years older than his wife.

As a final example, Figure 3 shows the distribution of deltas between a birth year and death year (for males, using specific dates). Note the spike in the first few years indicating a historically high rate of infant mortality.

In addition to dropping outliers, it was also decided that the deltas for an individual's own birth and death based on their own birth and death would be forced to be 0, even though some self-contradictory data might have suggested otherwise. For example, given two individuals with year-only dates that are 5 years apart, it actually is not impossible that these refer to the same person, because each date could be an estimate that is off by a few years. However, it was decided to allow a matching algorithm to account for this variability rather than allow the date range

Figure 3. Distribution of deltas between birth and death for males, using specific dates.

			Year-only date						Specific (day, month, year) date					
myEvent	**Relative**	**Event**	**Male**		**Female**		**Either**		**Male**		**Female**		**Either**	
Birth	Ind	Birth	0	0	0	0	0	0	0	0	0	0	0	0
Birth	Ind	Chr	-6	44	-7	36	-6	42	0	5	0	4	0	5
Birth	Ind	Death	0	97	0	100	0	99	0	92	0	94	0	93
Birth	Ind	Marriage	12	64	8	55	10	60	17	63	14	52	15	60
Birth	Father	Birth	-62	-14	-61	-14	-61	-14	-56	-19	-56	-19	-56	-19
Birth	Father	Chr	-63	21	-62	25	-63	22	-61	-1	-60	-3	-60	-1
Birth	Father	Death	-6	70	-5	69	-5	70	-1	65	-1	65	-1	65
Birth	Father	Marriage	-30	17	-29	16	-30	16	-25	5	-25	5	-25	5
Birth	Mother	Birth	-52	-10	-51	-10	-51	-10	-45	-17	-45	-17	-45	-17
Birth	Mother	Chr	-51	32	-53	32	-52	32	-46	0	-47	-1	-47	-1
Birth	Mother	Death	-5	75	-5	75	-5	75	0	70	0	70	0	70
Birth	Spouse	Birth	-12	29	-29	12	-24	24	-10	24	-24	10	-21	21
Birth	Spouse	Chr	-13	38	-31	31	-27	35	-14	39	-27	29	-25	35
Birth	Spouse	Death	21	110	18	100	19	107	25	106	23	93	24	103
Birth	Child	Birth	10	60	7	51	8	57	18	55	16	44	16	52
Birth	Child	Chr	17	63	14	54	15	59	18	61	16	52	16	59
Birth	Child	Death	24	139	21	133	22	136	23	136	20	129	21	133
Birth	Child	Marriage	28	101	25	95	26	98	37	94	34	88	35	92

Table 1. Deltas for calculating estimated birth date ranges from relatives' events.

propagation to take a known date and widen it.

Similarly, while a person can indeed be married more than once, it was decided to not allow a source marriage event on an individual to generate an estimate that was wider than the original marriage event.

Table 1 shows the deltas computed from the empirical data for calculating the estimated birth range; Table 2 shows the deltas for calculating the marriage ranges; and Table 3 shows the deltas for computing the death ranges (separated between results for those who "look married" and those who do not).

			Year-only date						Specific (day, month, year) date					
myEvent	**Relative**	**Event**	**Male**		**Female**		**Either**		**Male**		**Female**		**Either**	
Marriage	Ind	Birth	-64	-12	-55	-8	-60	-10	-63	-17	-52	-14	-60	-15
Marriage	Ind	Chr	-69	19	-56	16	-65	18	-66	0	-54	0	-62	0
Marriage	Ind	Death	-2	78	-1	81	-1	80	0	70	0	75	0	73
Marriage	Ind	Marriage	0	0	0	0	0	0	0	0	0	0	0	0
Marriage	Father	Birth	-106	-34	-96	-30	-102	-31	-100	-40	-91	-37	-97	-38
Marriage	Father	Chr	-107	-6	-99	2	-106	-1	-107	-24	-97	-22	-104	-23
Marriage	Father	Death	-50	46	-42	50	-47	48	-44	43	-34	46	-41	44
Marriage	Father	Marriage	-78	-2	-68	0	-75	-1	-74	-14	-64	-11	-70	-12
Marriage	Mother	Birth	-100	-30	-90	-26	-96	-28	-94	-37	-84	-34	-90	-35
Marriage	Mother	Chr	-99	-2	-92	3	-97	1	-101	-18	-89	-12	-97	-15
Marriage	Mother	Death	-51	51	-43	54	-48	53	-42	48	-33	51	-38	50
Marriage	Spouse	Birth	-61	1	-66	-9	-64	-2	-60	-6	-65	-15	-63	-9
Marriage	Spouse	Chr	-64	19	-72	19	-69	19	-62	4	-68	2	-66	3
Marriage	Spouse	Death	-14	83	-11	78	-13	80	-10	78	-7	71	-9	76
Marriage	Child	Birth	-34	31	-27	27	-31	29	-31	26	-22	23	-28	25
Marriage	Child	Chr	-33	47	-25	45	-30	46	-31	31	-22	27	-27	29
Marriage	Child	Death	-1	111	1	110	0	110	-7	109	0	107	-3	108
Marriage	Child	Marriage	-15	74	-8	73	-12	74	-12	69	-5	67	-9	68

