Why Are There Two Aunt Sally's?

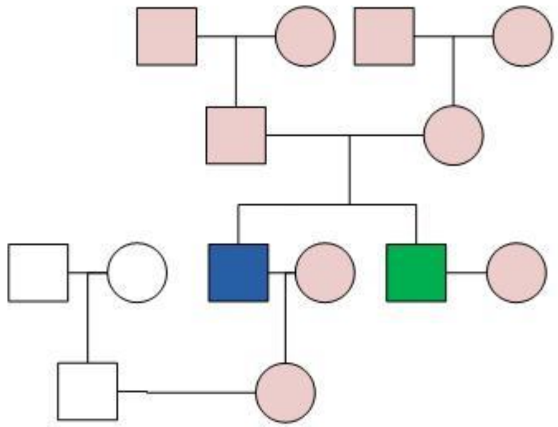
Finding Duplicates Using GenMerge and GenMergeDB

Sue Dintelman
Tim Maness

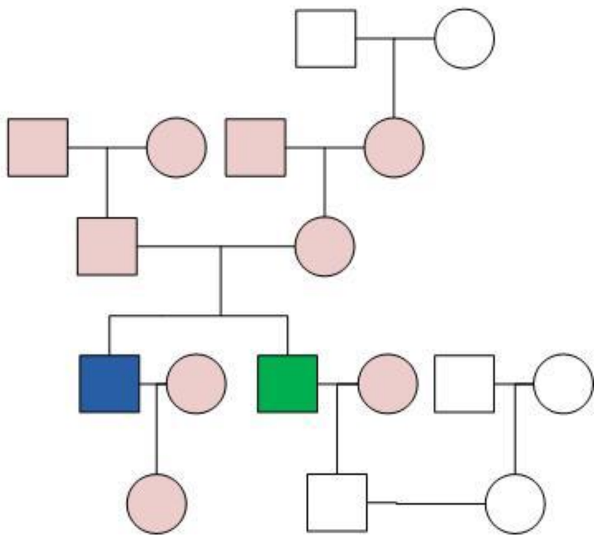
Agenda

- Why ARE there two aunt Sally's?
- How do I find them?
  - Intro to probabilistic record linking
  - Scoring
  - Steps of a linking job
- Case studies

Why do duplicates exist?



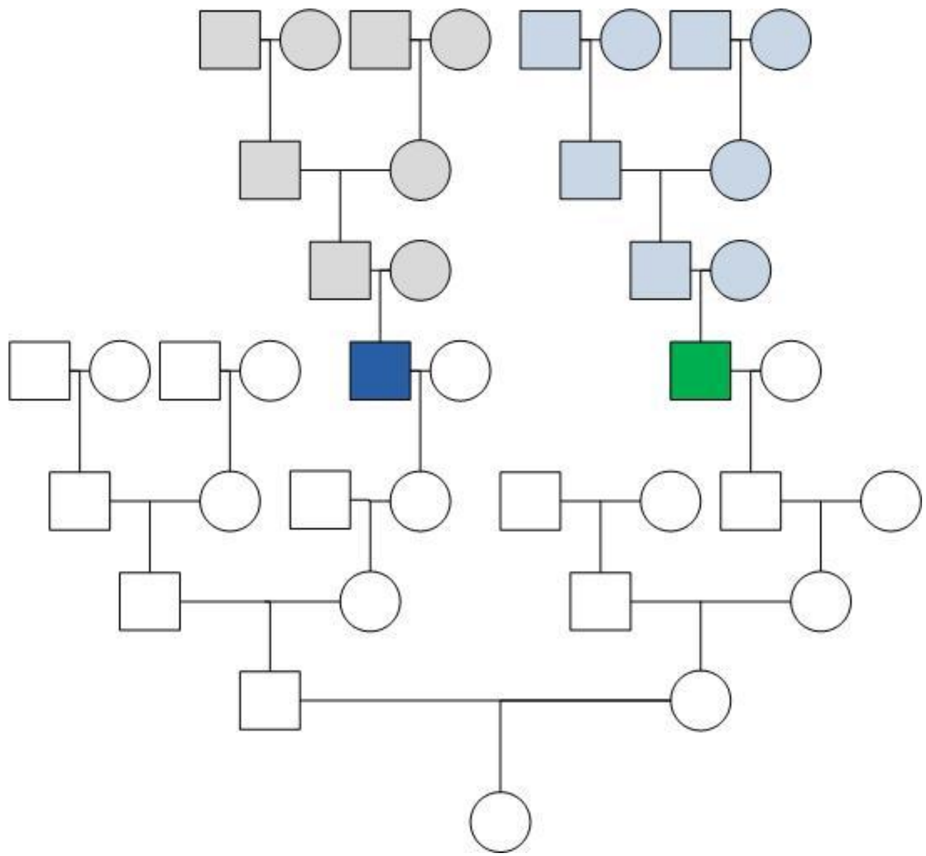
1. Add data to your family file from a relative or on-line site



