Probabilistic Record Linkage for Genealogical Research

John Lawson, David White, Brenda Price, and Ryan Yamagata

With increased interest in family history research, there is a great need for improvement in procedures for generating genealogical information. One of the most time-consuming parts of the work is searching through records (such as civil records, church records, census records, immigration records, wills, deeds, and certificates of births, marriages, and deaths) for information about an individual. When multiple records are searched, an individual may appear numerous times. Each of these occurrences may contain identical or unique information about the individual. More complete information (such as pedigree) can be constructed for an individual by combining or linking all the records about that individual, especially when in one record the individual appears as a child and in another record as a parent.

Presently, when a genealogist searches through records he or she usually links records manually. This process entails looking at the individual records and comparing the information within each record. The genealogist then decides if any records are *matches,* representing the same individual. Done on a record-by-record level, this is a time-consuming and expensive process.

By comparison, in today's information age most records on individuals (such as financial and medical records) are stored electronically to facilitate quick computer searches. If civil records, useful for genealogical research, could be stored electronically, entire files could be searched in seconds instead of hours or days.

However, it would take more than just storing civil or church records electronically to allow genealogical researchers to use them optimally.

Matching or linking records on one individual is usually accomplished by using a unique identifier such as the social security number. Older records do not contain unique identifiers such as social security numbers to aid in computer searches. Programs written for simple searches would have to match on information such as surname, given name, and date of birth. Herein problems lie. Early civil and church records may use different spellings of names in different records of the same individual. Nicknames may be used, dates may be misreported, or day and month may be interchanged. Needed information may be missing. Programs written for simple searches will miss many matches because these algorithms require fields to be matched identically. The slower but surer trained genealogist will match many more records and compile a much more complete history of an individual by recognizing human variations, catching errors in names and dates, and realizing that various fields do not need to match exactly but be "close."

Procedures grouped under the classification of *probabilistic record linkage,* which links records that are not necessarily identical but close in some fields, have been developed by researchers in the U.S., England, and Canada. Probabilistic record linkage allows a computer to mimic some of the decision-making processes a genealogist may use to recognize valid variations in the data. Although these methods are not intended for genealogical research, The Church of Jesus Christ of Latter-day Saints Family History Department has adapted these procedures for use in the computer program TempleReady, which is used to identify ordinance work that has already been performed for an individual.[1]

In this paper we describe the approach to probabilistic record linkage used in TempleReady based on a method of weighting that is described by David White,[2] and we show its application to genealogical research using a set of civil and church records of Quakers in Perquimans and Pasquotank Counties, North Carolina. The results of our study are very promising. Probabilistic record linkage has the potential of dramatically increasing the productivity of genealogical researchers. This paper is a report of a work in progress and describes what has been done to the present and outlining some of the tasks yet to be addressed.

Historical Overview of Record Linkage

Record linkage is a relatively modern concept. Halbert Dunn, chief of the U.S. National Office of Vital Statistics, introduced the term "record linkage" in 1946.[3] Dunn used the term to describe a process that joins separately recorded pieces of information for a particular individual or family. During

the 1950s the idea for computerized record linkage was born, and in 1959 H. B. Newcombe and others[4] were the first to make probabilistic linkages of vital records in order to track hereditary diseases. This method used the mathematical probability of agreement or disagreement in a certain field as the classification factor.[5] Unfortunately, computing capability at that time limited the efficiency and practicality of this method.

In the 1960s, mathematical theory for record linkage began to appear in the literature. Papers by N. S. D'Andrea Du Bois,[6] Gad Nathan,[7] Benjamin J. Tepping,[8] and Ivan P. Fellegi and Alan B. Sunter[9] laid a theoretical foundation for record linkage methodology. Fellegi and Sunter's paper emerged as the theoretical approach most often cited and as the basis for most current methods of record linkage. It was developed along the lines of classical hypothesis testing using a likelihood-ratio-type statistic. The logarithm of the likelihood ratio is a sum of *weights,* one weight for each *field,* used to compare records. The objective of the linkage is to minimize the number of records that are misclassified, which is achieved by establishing threshold values for decision-making based on the log likelihood ratio.

In the past few decades, advances in computers and computational methods have improved the methods and speed of record linkage. Record linkage software such as CANLINK, developed at Statistics Canada by Nancy J. Kirkendall;[10] CAMLIS, developed at the University of California at San Francisco by Max A. Arellano and others;[11] and LinkPro, developed by A. Wajda and others at the University of Manitoba,[12] are based on the Felligi-Sunter model. In addition, a wealth of recent literature focuses on how to apply the Fellegi-Sunter model to specific types of data.