Table 2. Deltas for calculating estimated marriage date ranges from relatives' events.

			Year-only date						Specific (day, month, year) date					
myEvent	**Relative**	**Event**	**Male**		**Female**		**Either**		**Male**		**Female**		**Either**	
Death(married)	Ind	Birth	-100	-21	-101	-18	-100	-19	-95	-25	-97	-20	-96	-22
Death(married)	Ind	Chr	-93	-10	-92	-14	-93	-13	-94	-2	-95	-2	-95	-3
Death(married)	Ind	Death	0	0	0	0	0	0	0	0	0	0	0	0
Death(married)	Ind	Marriage	-78	2	-81	1	-80	1	-70	0	-75	0	-73	0
Death(married)	Father	Birth	-140	-47	-140	-44	-140	-45	-138	-52	-141	-47	-140	-50
Death(married)	Father	Chr	-136	-36	-134	-32	-135	-35	-140	-38	-141	-36	-141	-37
Death(married)	Father	Death	-83	25	-83	29	-83	27	-80	24	-82	29	-81	26
Death(married)	Father	Marriage	-111	-23	-112	-18	-111	-20	-109	-26	-112	-22	-110	-23
Death(married)	Mother	Birth	-133	-43	-133	-40	-133	-41	-130	-49	-133	-44	-131	-46
Death(married)	Mother	Chr	-127	-38	-127	-35	-127	-37	-132	-32	-133	-30	-133	-31
Death(married)	Mother	Death	-84	29	-84	34	-84	31	-80	28	-82	33	-81	30
Death(married)	Spouse	Birth	-100	-18	-110	-21	-107	-19	-93	-23	-106	-25	-103	-24
Death(married)	Spouse	Chr	-88	-15	-99	-19	-96	-16	-93	-16	-105	-18	-102	-17
Death(married)	Spouse	Death	-53	56	-56	53	-55	55	-51	53	-53	51	-52	52
Death(married)	Child	Birth	-76	6	-80	4	-78	5	-67	0	-72	0	-70	0
Death(married)	Child	Chr	-67	17	-70	12	-68	15	-65	6	-67	3	-66	5
Death(married)	Child	Death	-52	84	-58	84	-54	84	-54	80	-59	80	-56	80
Death(married)	Child	Marriage	-51	50	-55	50	-53	50	-46	40	-51	39	-49	39
Death(single)	Ind	Birth	-99	0	-100	0	-100	0	-90	0	-93	0	-91	0
Death(single)	Ind	Chr	-96	0	-97	0	-96	0	-83	0	-84	0	-83	0
Death(single)	Ind	Death	0	0	0	0	0	0	0	0	0	0	0	0
Death(single)	Father	Birth	-139	-22	-140	-22	-139	-22	-134	-22	-137	-23	-135	-22
Death(single)	Father	Chr	-140	-3	-144	-9	-142	-6	-132	-21	-134	-20	-133	-21
Death(single)	Father	Death	-79	59	-81	60	-79	60	-73	59	-76	59	-74	59
Death(single)	Father	Marriage	-110	0	-112	0	-111	0	-103	0	-106	0	-105	0
Death(single)	Mother	Birth	-132	-19	-134	-19	-133	-19	-125	-19	-128	-19	-127	-19
Death(single)	Mother	Chr	-133	-3	-139	-6	-135	-4	-125	-17	-127	-17	-126	-17
Death(single)	Mother	Death	-79	64	-82	65	-80	65	-73	64	-75	64	-73	64

Table 3. Deltas for calculating estimated death date ranges from relatives' events.