Pleiades Software Development, Inc.

Why do duplicates exist?

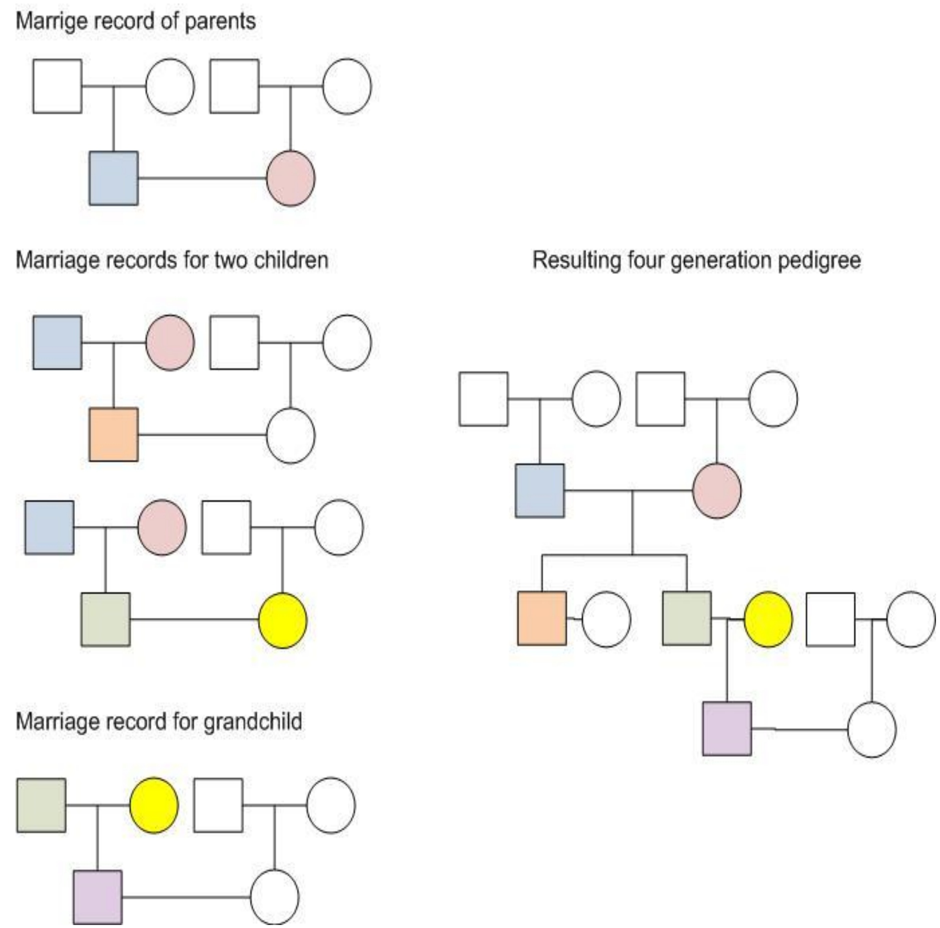
2. Common ancestors



Why do duplicates exist?

3. Population Reconstitution

Birth/Christening records
Marriage records
Death/Burial records
Census records
Probate records



Pleiades Software Development, Inc.