Description of Record Linkage for Genealogical Research

The first step in record linkage for genealogical research is to manually enter the records on magnetic storage media (computer disks) as a GEDCOM file.[13] The data should be entered using the "Family Records" option. This option allows for the following fields to be entered for an individual: surname, first and second given name, title, birth and death dates, congregation, town, country, and state. It also allows for family units of parents and children to be entered along with marriage information.

To link records, a comparison is made of pairs of records selected from the file. The entries for corresponding fields may be the same, may be different, or one or both entries may be missing. For most linkages of this type, it is anticipated that the number of missing entries may be large, but missing entries are taken into account in this methodology. Positive and negative weights are assigned in advance to each field. David White describes the

details for computing these weights.[14] When two records are compared, the positive weight for a field is used if the records match on that particular field; the negative weight is used if the two records do not match on that field; a zero weight is used if the field is blank in one or both records. A score equal to the sum of weights (over all the fields) is then calculated for each pair of records compared. Large positive scores indicate the pair of records represents the same individual, and large negative scores indicate the pair of records does not represent the same individual.

Initially, a *training set* of records, which could be a subset of the records in the file, is used to estimate the weights. The records in the training set are sorted, using a field or combination of fields that are considered to be useful in identifying matches (pairs of records that are highly likely to represent the same person). An example would be to sort first on surname and then on given name, since records representing the same person would most often have the same name. A set of records having the same given name and surname is then defined as a *block* (more generally, a set of records with the same value for the sort field or fields is defined as a block). Next, a genealogist looks at the blocks of records and identifies matches.

From the matched records the weights are determined as the log odds in favor of a pair of records being a match given agreement or disagreement on a particular field. The odds for agreeing fields are estimated by counting the proportion agreements on particular fields within records considered a match by the genealogist, divided by the proportion of agreements among randomly paired records. Once the weights are established for each field, the score or sum of weights is calculated for every pair of records in each block. Pairs with a large positive score are considered linked, and pairs with a large negative score are not linked.

Measuring the Effectiveness of Record Linkage

There are two kinds of errors that can be sustained when using record linkage.

a. A false negative: Concluding from the score that a pair of records do not represent the same individual, when by manual inspection, they do. The probability of this error is defined as λ.

b. A false positive: Concluding from the score that a pair of records do represent the same individual, when, again by manual inspection, they do not. The probability of this error is defined as μ.

A third situation, which deserves a probability, occurs when there is insufficient information to make a decision. The probability of this is defined as γ.

The probabilities λ and μ in the training set can be controlled by choice of the upper threshold values T_μ and the lower threshold value T_λ. If the score determined by comparing all the fields on a pair of records exceed T_μ, the pair of records is linked. If the score is less than T_λ, the pair is not linked. If the score falls between T_μ and T_λ, there is insufficient evidence to make a decision. The smaller T_λ is chosen to be, the lower the probability, λ, of failing to link known matches. The larger T_μ is chosen to be, the smaller the probability, μ, of falsely linking a pair that is not a match. In accordance with normal statistical practice, this choice should be made such that μ (the probability of a false positive) and λ (the probability of a false negative) are both less than 0.05. Relative effectiveness of specific record linkage projects can be assessed by comparing the probability of no decision, γ, with the thresholds adjusted so that λ and μ are nearly the same for each data set.

The use of thresholds is illustrated in figure 1 below, which shows frequency histograms of the scores of matched pairs and nonmatched pairs in a hypothetical set of records. The upper and lower threshold values are shown on the graph. The probability μ, shown on the graph, is the proportion