2.2 Calculating ranges from a year and a delta

Given the above deltas, a *target date range* is estimated from a known range by looking up the appropriate delta (i.e., *delta*(*targetEventType, relativeType, sourceEventType, genderType, isSpecific*)), and then using

range.min = eventYear – delta.max
range.max = eventYear –delta.min

For example, in order to calculate an estimated birth range for a male from a father's death date of January 23, 1800, we look up *delta(birth, father, death, male, specific)* = -1..65. This gives us a *range.min* = 1800 - (65) = 1735; and *range.max* = 1800 - (-1) = 1801, for a resulting range of 1735..1801.

2.3 Further iterations: calculating ranges from a **range** *and a delta*

During the first pass of a date propagation algorithm, most source events consist of a single year. However, in a lineage-linked database, it may be necessary to iterate in order to estimate ranges for individuals who are two or more generations away from a relative with events.

Often a range must be calculated not from a specific year, but from another range (which in turn may have been calculated from another year or range). This is done by modifying the above formulas to be:

range.min = eventYearRange.min - delta.max
range.max = eventYear.max - delta.min

which is equivalent to the earlier equations when *eventYear = eventYear.min = eventYear.max*.

For example, assume that a person's father had an estimated birth range of 1800..1820. Then to calculate their death range, we could again use

delta(birth, father, death, male, specific) = -1..65

which yields an estimated range of

(*range.min-delta.max*)..(*range.max-delta.min*)
= (1800 - 65)..(1820 - (-1)) = 1735..1821

Note that the range is wider by 20 years than the above example (1735..1821 instead of 1735..1801) because the "source range" was wider.

In general, as ranges are propagated from one generation to the next, they get wider as there is less certainty.

2.4 Combining evidence: Intersecting ranges

Quite often there is more than one date or range available for use in estimating a particular event. For example, if a person has a death date of 1890 and a father's estimated birth date range of 1840..1850, we get two overlapping ranges: (a) The death date yields a range of 1890-92..1890-0 = 1798..1890. (b) The father's birth date yields a range of 1840-(-19)..1850-(-56) = 1859..1906.

These two ranges can be looked at as *constraints* on what years are reasonable for the person's birth event. Given the person's death year, it is reasonable that they were born between 1798 and 1890. But given when the father was born, it is only reasonable that they were born between 1859 and 1906.

To satisfy both constraints, we take the *intersection* of these two ranges, by using:

$range.min = max(\text{range}_i.min)$
$range.max = min(range_i.max)$

i.e., we build the resulting range by taking the maximum of the minimums, and the minimum of the maximums. In this case, the intersection of 1798..1890 and 1859..1906 is 1859..1890.

2.5 Ignoring conflicting data: Voting

Occasionally the multiple ranges that contribute to an estimated range conflict with each other. In particular, when there is a bogus (or at least surprising) piece of data, one of the ranges may not overlap with the others, making the intersection empty. This means that it is impossible to satisfy all the constraints.

For example, given several events on a person and their relatives, we may find ranges of 1800..1850, 1825..1860, and 1920..1940. Perhaps the last range was due to a typographical error. Using the above equations to find the intersection results in an "upside-down" range of 1920..1850, which is nonsense.

After examining several real-world cases where data like this appeared, a technique that resolved the problem most of the time was to use a voting scheme to identify which range or ranges were the most likely to be wrong. This is done as follows.

1. Each range is intersected with all of the ranges it overlaps with. This often yields one resulting range, but can result in multiple disjoint ranges.
2. For each of the resulting disjoint ranges, a count is made of how many of the ranges overlap each range.
3. The range with the most overlapping ranges is kept as the most likely one. All others are ignored.
4. If there is a tie, then the union of the tied ranges is used to reflect the uncertainty for that range.