Probabilistic Record Linking

The score for two records is an estimate of the conditional probability that the two records are the same (L) divided by the probability that the two records are not the same (~L).

$$\frac{P(L \mid O_1 O_2 O_3 \ldots)}{P(\sim L \mid O_1 O_2 O_3 \ldots)}$$

$$= w_1 + w_2 + w_3 \ldots + \log_2(P(L)/P(\sim L))$$

Weights

Each field level weight is estimated by

$$w_i = \log_2(p/a_i)$$

where p is the population size and a_i is the absolute frequency of the value i in the population

$$\text{Smith} = \log_2(28{,}575{,}300/283{,}207) = 6.6$$

Scale by 10 so the scores are integers

Scoring

Common

John Smith

John Smith

35 + 66 = 101

Uncommon

Benedictus Allphin

Benedictus Allphin

169 + 158 = 327

Total Score = Individual score + Family score

- John Smith
- John Smith

Individual score 101

Fathers:	Mark Alma Smith, b. 6/4/1955 Rock Springs, Sweetwater, WY
Mothers:	Salli Bolen, b. 7/31/1953
Sibs:	Nathan Mark Smith Jared Willis Smith Cullen Smith Garth Smith

Family score 537

Total score 638

Partial Matches

Fields often don't match exactly

Historical changes

Recording errors in original documents

Transcription errors in digitizing

- Names
  - Most important fields for matching, need to get the most out of every name
- Dates
  - Some dates are estimates and are often wrong, even if self reported
- Places
  - Same issues as other string fields, plus changes over time

Name Matching

- Phonetic algorithms
  - NYSIIS
  - Utah Phonetic Transducer
- String similarity measures
  - Jaro-Winkler
- Maiden names
- Patronymics, toponymics

Examples

Phonetic

Quinn	QUAN, KAN, KAM, KM	String Comparator: .633
Kwin	KWAN, KAN, KAM, KM	
Snyder	SNADAR, SNADAR, SMADAR, SMDR	StringC: .756
Schneider	SCHNADAR, SHNADAR, SMADAR, SMDR	

String matching

Anderson	ANDARSN, ANDARSN, AMDARSM, AMDRSM	StringC: .933
Ancerson	ANSARSN, ANSARSN, AMSARSM, AMSRSM	

Linking Process

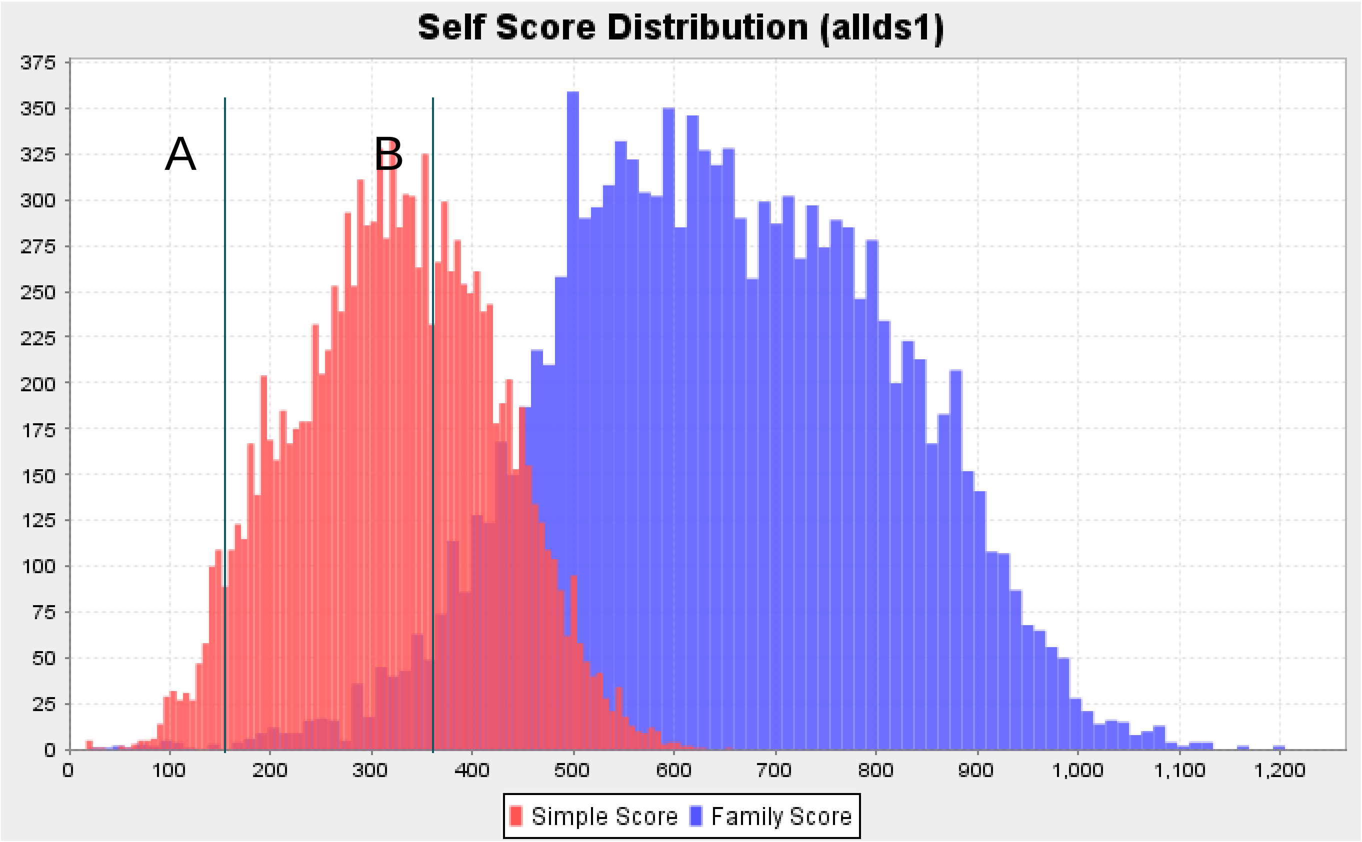
1. Create weights
2. Set cut-off scores
3. Bin
4. Compute scores
5. Choose best links
6. Process linked records to merge family members