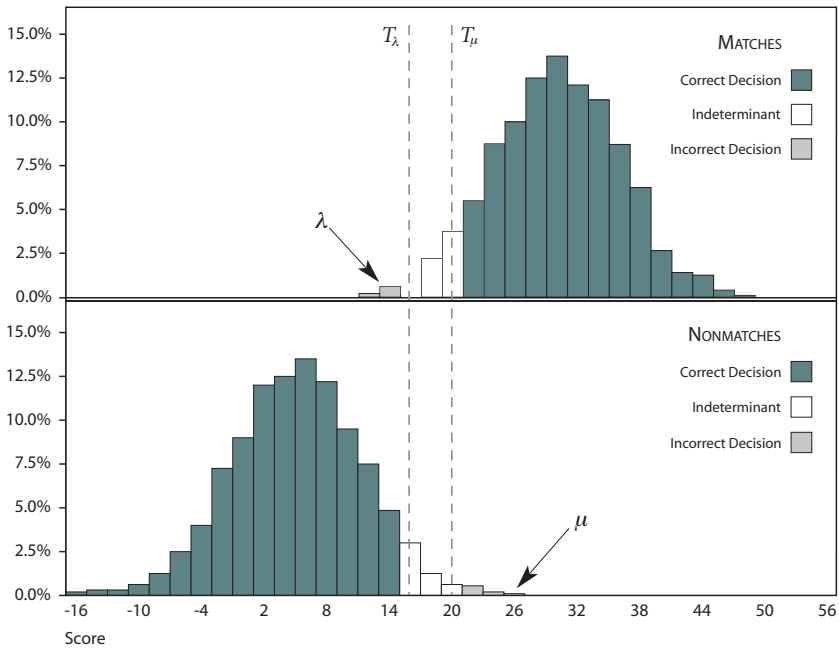

Fig. 1. Hypothetical relative frequency histogram of scores for pairs of matches and nonmatches with T_λ and T_μ

of scores in the lower histogram above the upper threshold, T_μ. The probability λ is the proportion of scores in the upper histogram below the lower threshold, T_λ.

An Application with North Carolina Records

The data used in this paper consist of a collection of records transcribed from handwritten documents recording the proceedings of Quaker congregation meetings[15] or county birth, death, and marriage records.[16] The Quaker records are a compilation of individuals mentioned in the minutes of the yearly Quaker congregation meetings of Perquimans and Pasquotank Counties. The individual information contained within these records varied greatly. Some records contain birth and death dates with parental and spousal information. For example, a family group record reads as follows:

> Benjamin C. Winslow, s. William & Julian, b. 3–5–1837, Chowan Co.
> Esther P. Winslow. (dt. Silas & Elizabeth Chappell, b. 2–10–1840, Chowan Co. p. 11-4)
> Ch: Harriett Ann b. 6–23–1862.
> William W. " 11–8–1864.
> James Claudius " 9–21–1873.
> Ora
> Henry[17]

From this entry one record would be made for each individual mentioned. Other records contained only limited information for a single individual, for example:

> Laden.
> 1880, 8, 7. Sarah (form Winslow) rpd m. (not m in mtg).

The county records were organized as records of events in which individuals were mentioned. An example of a birth record reads:

> George Durant son of George & Ann Durant was borne the 24th December 1659[18]

There were a total of 9,279 individual records for comparison in these sources.

The format of the printed records required that the information to be manually entered into a computer database. This was done using Personal Ancestral File (PAF) Release 2.3.1, a software package produced by The Church of Jesus Christ of Latter-day Saints for the recording of genealogical data.[19] The format used by PAF is such that entering the records was a simple task, and all family relationships were maintained as recorded in the printed records.

The data was entered into PAF using the "Family Records" option. This option allows the following fields to be entered for an individual: surname[20]; first and second given name; title; birth and death date; and town, county, and state of the congregation. It also allows for family units of parents and children to be entered along with additional marriage information, if desired. Any additional information can be entered by selecting the "Create Notes?" option. The notes option was used to enter information for fields which were not available, specifically information about the related event that was recorded. For example, if the record was a birth record, the child's birth was entered in the notes for each parent, with the associated date and place.

Using a procedure in PAF, a GEDCOM file was then created from the input information that contained all the information of all the records. The GEDCOM file contains two sections: The first section contains only an individual's information. The second section contains all the family information.

The individuals section of the GEDCOM file lists each individual. Each record was assigned a Record Index Number (RIN). This unique identifying number was further used in the family section. The individuals were listed by RIN in sections of five to ten lines that include all personal information.

Each family group was assigned a Marriage Record Index Number (MRIN). The family section of the GEDCOM file consists only of RINs, MRINs, and marriage information, such as date and place, if available. The family groups were listed by MRIN. Within each group the RINs associated with the father, mother, and each child were identified. Therefore, to include information about family relationships, the family section was referenced, and, to retrieve individual specific information, the individuals section was used. Both were needed to construct each record's information.