This technique has been effective at identifying and excluding range estimates that are due to erroneous data or exceptionally rare cases.

3. Using Date Propagation in Practice

FamilySearch has used the above date propagation algorithms and data in several situations.

3.1 Person Matching

As part of the person matching (or "record linkage") algorithm used by "new FamilySearch" (the common Family Tree), date propagation is used to detect when two individuals are unlikely to be the same person because their events are unreasonable (Wilson, 2011). For example, if one person was born in 1820 and the other has a child born in 1821 or 1898, then the second person would be too young or too old to be reasonably having children.

3.2 Set classifier

In addition to simple person matching, a *set classifier* was used during a bulk person linking project involving over one billion individuals. After many pairs of likely matches were found, all of the individuals that would be brought together by combining all the pairs were listed in a set, and a *set classifier* used various features to determine whether the entire group looked like it really did represent one real person. Date propagation was used to estimate a birth, marriage and death range for each individual, using their own events and those of their parents, spouses and children. Disjoint ranges indicated that there were conflicts, resulting in splitting the sets to avoid bad merges.

3.3 Search

Billions of individuals from indexed records or submitted trees are searchable at *familysearch.org*. When this data is loaded, the above algorithms are used to estimate reasonable ranges for birth, marriage and death years for each individual, including those who don't have their own dates for these events. Then search queries can filter out unreasonable results. For example, if I search for James Gray born in 1800, I will not get someone who died in 1790, nor someone who has children born in 1800.

3.4 Living calculation

In order to protect the privacy of living individuals, it is important to be able to tell which individuals in a submitted tree might be living. On the other hand, it is good to share as much data with the public as can be done safely. The above algorithms were used to estimate a birth and death range for everyone in the submitted trees at *familysearch.org*. In addition, a "110-year-rule" was applied so that after all the date propagation was done, someone was assumed to be "possibly still living" if they were born less than 110 years ago.

The result of running the algorithm was that each individual had a death range. If the latest reasonable death year was earlier than the current year, then the person is likely dead. If not, then they might still be alive, and are thus hidden. (If there was no connection to even a distant relative with an event, then there is no estimated death event, so the person is assumed to be possibly living, and is thus restricted as well).

This algorithm was compared with simpler algorithms, such as doing propagation using the "most common delta" instead of a delta range (e.g., assuming that a male is 25 when married instead of using a range of 16..65), followed up by the same 110-year rule. It was also compared with an even simpler rule that only used the person's own events and those of their spouses and children to estimate a death year (since this was an algorithm in use in another system that we wanted to compare with).

3.5 Simulated labeled data

It is not trivial to come up with a good way to evaluate one date propagation algorithm against another. We had available a corpus of over 100,000 submitted GEDCOM files from the Pedigree Resource File (PRF). However, each person in the corpus either has a death date, in which case we can assume they are deceased; or they do not, in which case we don't really know if they are deceased or living, i.e., we can't tell the difference between those who don't have a death date because it simply wasn't entered and those who don't have a death date because they are still alive.

Therefore, the author devised a technique for simulating labeled data using the data at hand.

Pruning the graph at 1900. To apply the technique, we select some year in the past such as 1900. Then we trim a GEDCOM file down to make it look like it would have if 1900 was the date it was submitted (well, ok, we don't put it on parchment, but still...). That is, we find anyone with a birth date after 1900 and remove them, their spouses and all their descendants. We also remove any events that happen after 1900. For death events, we keep track of what the death year was for people whose death date was removed.

The result is that we have a graph that looks much like it would of if it only had data that could have been known up to 1900, but we still secretly know what the death date of many of the people are, i.e., everyone who had a death date listed that was after 1900.

(Incidentally, this was tried using several cut-off years, and it was found that from 1900 on back the results looked fairly stable, but after 1900 they changed more and more, since many of the people in the database then ended up being young enough that it wasn't known what their eventual death date was going to be. In other words, by using pruning the data to 1900, everyone in the database is dead, and some of them have known death dates. But using a later date, more and more of them really aren't dead, so the distribution is skewed. So 1900 seemed to be the best cut-off date to use.)