Step 1. Create Weights/Discriminating power

Field	Distinct Values	Distinct Value (25% of Pop)	Discriminating Power	%Complete
First Name	2112/1656	1/2	3.8/4.4	99.3/99.5
Last Name	6014	6	5.87	96.96
Middle Name	1088/1238	1/4	3.4/5.6	17.94/13.71

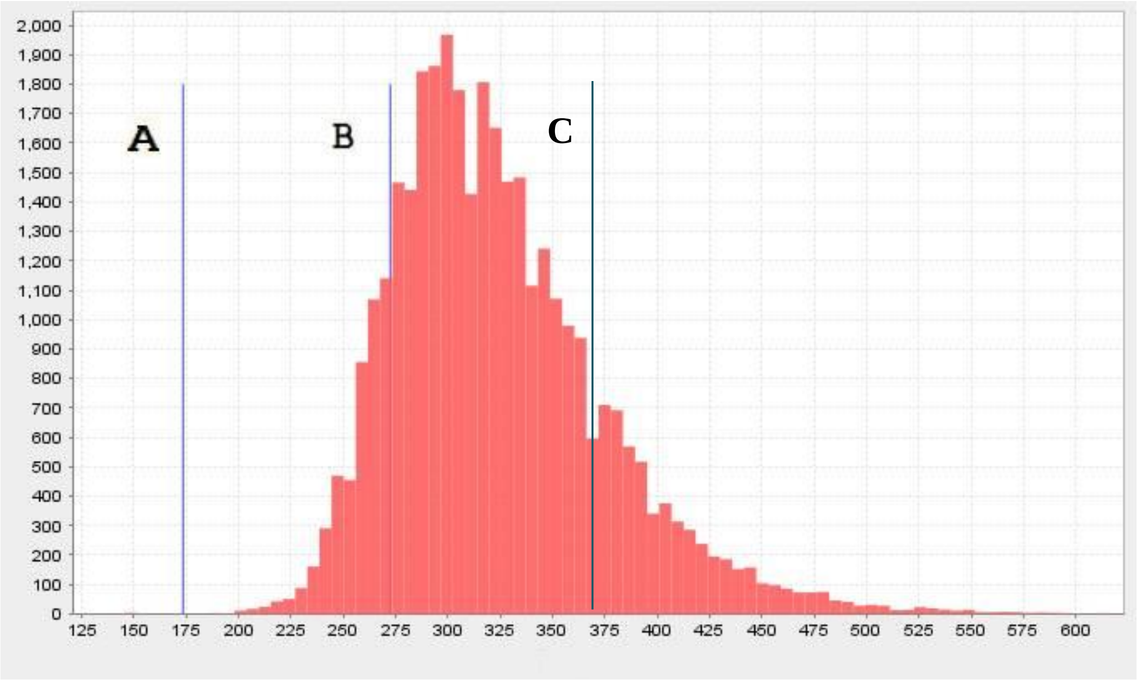
25% of the men are named John
25% of the women are named Mary or Catherine
25% of lastnames are:
McDonald, MacDonald,
McNeil, MacNeil,
Chisholm, Gillis