These GEDCOM files needed to be converted to flat files[21] in order to simplify the linkage process. The conversion of these GEDCOM files to a flat file was done using Microsoft Visual Basic.[22] The Visual Basic program used the GEDCOM files to gather all the personal and family information for each record. It then created a flat file that assigned each record a single line. On that line, each piece of information was placed into a single field. For each record there were 21 fields, although many of the fields were blank for any given record. The fields present were surname, first given name, sex, father's given name, father's surname, mother's given name, mother's surname, spouse's given name, spouse's surname (or maiden name), birth town, birth county, birth state, birthday, birth month, birth year, death town, death county, death state, death day, death month, and death year. The complete flat file contained multiple records for many individuals.

A training set constructed by matching of records representing the same individual was done manually in Microsoft Excel. Performing various sorts and searches and using the original records as a reference found additional matches from the amended data. In our 9,279 records, a total of 880 individuals were found to have more than one record in the file. This training set was used to calculate the weights for probabilistic record linkage. Records were paired in order to calculate the log odds of agreement or disagreement of each field, given that the pair was a match or not a match.

To reduce the number of pairs to be considered, blocking was done to find a restricted subspace. Two different blocking methods were used for comparison. The first method used surname and sex as the blocking factors, leaving 19 fields available for comparison. Of the 9,279 records, 1,875 did not have a surname listed and thus were not considered. These records consisted mainly of married females without record of their maiden name. This left 7,404 records to be blocked for comparison. After blocking, there were 220,931 pair-wise comparisons to be classified, much fewer than blocking only on surname. Of these, 2,118 were known matches and 218,813 were considered nonmatches.

The second method blocked on surname only. Those records with missing surnames were considered a block and paired within that block for consideration. After blocking, there were 1,961,004 pair-wise comparisons to be classified. Of these, 3,692 were known matches and 1,957,312 were known nonmatches. Using this method, there were 20 fields available for comparison.

All blocking was performed using Visual Basic. The Visual Basic program simply paired all records and then output each pair, with all fields, that satisfied the blocking criteria as a line in a flat file.

The weights for the individual fields were estimated as previously described and for the second case were blocked on surname only. The results are shown in table 1.

For each field, two weights were calculated: $w_i(S)$ was used if records being compared agreed on the field; $w_i(D)$ was used if the records were not in agreement for the field. If the field was missing for either record, then a weight of zero was assigned. Death town was given a weight of zero since for every matched pair of records death town was missing from one or both records.

Using the blocked data defined earlier, a score was then calculated for each pair of records within the block. Each pair of records was compared field by field. Using the weights given in table 1, each field present in both records was given a weight based on the field's agreement status. The score was then found by summing all of the weights. This score reflected the likelihood that the two records were a match. A large value indicated the records should be linked. Conversely, a small value indicated the records should not be linked.

Table 1

Calculated Weights for the Individual Fields

Field No. (i)	Variable	Calculated Values $w_i(S)$	$w_i(D)$
1	Given Name	3.47715	-2.81401
2	Sex	0.69078	-8.16280
3	Father's Given Name	2.83686	-2.54161
4	Father's Surname	3.89474	-2.44506
5	Mother's Given Name	2.09498	-1.64660
6	Mother's Surname	3.04619	-8.16280
7	Spouse's Given Name	3.30857	-2.58610
8	Spouse's Surname	4.39975	-3.06505
9	Birth Town	0.00176	-8.16280
10	Birth County	0.55256	-1.57191
11	Birth State	0.00604	-8.16280
12	Birthday	3.43841	-2.16826
13	Birth Month	1.98113	-0.91975
14	Birth Year	4.60908	-1.09195
15	Death Town	0.0	0.0
16	Death County	0.59431	-8.16280
17	Death State	0.0	-8.16280
18	Death Day	3.47962	-1.70889
19	Death Month	2.28891	-2.04636
20	Death Year	4.41364	-2.12932

When blocking by surname and sex, and including the fields of father's given name, father's surname, mother's given name, mother's surname, spouse's given name, and spouse's surname, the distributions for matches and nonmatches were separated as shown in figure 2. Setting $T_\mu = 7.88$ and $T_\lambda = 4.40$ yielded values for μ and λ of 0.0187 and 0.0165 respectively. These threshold values also resulted in low unclassified rates. Only 7.71% of the nonmatches and 17.52% of the matches are between the threshold values and classified as indeterminable status.