Leave-one-out sampling. The other trick employed was to take one person at a time who has a known death date (either one that was before 1900 or one that was after and thus removed), and temporarily remove that death date (if it was before 1900 and thus not already removed). Then the date propagation algorithm is run on the whole relationship graph to see what the latest estimated death date is for the person in question.

This resulted in a mapping of known death dates to estimated death dates. From this mapping, four counts were gathered:

1. Estimated and actual death year both before 1900 => "correct dead"
2. Estimated and actual death year both after 1900 => "correct living"
3. Estimated death year < 1900, actual > 1900 => "false dead" / "leaked living data"
4. Estimated death year > 1900, actual < 1900 => "false living" / "(unnecessarily) hidden data"

Empirical results. Experiments were run on 1000 randomly-selected PRF GEDCOM files using a cut-off (filter) year of 1900 as described above. 5,590 individuals were randomly selected among those who had their own death date (which was of course hidden from the algorithm). 1,528 of the 5,590 sampled individuals died after 1900 and 4,062 died before 1900.

For the purpose of the comparison, the 110-year rule was initially *not* applied, thus helping to better see the difference between the range propagation and simple "most common year" propagation.

Table 4 shows a summary of the results on the simpler "most common year-propagation" approach, compared to the proposed range-propagation method.

	Year		**Range**	
	count	percent	count	percent
Correct dead	1492	26.69%	1517	27.14%
Correct living	3458	61.86%	3194	57.14%
False dead ("Leaked living")	36	0.64%	11	0.20%
False living ("Hidden dead")	604	10.81%	868	15.53%

Table 4. Results of "year propagation" vs. "range propagation" (without 110-year rule applied).

As can be seen from the table, the range-propagation was somewhat more conservative, hiding about half again as many dead people compared to the year approach; but it "leaked" only about a third as many living individuals, which is the greater concern in this application.

When the 110-year rule was applied to both approaches, however, the results were much more similar, and both approaches "leaked" only one individual, with the trade-off that both approaches hid about 22% of the individuals unnecessarily.

The results had small numbers of leaked individuals identified, so a repeat of the experiments with a larger sample size would help draw conclusions that are more statistically significant. By adjusting the width of the ranges, trade-offs between leaking and hiding would be viewable to see if there a point where range-propagation provides both less leaking and less unnecessary hiding at some level.

Only 30.3% of the individuals in these files had their own death date, and 67% had their own event of any kind, so date propagation in general was able to salvage much of the data that would have had to be hidden without it.

4. Future work and Conclusions

One potential problem with range propagation is that it treats all years within a range with equal probability, whereas the underlying distribution is such that some years are much more likely than others. As ranges propagate across several generations, the probability of the edges of the ranges tends to become very small, but the ranges don't take that into account. The ranges are therefore probably wider than they need to be.

One possibility is to propagate *probability distributions* instead of flat ranges. Instead of intersection, *convolution* could be used to determine the probability for each year in the resulting range, and the final ranges could be trimmed according to some probability (e.g., trim the top and bottom 1%) to result in a final estimated range. After some discussions about this idea, Mayfield (2009) did some work using a similar idea with some success on some related problems, which makes this approach appear promising in this particular problem as well.

Date propagation has been a powerful tool in making matching and searching more accurate and in restricting access to data about living individuals without hiding too many people unnecessarily.

References

Brox, Vegard, and Brian Randell, 2000. "Date Estimation in Lineage-Linked Databases," B.S. Dissertation, University of Newcastle upon Tyne. (*http://homepages.cs.ncl.ac.uk/brian.randell/Genealogy/Brox/dissertation/dissertation.html*)

Despain, Bruce, 2001. "Integration of Genealogical Information", *Family History Technology Workshop 2001. (http://fht.byu.edu/prev_workshops/workshop01/final/Despain.pdf*).

Mayfield, Chris, Jennifer Neville, Sunil Prabhakar, 2009. "A Statistical Method for Integrated Data Cleaning and Imputation," *Computer Science Technical Reports*, Department of Computer Science, Purdue University, Paper 1723. (*http://docs.lib.purdue.edu/cstech/1723*).

Wilson, Randy, 2011. "Genealogical Record Linkage: Features for Automated Person Matching," *RootsTech 2011*, pp. 331-340, February 2011, Salt Lake City, Utah.