Step 2. Set cut-off score



Pleiades Software Development, Inc.

Step 2. Set cut-off score



Self Scores

Pleiades Software Development, Inc.

Step 3. Binning

- Not practical to compare every record to every other record
- There is no perfect binning criteria

- The trade off:
  Minimize comparisons w/o
  missing potential duplicates

Step 4. Score
Step 5. Choose best links

Step 6. Process Linked Records

Jonathan Andrews 5/11/1842

Father: James Andrews
Mother: Anna Franklin
Spouse: Maryann

John Andrews 1842

Father: James Andrews
Mother: Anna Franklin
Spouse: Mary Ann Anderson

Linking Process

1. Create weights
2. Set cut off scores
3. Bin
4. Compute scores
5. Choose best links
6. Process linked records to merge family members

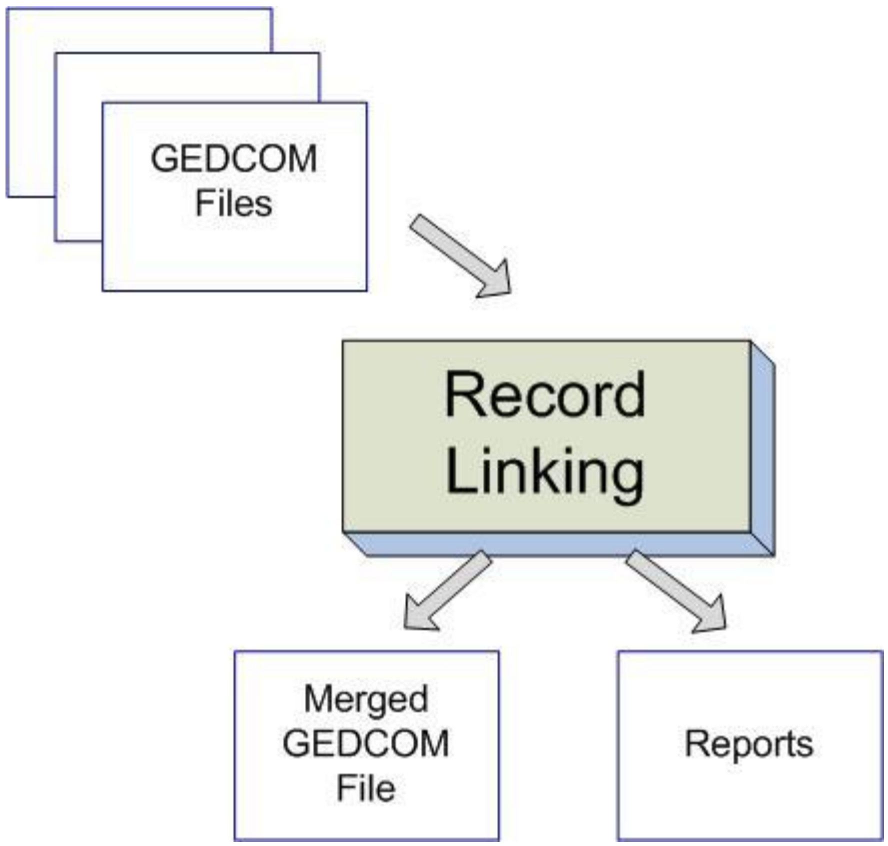
GenMerge

- Desktop product
- GEDCOM files
- Maximum 200,000 records
- Weights based on large Indo-European population (27 Million+)
- Cut-off values pre-set