Blocking by only the surname allowed one more field to be used for comparing records. In addition to the six family-related fields previously used, sex was also considered as matching criteria. This method of blocking also found the distributions of matches and nonmatches to be sufficiently separated. In this case it is sufficient to set only one threshold. Setting $T_\mu = T_\lambda = 2.28$ yields error rates of 0.0239 and 0.0496 for μ and λ respectively. This can be seen in figure 3. In this situation, the error rates are still lower than 0.05, though they are both higher than in the previous method. But by having the slightly higher error rates, the unclassified rates are now both zero. Thus a decision is made for each pair of records examined.

The Future of Probabilistic Record Linkage for Genealogical Research

The results of probabilistic record linkage for genealogical research described in this paper are very promising. Once the weights are established through a training set, all the records representing the same individuals

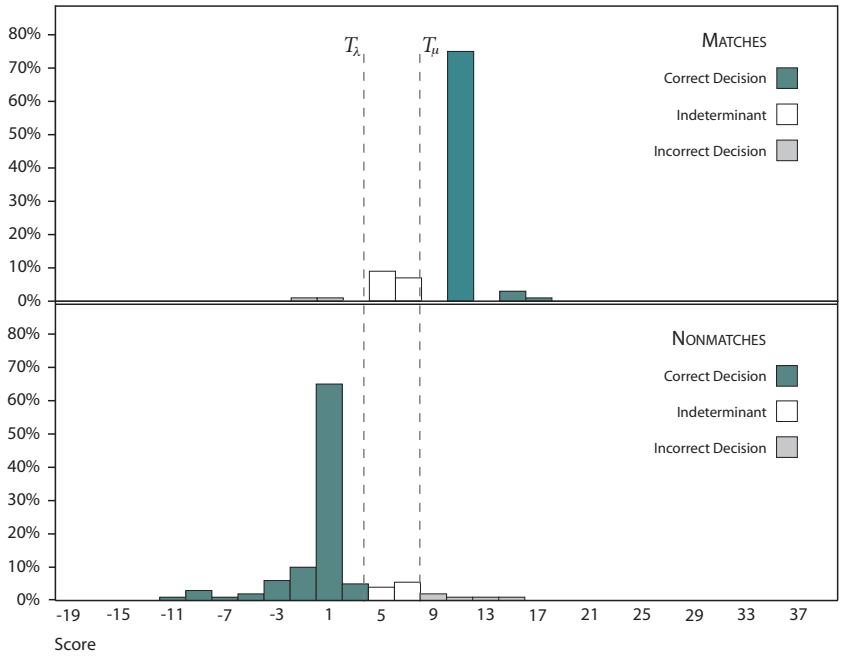

Fig. 2. Relative frequency histogram with thresholds when blocked by surname and sex

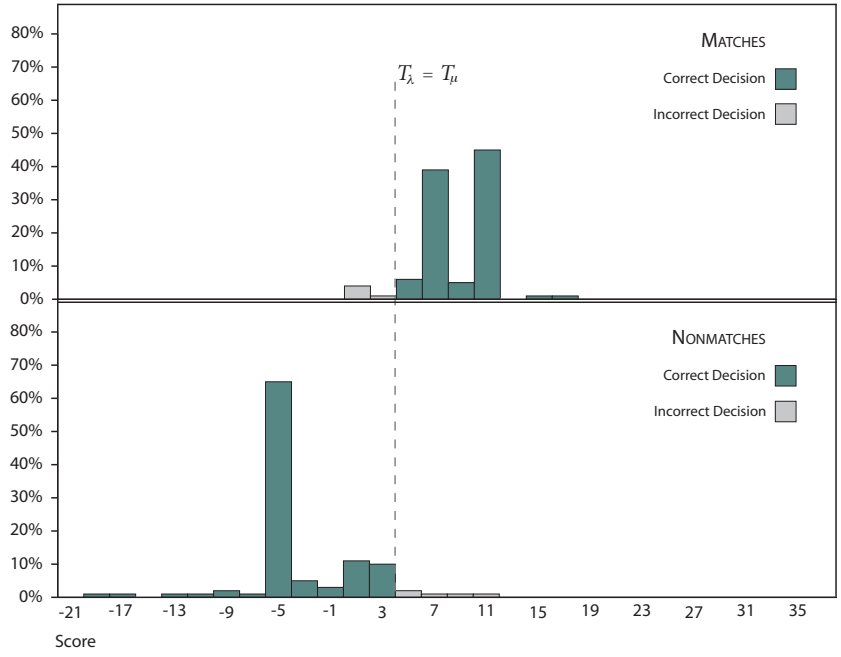

Fig. 3. Relative frequency histogram with thresholds when blocked by surname only

in a large GEDCOM file can be linked simultaneously in seconds using this technology, rather than having a genealogist spend hours or days to link the records relating to just one individual. But this research is still just the tip of the iceberg for what can be done. In this section we describe what we plan to do in the immediate future and then discuss what could be accomplished in genealogical research more universally through use of probabilistic record linkage.

In the research described in this paper (which was the result of two master's projects[23] in the Brigham Young University Statistics Department) a training set was formed consisting of 9,279 records from Perquimans and Pasquotank Counties, North Carolina. A GEDCOM file of the results was converted to a flat file of pairs of records using a Visual Basic program. The flat file was then read into Statistical Analysis System (SAS) where weights were calculated and records were linked using probabilistic record linkage, with less than 5% false positives and false negatives.

Although the results of this research were excellent, an immediate question comes to mind: How well will the weights created in the training set do in linking records that are not in the training set? One indication from our study that the results will be good comes from the fact that the weights didn't change much when the training set was expanded from the Perquimans County records to include the Pasquotank records as well. One of the next steps in our continued research is to test the question. We need to obtain more data, determine how well weights calculated from a subset of the data (or training set) do in linking records from the complete file, and see how weights change from one data set to another.

The linkage and calculation of field weights reported in this study were done using SAS. However, with some programming effort all of these tasks could be included in the portable stand-alone Visual Basic program that converted the GEDCOM file to a flat file. This is another item on our agenda for continued research. Weights could be calculated from a training set by this program or could be supplied by the user at a prompt. The program could then calculate the links for any GEDCOM file, write a modified GEDCOM file by combining all the linked records, and include any new family ties found through the linking process in the family section of the file.

This method would be of great benefit to those doing genealogical research. Instead of searching a GEDCOM file of somewhat unrelated records of births, deaths, wills, deeds, and so on for any information they could find on a particular individual, genealogists could simply read the modified GEDCOM file into PAF or a similar genealogy program. Then they could simply search for any individual and immediately view his or her entire family tree, spouse, children (in other words, the results of the prob-

abilistic linkage), as is now done in the Ancestral File, available through The Church of Jesus Christ of Latter-day Saints.

Having a quick stand-alone program to link the records in a GEDCOM file could change the whole emphasis in genealogical research. Instead of laborious searching of original records, the emphasis would shift to getting original records into GEDCOM files, running them through a probabilistic record linkage, and cataloging the results where they would be available to other researchers. Then the genealogical research would be almost as simple as it is today to look up an individual's credit history in a large database of linked financial records. Research could be automated and done in seconds.

Many other questions are yet to be answered as we learn more about applying probabilistic record linkage to genealogical research. Certainly the fields, weights, and threshold values that are effective in linking records will change depending on the locality and age of the records being linked. Is there any pattern to the changes? Will there be a way to predict what the field weights and thresholds should be without doing manual matching in a training set? As more resources and data are available we will research these questions.

In the study reported here, weights were developed for only two cases, where the fields are either the same or different in a pair of records. This weighting should be expanded to the case of "different but close." For example, for dates, the weight could be a function of the difference between two dates, possibly with higher weights given for transposed numbers. For names, positive weights could be given matching names, matching soundex code for name, or a reasonable nickname or initial.

Many similar questions remain, making probabilistic record linkage for genealogical research a fertile ground for research. We have investigated only one method of record linkage using the same method of weighting as used in TempleReady. Perhaps other schemes for developing weights or entirely new methods of record linkage based on theory of fuzzy sets may be more effective. These are all open questions that should be investigated in order to improve the methods that could revolutionize and automate genealogical research. Combined with computer automated methods of transferring original records to GEDCOM files, probabilistic record linkage is a method that has the potential of allowing interested people, even those with little formal training in research methods, to become highly productive in genealogical research work.

John Lawson (lawson@byu.edu) is Associate Professor of Statistics at Brigham Young University. He received a Ph.D. in applied statistics at Polytechnic Institute of New York.

David White is Professor Emeritus of Statistics at Utah State University. He received a Ph.D. in mathematics at Oklahoma State University.

Brenda Price is a statistician at Claritas Inc. in San Diego, Calif. She received an M.S. in statistics at Brigham Young University.

Ryan Yamagata is a statistician at Wyeth-Ayerst Pharmaceutical Corporation in Pearl River, N.Y. He received an M.S. in statistics at Brigham Young University.

1. Nancy P. NeSmith, "Record Linkage and Genealogical Files," in *Record Linkage Techniques—1997, Proceedings of an International Workshop and Exposition* (Washington, D.C.: Federal Committee on Statistical Methodology, Office of Management and Budget, 1997), 358–61.

2. David White, "A Review of the Statistics of Record Linkage for Genealogical Research—As Used for the Family History Library, Church of Jesus Christ of Latter-day Saints," in *Record Linkage Techniques—1997: Proceedings of an International Workshop and Exposition* (Washington, D.C.: Federal Committee on Statistical Methodology, Office of Management and Budget, 1997), 362–73.

3. Halbert L. Dunn, "Record Linkage," *American Journal of Public Health* 36 (December 1946): 1412–16.

4. H. B. Newcombe, J. M. Kennedy, A. P. James, and S. J. Axford, "Automatic Linkage of Vital Records," *Science* 130 (October 1959): 954–59.

5. The probability of agreement or disagreement is $\log_2$.

6. N. S. D'andrea Du Bois, "A Solution to the Problem of Linking Multivariate Documents," *Journal of the American Statistical Association* 64 (March 1969): 163–74.

7. Gad Nathan, "Outcome Probabilities for a Record Matching Process with Complete Invariant Information," *Journal of the American Statistical Association* 62 (June 1967): 454–69.

8. Benjamin J. Tepping, "A Model for Optimum Linkage of Records," *Journal of the American Statistical Association* 63 (December 1968): 1321–32.

9. Ivan P. Fellegi and Alan B. Sunter, "A Theory for Record Linkage," *Journal of the American Statistical Association* 64 (December 1969): 1183–1210.

10. Nancy J. Kirkendall, "Weights in Computer Modeling: A Method for Determining the Best Blocking Strategy," in *Record Linkage Techniques—1985: Proceedings of the Workshop on Exact Matching Methodologies* (Washington, D.C.: Federal Committee on Statistical Methodology, 1985), 189–97.

11. Max G. Arellano, Gerald R. Petersen, Diana B. Petitti, and Roger E. Smith, "The California Automated Mortality Linkage System (CAMLIS)," *American Journal of Public Health* 74, no. 12 (December 1984): 1324–30.

12. A. Wajda, L. L. Roos, M. Layefsky, and J. A. Singleton, "Record Linkage Strategies, Part II: Portable Software and Deterministic Matching," *Methods of Information in Medicine* 30 (1991): 210–14.

13. GEDCOM stands for Genealogical Data Communications and is a standard format used to exchange genealogical information between different computer programs.

14. White, "A Review of the Statistics," 367–68.

15. William W. Hinshaw, "Perquimans Monthly Meeting," *Encyclopedia of American Quaker Genealogy*, 7 vols. (Ann Arbor, Mich.: Genealogical Publishing, 1948), 1:1–90; William W. Hinshaw, "Pasquotank Monthly Meeting," *Encyclopedia of American Quaker Genealogy*, 1:91–178.

16. James R. B. Hathaway, ed., "Births, Deaths and Marriages in Berkeley, Later Perquimans Precinct, N.C.," *North Carolina Historical and Genealogical Register* vol. 3, nos. 2 and 3 (1903): 199–220, 363–410. (The original can be found in the courthouse at Hertford, N.C.)

17. Hinshaw, "Perquimans Monthly Meeting," 27.

18. Hathaway, "Births, Deaths and Marriages in Berkeley," 205.

19. Personal Ancestral File Release 2.3.1, The Church of Jesus Christ of Latter-day Saints. Direct inquiries to the Family History Department at 1-800-346-6044.

20. Maiden names were usually not given in the records we used and thus were not included.

21. In a flat file, all the information must be recorded on one line.

22. Microsoft Visual Basic, Microsoft Corporation, Redmond, Wash.

23. Brenda L. Price, "Probabilistic Methodology for Record Linkage: A Review of Theory and an Example" (master's project, Brigham Young University, 2000); Ryan T. Yamagata, "Probabilistic Methodology for Genealogical Record Linkage: Increasing Classification Rates and Decreasing Unclassified Rates" (master's project, Brigham Young University, 2001).