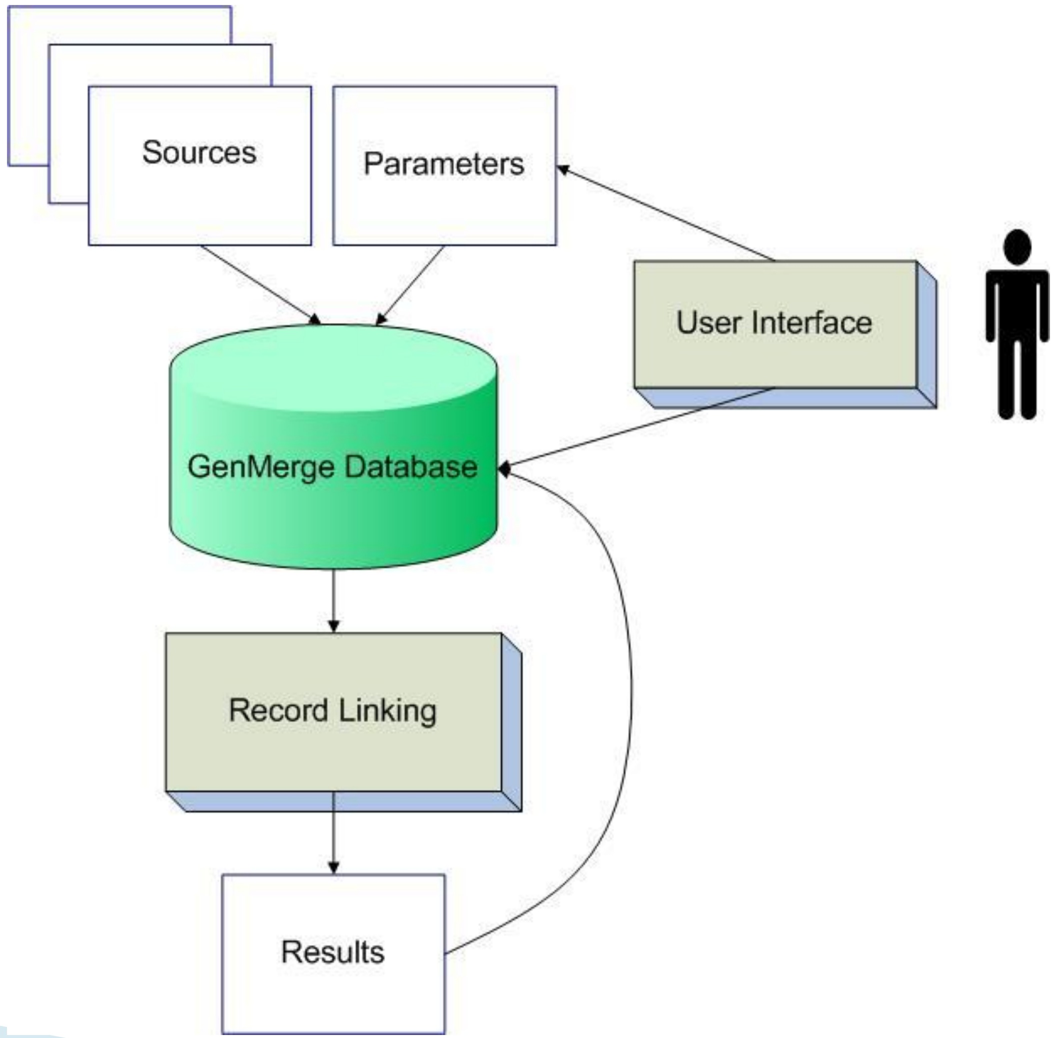
Find duplicates in a single file

Merge multiple trees together

GenMerge



GenMergeDB



Pleiades Software Development, Inc.

Case Study 1: Merge two trees

Case Study 2: Evaluate trees

Case Study 3: Population
Reconsitution

What's next

- GenMerge 3.0 Shipping Feb 2011
- Linking as a service (Laas)
- New Family Search application
- Continued development of linking process
  - Improve linking
  - Improve data preparation tools
  - Improve analysis tools

Thank You.

Sponsored by: