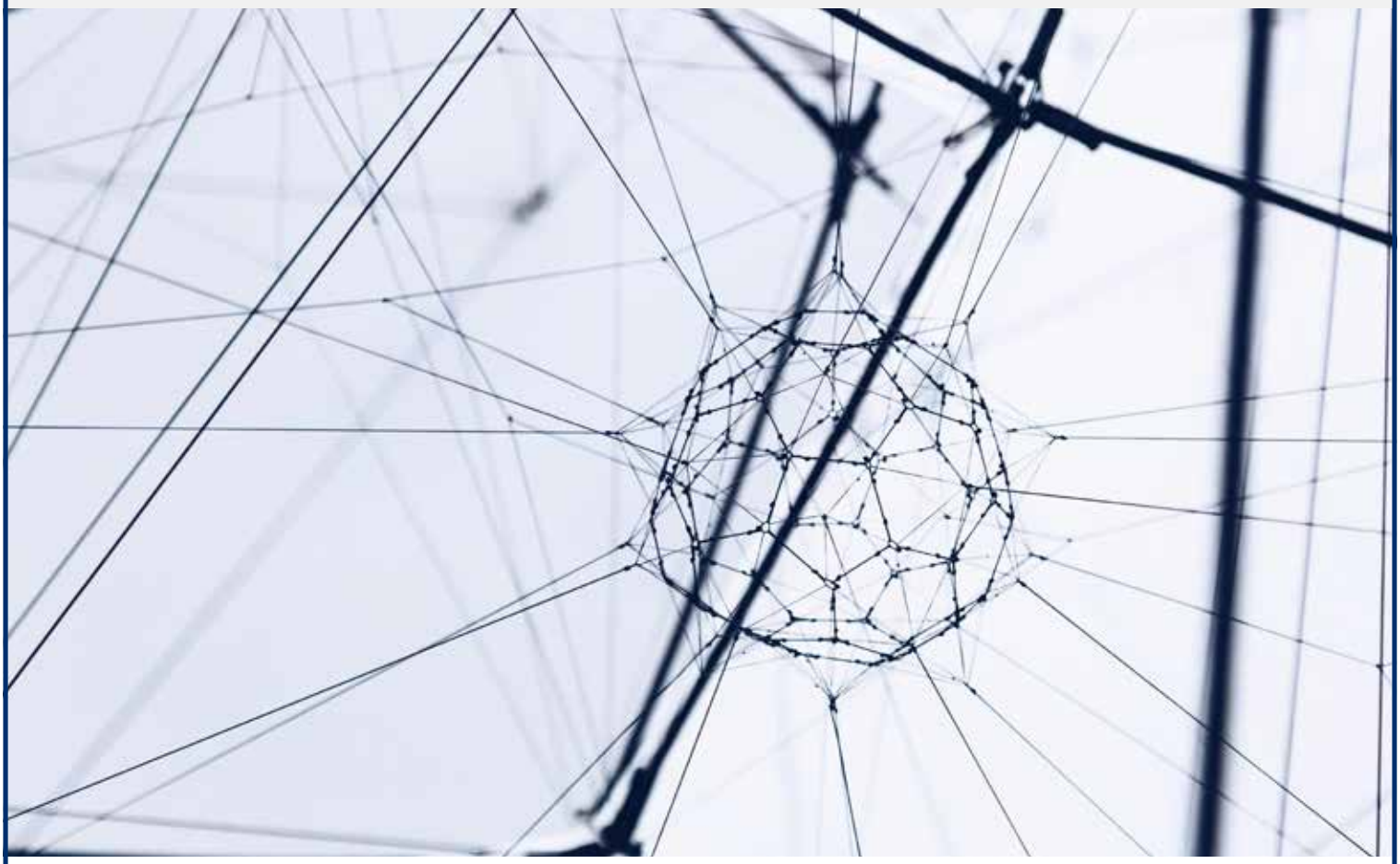
THE CONSORTIUM
stimulating self-sufficiency & stability scholarship

RECORD LINKAGE INNOVATIONS FOR THE HUMAN SERVICES

EMILY R. WIEGAND and ROBERT M. GOERGE | Chapin Hall at the University of Chicago



While an array of technical solutions and an extensive research base address problems of record linkage or deduplication, research and innovation tend to occur in discrete spaces, and often studies are published in very technical language. This report characterizes the landscape of approaches to record linkage, with particular attention to recent innovations. We aim to assist users and analysts working to combine state and local administrative data sources for analyses of family well-being and family self-sufficiency. We particularly highlight methods with the potential to increase either the accuracy or the efficiency of those matches.

FSSDC
Family Self-Sufficiency Data Center

CHAPIN HALL
AT THE UNIVERSITY OF CHICAGO

THE UNIVERSITY OF CHICAGO
Harris Public Policy

This research was supported by the Family Self-Sufficiency Research Consortium, Grant Number #90PD0272, funded by the Office of Planning, Research, and Evaluation in the Administration for Children and Families, U.S. Department of Health and Human Services. The Family Self-Sufficiency Data Center (FSSDC) facilitates the use of administrative data by researchers and administrators to improve understanding of and identify methods for increasing family well-being. The authors would like to thank Neil Miller, Shen Han, and David McQuown for research assistance. We also thank Julia Lane, Steven T. Cook, and Nick Mader for their review and insightful feedback on an early version of this report. The contents of this publication are solely the responsibility of the authors and do not necessarily represent the official views of the Office of Planning, Research, and Evaluation, the Administration for Children and Families, or the U.S. Department of Health and Human Services.

Recommended citation:
Wiegand, E. R. & Goerge R. M. (2019). *Record linkage innovations for the human services*. Washington, DC: Family Self-Sufficiency and Stability Research Consortium.

EXECUTIVE SUMMARY

Answering key questions in human services policy and research requires compiling data on the same individuals across a number of public sector programs, each with its own data system. Combining individual records across public sector data systems—such as those tracking public assistance programs, child welfare, education, criminal justice, and employment—can create data sources that fully contextualize a family's experience with public services, as well as their ultimate outcomes (Hotz, Goerge, Balzekas, & Margolin, 1998). The challenge in the United States is that a range of different agencies within state and local government administers these programs. This creates fragmentation of different data systems across programs and jurisdictions. Record linkage, the process of joining multiple records for the same entity, provides a methodological solution to this fragmentation.

Despite strong interest and increased investment in using record linkage for solving the problem of siloed data, there is little discussion of methods used to link data in the human services research community. However, there are rich histories of research on record linkage practices in public health and medical research and in large national data systems, as well as an extensive literature on the underlying theories in statistics and computer science. In this report, we translate between the scholarly and technical communities investigating record linkage methodologies and the community of researchers and practitioners seeking to use these methods in linking state and local data sources to study social and human services policy topics.

Motivation. During our interviews with subject matter experts, respondents described the struggle to balance accuracy, computational efficiency, and the resources required to execute record linkage.

Overall, respondents sought to reduce manual involvement in matches. A more streamlined, less manual process facilitates more frequent updates and even allows for sensitivity testing. Automation also reduces the chance for error or inconsistencies introduced by human reviewers. In most implementations represented in our interviews, 5–10 datasets are combined in a single, integrated database to allow for analyses that cut across a range of systems. This process requires particularly careful consideration of methodological changes, since decreases in efficiency will quickly accumulate.

However, just as inefficiencies can accumulate at scale, so can potential inaccuracies. For this reason, respondents often sought ways to better incorporate more data elements in links, adding to the richness of what the algorithm considers in an attempt to supplant a certain amount of manual review or consideration. Potential new fields to be added included multiple versions of the same data point (e.g., aliases), address history, and relationships and family connections. Currently, methodological requirements, computational limitations, and a lack of good models for how to incorporate these elements prevent the use of these additional fields. However, using additional fields could theoretically improve match quality.

We reviewed the literature and commercial solutions for examples of similar ideas or of methods that spoke to these linkage challenges.

Methodology. To collect the relevant literature, we searched a number of prominent research compilers, putting a particular emphasis on identifying reviews, recent papers (published since 2014), and research on new and different methodologies. Recognizing that the concept of record linkage has many names, we used an array of keywords popular in different disciplines. We also sourced citations from professional contacts, related research projects, and the U.S. Census Bureau's extensive record linkage resources.

Finally, we searched Google for record linkage or deduplication software solutions and reviewed any publicly available methodological documentation or references for those tools.

Background. In the absence of a single identifier with which to join two datasets, it is necessary to compare known characteristics for individuals between the datasets to identify likely pairs.

Deterministic record linkage uses a series of logical rules to define matches based on the level of similarity between different data elements in the two records. One of the great strengths of deterministic matching is that, because it is strictly rule-based, the rules can be clearly articulated. Furthermore, these algorithms scale well and can be run on large record sets; they can also be fully automated and routinely rerun. However, deterministic matching algorithms are very specific to their underlying datasets and generally must be custom written for implementation on any collection of data.

Traditional probabilistic matching attempts to address these limitations by giving the analyst the ability to describe component fields with parameters, which the model then applies to all comparisons of records, ranking the relative match quality of the results. The model may identify high-quality match combinations that would not have occurred to the author of a deterministic matching algorithm. In other words, probabilistic methods are better able to handle unexpected variation in data. They can be applied to new or changed datasets with less customization than deterministic matches. However, in exchange for this flexibility the analyst forfeits some control over what constitutes a match. Additionally, traditional probabilistic record linkage assumes there is no correlation between any of the data elements used in the comparison, which limits the range of fields that can be used in a single match.

Results. We organize the additional linkage methodologies we identified into two broad buckets: classification approaches and collective matching approaches. We also briefly discuss privacy-preserving record linkage.

Classification approaches are still focused on identifying pairs of records as matches or nonmatches: within the universe of all possible record pairs, there are a class of pairs that are matches, a class that are nonmatches, and (in many implementations), a class that are partial matches. Classification is a common statistical and machine learning problem. There are a variety of ways it has been addressed that go beyond traditional probabilistic record linkage. Unlike traditional probabilistic record linkage, classification approaches borrowed from machine learning do not assume that data elements being compared are independent.

The more commonly used classification algorithms in the record linkage space are supervised approaches. A supervised algorithm is one that requires a set of training data—in this case, a set of potential pairs of records that have already been identified as matches or nonmatches that the model can use to learn how to identify a match. Record linkage is an unusual classification problem in that the vast majority of record pairs are nonmatches so creating training data for record linkage is a major area of research. In particular, active learning methods identify the record pairs that a classifier is currently not able to classify—those pairs that theoretically would teach the classifier the most—and prioritize those pairs for manual review and determination.

We found limited evidence that classification methods have reached mainstream commercial application. There are few options and little guidance for analysts seeking to select or apply a classifier.

Collective matching approaches shift away from looking at record linkage as a strictly pairwise challenge, instead viewing the universe of records across the datasets to be linked as nodes in a graph (i.e., as a network). Some of those nodes represent the same individual and should be connected; isolated clusters within the graph represent unique individuals. These methods often use pairwise similarity measures to identify potential connections between any two records. However, decisions about which records are identified as a single entity are informed not only by the similarity of a given pair but also by things like the number of overall records that share that level of similarity. Steorts, Hall & Fienberg (2016) propose a linkage structure that goes beyond clustering existing records; instead, the structure matches records to inferred latent individuals. Thus, records are only indirectly matched to other records when two records are linked with the same latent entity.

Collective matching methods allow for simultaneously addressing both de-duplication (within one file) and record linkage (across more than one file). They also promote transitivity in the output of a record linkage. "Transitivity" is the property stating that if record pairs A-B and B-C are identified as matches, then A and C should match as well.

The biggest limitation to these types of match methodologies is that they are extremely computationally intensive. In general, while collective matching seems to be a promising area of research, it is still under development. Efficiencies and methods to scale for implementation are current research topics for the field (Mamun, Aseltine, & Rajasekaran, 2016).

We found no evidence that collective matching methods have reached mainstream application. Instead, these methods appear to be largely theoretical.

Privacy-preserving record linkage methods attempt to create linkage approaches that do not require direct access to identifiers. These methods thus may be uniquely suited to situations where data governance or privacy restrictions make access to the raw identifiers impossible. In general, privacy-preserving record linkage methods find ways to obscure identifying or sensitive information while still permitting comparison of records from different sources.

Discussion. Our review of recent research and innovations in record linkage theory suggest that developing methods could ultimately increase accuracy and efficiency for analysts linking human services data sources.

Classification approaches borrowed from machine learning do not retain the same assumptions as the probabilistic record linkage approach. This difference opens the door to a greater range of comparisons, including using data points that have high potential for natural correlation. As is usually the case when adding more and more complex predictors to a model, these techniques will have trade-offs in transparency and interpretability. By definition, classification approaches do not address scalability concerns, since these methods oftentimes still identify a class of potential matches requiring clerical review. Furthermore, a quality match with a supervised classifier requires good training data, and a number of unknowns about how to create and maintain high quality training data remain. Finally, while the literature contains a wide array of examples of different classifiers applied to record linkage and different approaches to developing training data, there are very few commercial implementations and little advice to help an analyst select the appropriate method.

While still very much under development, collective matching approaches to record linkage have particular potential for adding new records on to existing integrated datasets, with benefits for scalability. Similarly, these methods represent a promising solution for analysts seeking to merge more than two datasets at a time. Collective record linkage methods also represent promising opportunities to use relationship data more broadly in matches. However, we found no tools to allow implementation of these methods at present.

While privacy-preserving record linkage does not speak to the methodological challenges identified by our interviewees, it potentially addresses governance restrictions on the long-term growth of integrated administrative data for research and program management. These approaches allow for some level of matching records across datasets in which governance considerations prevent matching identifiers. In terms of allowing for maximum nuance in the match, however, even the best privacy-preserving approaches are handicapped compared to record linkage with full identifiers.

Recommendations. This report outlines a number of exciting alternatives to traditional record linkage that have the potential to address particular concerns for analysts linking human services datasets. However, options for application are severely limited. We are convinced that the field badly needs an active community of practice dedicated to record linkage methodology specific to this use case to guide practitioners seeking new approaches and document best practices for ensuring quality. We provide more information about our rationale and guidance for this recommendation in a companion statement (Wiegand & Goerge, 2019a).

INTRODUCTION

Record linkage, the process of joining multiple records for the same person across data sources, and deduplication, the specific case of record linkage applied to finding and merging multiple records for the same individual within a data source, are widespread challenges in the digital age. Librarians and archivists maintaining digital catalogs or combining catalogs across institutions wrestle with the best way to approach identifying unique records. Likewise, businesses who source customers from multiple sources and want a single source of contacts face many of the same challenges. As analysts seek to unlock new insights from data, the push to draw connections between data sources is growing.

The public sector maintains an extensive array of information systems, particularly in the human services, where program management necessitates creating databases of individual or family participation records. Researchers and policymakers are increasingly examining these administrative data sources, in isolation or in conjunction with conventional survey data, as rich mines of data on families, communities, and government services (Commission on Evidence-Based Policymaking, 2017).

Analyses of social policy topics related to family well-being and family self-sufficiency rely on data from assistance programs (income/cash, food, housing, medical, and child care), the child welfare system, education system, and criminal and juvenile justice systems, along with data sources on employment and earnings collected to administer programs like unemployment insurance and child support. In the United States, these data sources present particular challenges for record linkage because these programs are almost exclusively administered at the state and local levels.[1] The necessary data are fragmented not only by topic but also by jurisdiction. Furthermore, these data are collected in disparate contexts, including changing data systems and data collection processes and policies, leading to variable quality and consistency both within and across data sources.

We explore this use case for record linkage in the human services in more detail in a complementary report and recommendations statement (Wiegand & Goerge, 2019a, 2019b). Together, all three documents are grounded in a series of interviews with practitioners experienced in linking human services datasets, as well as the authors' own experience both conducting and providing technical assistance for these projects.

This report focuses on the details of record linkage methodology as it is applied to these data sources. Despite the strong interest in use of record linkage for these datasets, there is little discussion of linking methodology in the corresponding research community. However, there are rich histories of research on record linkage practices in health and in large national data systems, as well as extensive literature on the underlying theories in statistics and computer science.

We seek to translate between the scholarly and technical communities exploring record linkage and the community of researchers and practitioners seeking to use these methods in linking state and local data sources to study human services policy topics. We reviewed the statistical and computational literature on record linkage to understand alternate approaches being explored in the record linkage space, with a particular emphasis on seemingly new or innovative approaches such as machine learning techniques, graphical approaches, and privacy-protecting record linkage. In particular, we sought record linkage techniques that can make use of a wide array of elements, are robust to variations in data quality, and are appropriately scalable for linking large numbers of small data sources. These three facets reflect key characteristics of state and local administrative data with implications for record linkage.

The relevant literature is extensive, crosses several domains, and is consistently expanding. It is also extremely technical. We by no means intend this review to be comprehensive. Instead, we have simply tried to organize some of the key areas of methodology and innovation in language accessible to lay readers. We direct an interested reader to a variety of other recent publications that provide good surveys of the literature (Christen, 2007; Elmagarmid, Ipeirotis, & Verykios, 2007; Gu, Baxter, Vickers, & Rainsford, 2003; Hettiarachchi, Hettiarachchi, Hettiarachchi, & Ebisuya, 2014; Vatsalan, Christen, & Verykios, 2013). Some articles also include surveys of available tools and software (Elmagarmid et al., 2007; Gu et al., 2003; Herzog, Scheuren, & Winkler, 2007).

This paper begins with an overview of the methodological challenges and constraints identified by our interviewees that motivated this literature review.[2] We follow with detail on our methodology for the review. We provide background on deterministic and probabilistic record linkage, the techniques most widely in use. We then summarize a variety of innovations at various stages of development and discuss their potential to address our motivating challenges.

Limitations

The paper does not discuss legality, privacy, or ethics regarding using and linking administrative data sets (i.e., data governance). These are extremely important topics. They become only more pressing when we recognize that record linkage requires access to personal identifiers such as name, birthdate, and Social Security number (SSN)[3] and that record linkage in the human services often focuses on vulnerable populations like children and poor families. There are a number of industry best practices to reduce both ethical and data security risks in record linkage, including routinely separating personal identifiers from other data so that no analyst views identifiers and program or outcome data at the same time. We suggest Actionable Intelligence for Social Policy as a source for introductory resources on these topics.[4] The 2017 report from the federal Commission on Evidence-Based Policymaking also focuses on many of these issues (Commission on Evidence-Based Policymaking, 2017).

In our discussions of record linkage in the human services we concentrate on linking data sources for analytical purposes. While some of our findings are transferable to the challenges of linking data systems in real-time for case management purposes, that use case is not the focus of our research.

MOTIVATION: METHODOLOGICAL CHALLENGES IN HUMAN SERVICES RECORD LINKAGE

When we interviewed human services record linkage practitioners, respondents reported using a variety of approaches and tools (including programming languages and both proprietary and commercial software) to link data. These implementations struggle to balance accuracy, computational efficiency, and the resources required to execute the record linkage.

In particular, respondents expressed a desire to reduce manual involvement in matches. A more streamlined, less manual process can facilitate more frequent updates and even sensitivity testing. It can also reduce the chance for error or inconsistencies introduced by human reviewers. Current approaches to record linkage balance a tension between scalability and rigor. This tension limits rapid implementation or expansion in number of datasets, population size, or match frequency. In most implementations represented in our interviews, 5–10 datasets were combined in a single integrated database and regularly updated to allow for analyses that cut across a range of systems. Record linkage approaches with heavy manual components are extremely resource intensive at this scale.

[1] The only federally administered exceptions are Medicare and Head Start.

[2] More detail about the interview process and methodology is discussed in Wiegand & Goerge (2019b).

[3] We do discuss privacy-preserving record linkage, a variety of techniques that test the assumption that identifiers must be available to conduct a match, in our review of record linkage methodologies (see "Results").

[4] See www.aisp.upenn.edu.

Respondents also recognize that their current, manual approaches are in response to the challenges and limitations of their data (Wiegand & Goerge, 2019b). Many of the datasets involved in these matches contained limited unique identifiers (like SSN or driver's license number) with variable data quality. Manual processes allow analysts to be very hands-on, monitoring and adjusting matches to ensure a high-quality result.

One approach that respondents thought would increase their confidence in algorithmically assigned links would be to use more detail from records in the match. Potential sources of additional information include:

- Alternate values for the same field: Many administrative systems track not only current values for fields like name, but also track historical information or aliases. Alternately, two records that have been identified as the same person may differ on a key identifier like SSN or birthdate. Assuming the analyst does not know the correct value, it would be useful to consider subsequent matches based on whether they match to either of the two potential values.
- Address data (current and historical): For an individual looking at two records clerically, identifying a common address could provide the missing link needed to confirm an otherwise borderline match.
- Relationships: Where relationship information has been included in matching, records are usually limited to the child, and mother or father's information is used as a match characteristic. If the dataset includes the full population, or exact relationships are not known, is it possible to capture common family members as a match characteristic? In theory, relationship information could be used both to identify borderline matches that represent true pairs and to distinguish individuals with strong similarity between their records (such as siblings and spouses).

Currently, methodological requirements, computational limitations, and a lack of good models for how these elements can be rigorously incorporated prevent the use of these additional fields in matches. As we reviewed the literature and commercial solutions, we looked for examples of similar ideas or of methods that spoke to these linkage challenges.

METHODOLOGY

Literature Review

We searched a number of prominent research compilers, such as SAGE journals, Springer, arXiv.org, and Google Scholar to explore existing methodological research on record linkage. We put a particular emphasis on identifying reviews, recent papers (published since 2014), and research on new and different methodologies. Recognizing that the concept of record linkage has many names and is called different things by different disciplines, we searched on a range of keywords: "record linkage", "entity disambiguation", "deduplication", "entity reconciliation," "entity matching," and "merge/purge problem." Because the practitioner community expressed interest in record linkage approaches using machine learning, we also did some specific searches for "machine learning record linkage" and for record linkage approaches using specific machine learning methods (e.g., "support vector machines record linkage"). We also sourced citations from professional contacts, related research projects, and the U.S. Census Bureau's extensive record linkage resources.

After our primary search, we followed references and citations within the articles to identify other relevant sources, again with an emphasis on finding reviews and recent articles.

Survey of Commercial Products

Finally, we searched Google for record linkage or deduplication software solutions and reviewed any publicly available methodological documentation or references for those tools. We did not do a comprehensive review of commercial record linkage software (something that is difficult to undertake because these solutions are often semicustom or presented by vendors as part of a broader data management solution). Our primary goal was to identify any major commercial players applying atypical or innovative techniques that had not been sourced in the literature. We did not find any software or tools that met these criteria.[5]

BACKGROUND: DETERMINISTIC AND PROBABILISTIC RECORD LINKAGE TECHNIQUES

For purposes of clarity, we begin with a brief discussion of traditional record linkage approaches.[6] Our interview respondents' implementations usually used a combination of tools including both probabilistic and deterministic methods.

A number of quality tutorials and overviews of these methods already exist, and our intention is not to recreate those resources. Instead, we aim to provide background for the other methods discussed.

In the absence of a single identifier on which to join two datasets, it is necessary to compare known characteristics for individuals between the datasets to identify likely pairs. As presented in Table 1, any record linkage process balances requiring strong evidence that two records represent the same person to prevent false matches (Type 1 error) with allowing for natural variation or data error between records that represent the same person to avoid missing correct matches (Type 2 error).

Table 1. Possible Results for Any Record Comparison (Confusion Matrix)

	Records are matched	**Records are not matched**
Person is the same	Successful match	Missed match (Type 2 error)
Person is not the same	False match (Type 1 error)	Successful non-match

The basic element of all record linkage techniques is the comparison between two values for the same field to determine if the values match: for example, comparing first name between two records, or comparing those records on gender. For fields like gender, this comparison is a straightforward operation with a true/false result. For fields like first name, however, the analyst has choices: names can be compared strictly (with a true/false result); they can be standardized using a technique like Soundex or NYSIIS to reduce the chance of spelling errors and then compared strictly (again with a true/false result); or they can be compared using various string metrics (e.g., Jaro-Winkler distance, Levenshtein distance, etc.) to identify partial matches. In the case of partial matches, the possible results are true (1), false (0), or some partial similarity measure between 0 and 1.

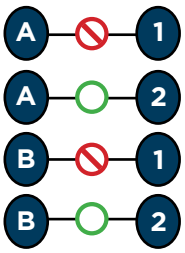
Record linkage techniques and algorithms represent different approaches to how these individual item-level comparisons are rolled up into record-level comparisons and which records are compared in the first place.

[5] Readers exploring record linkage software options may find it useful to read Chapter 19 of Herzog et al., 2007, which in addition to listing some software packages provides a checklist for evaluating and thinking about record linkage software.

[6] The distinction between deterministic and probabilistic matching has become blurred in the literature, in part by the introduction of the term "fuzzy matching." "Fuzzy matching" is used in reference both to traditional probabilistic record linkage techniques and to deterministic matching that uses string metrics and accepts partial matches. To more cleanly demarcate these approaches, we limit the term "probabilistic record linkage" to a series of methodologies that are derived from the formal record linkage theory developed by Howard Newcombe, Ivan Fellegi, and Alan Sunter. See Newcombe, Kennedy, Axford, & James (1959) and Fellegi & Sunter (1969) for seminal articles or Jaro (1989) and Winkler (1999) for good general discussions of this theory.

RECORD LINKAGE METHODS

Deterministic Record Linkage

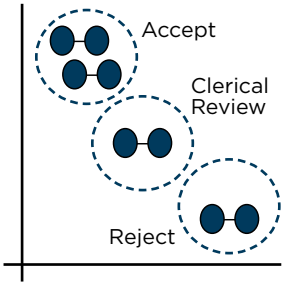


Pairs are individually reviewed and matched based on defined rules.

Pros: Transparent and scalable

Cons: Logic must be custom written for each match and changed if data change

Classification Approaches

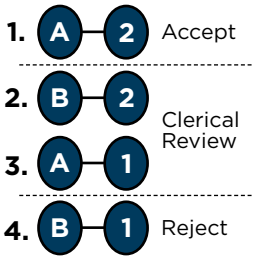


Pairs are classified according to a model derived from training data.

Pros: Can incorporate more data elements without assumption of independence

Cons: No best practices in selecting model or developing training data; limited commercial options; may require clerical review

Probabilistic Record Linkage

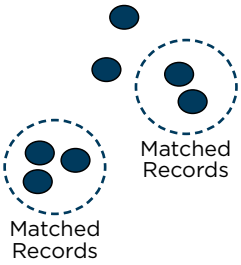


Pairs are scored and ranked based on likelihood of a match. Analyst sets cut-offs.

Pros: Requires less customization

Cons: Requires that data elements compared between records are statistically independent; may require clerical review

Collective Matching Approaches



Match determinations reflect not only the similarity between a pair of records but also how similar other records are to those.

Pros: Can be used to deduplicate and link at the same time or to link multiple datasets at once

Cons: Challenges of computational efficiency; still largely theoretical

Deterministic Record Linkage

The simplest deterministic match is a database join on a single value: a single item (such as SSN) is compared between two records, and then a true match on SSN indicates a match and a false match on SSN indicates a nonmatch. This deterministic match could be made more complex by allowing for a partial SSN match with a Jaro-Winkler distance (or similar) below a certain value. These matches become still more complex by including comparisons of other fields and a series of logical rules to distinguish matches. At their most complex, deterministic matching algorithms may compare a wide array of variables across record sets and accept various combinations of partial or true matches. For example, records could be considered linked if they have:

- any full SSN match; OR
- an SSN match under a certain threshold combined with a perfect match on Soundex of first and last name; OR
- at least one null SSN and first and last name matches under a certain threshold, combined with a birth day and month match and birth year values within +/- two years of one another.

From this example, it is easy to see how a deterministic matching algorithm can grow to significant length and detail.

One of the great strengths of deterministic matching is that, because it is strictly rule-based, the rules can be clearly articulated (if at some length) for discussion or documentation. Narrow rules may be added to address specific quality issues. Because these algorithms perform minimal calculations, they scale well and can be run on large record sets. They can also be fully automated and routinely rerun (although users may wish to change the logic behind the algorithm to react to new or unanticipated data quality challenges).

However, deterministic matching algorithms are very specific to their underlying datasets and generally must be custom written for implementation on any collection of data. Creating a finely tuned deterministic approach requires a deep knowledge of data quality and contents for the elements that will be included in the match and the ability to articulate rules that appropriately balance Type 1 and Type 2 errors within those datasets. If there is a significant change in the data contents or quality, the matching algorithm must be revisited.

Probabilistic Record Linkage

Traditional probabilistic matching attempts to address these limitations by giving the analyst the ability to describe component fields with parameters, which the model then applies to all pairs of records, ranking the relative match quality of the pairs.[7] Instead of itemizing all possible definitions of a match (as in the deterministic example), the analyst assigns match probabilities to the fields (i.e., SSN, first name, last name, and birth day). Then, the model uses these parameters to rate the likelihood of a true match for every pair based on the similarity of all fields, weighted by the assigned probabilities.

The model may identify high-quality match combinations that would not have occurred to the author of a deterministic matching algorithm, as well as pairs that would have met a set of deterministic criteria but for other reasons represent a poor quality match. In other words, probabilistic methods are better able to handle unexpected variation in data. They can be applied to new or changed datasets

[7] How the input parameters are defined varies with the technical implementation of probabilistic record linkage. Common methods include assignment using simple probabilities from the raw data, calculation from a small sample of manually reviewed matches, and assignment based on prior experiences with similar datasets. In any of these cases, parameters are refined iteratively based on match results. Alternately, the expectation maximization (EM) algorithm iterates using maximum likelihood estimation to optimize input parameters.

with less customization than deterministic matches. In exchange, the analyst forfeits some control over what constitutes a match. The key decision points for a probabilistic match are:

1) What is compared: Because the probabilistic model requires a calculation for every pair of records compared, it is usually infeasible computationally to compare every record to every other record in a dataset. The analyst defines blocking passes—record sets of comparison (e.g., "compare all records that have the same SSN"). In practice, defining blocking criteria and interpreting the results of each pass with more or less strictness may allow the analyst to impose deterministic-style criteria on the match results, with the attendant strengths (transparency, efficiency) of a rule-based match as well as the weaknesses (e.g., limitations on application to new data sources).[8]

2) What comparisons are accepted: For each blocking pass, the probabilistic model ranks the comparison set of results by strength of match. The analyst decides which matches are accepted. This decision can be made by setting a cut-off, with matches greater than the cut-off being accepted and those less than the cut-off being rejected. But individuals implementing record linkage often speak of the grey area: records that are above a certain point are accepted without question; records that are below a certain point are rejected without question; and, in many implementations, records that fall between these two cut points are subject to clerical review. In this case, a human reviews each pair and decides to accept or reject the pair. Clerical review is, obviously, an extremely manual process. There are no options for scalability and there may be questions of inter-reviewer reliability.

The approach taken on these two questions determines the overall scalability and generalizability of the model, the extent to which it is customized to features of the data, and the transparency and interpretability of the result.

One particular limitation of traditional probabilistic record linkage is the assumption of independence among variables included in the model. In theory, there should be no correlation between any of the variables used in the comparison. This assumption means, for example, that it would be inappropriate to use both first name and nickname in the match, since a match on one of these fields is very likely to correlate with a match on another field. This would effectively increase the weight placed on first name.

RESULTS

In this section, we summarize findings from our literature review and review of commercial options. Much of the taxonomy we present below is derived from a particularly helpful 2013 tutorial (Getoor & Machanavajjhala, 2013).

While practitioners have expressed interest in learning about machine learning methods for record linkage, it is important to emphasize the breadth of the term "machine learning" and the potential variation in associated techniques. Machine learning algorithms develop models based on samples of data. Ultimately, a "machine learning" method could look extremely similar to traditional probabilistic record linkage. The expectation-maximization algorithm frequently used to optimize parameters for probabilistic matching (see Jaro, 1989, for an early use of this algorithm) is a statistical model commonly used in machine learning and data mining applications. Furthermore, the Felligi-Sunter algorithm, which underlies traditional probabilistic record linkage, is generally equivalent to machine learning's naïve Bayes classifier (Wilson, 2011; Winkler, 2002). As a result of this ambiguity, we do not use the term "machine learning" as part of how we classify or define record linkage approaches. However, we note that analysts are most commonly thinking of supervised classifiers when they say "machine learning."

Classification Approaches

Like traditional record linkage methods, classification approaches compare pairs of records to identify those pairs as matches or nonmatches. Within the universe of all possible record pairs (or a subset of pairs defined by some blocking criteria), there are a class of pairs that are matches, a class that are nonmatches, and (in many implementations), a class that are partial matches. Classification, the idea of finding underlying groups in disparate data points, is a common statistical and machine learning problem; there are a variety of ways it has been addressed that go beyond traditional probabilistic record linkage.

Unlike traditional probabilistic record linkage, classification approaches borrowed from machine learning do not assume that the features (fields) being compared are independent of one another (e.g., that there is no correlation between the likelihood that one field matches and the likelihood that another field matches).

The more commonly used classifiers (classification algorithms) in the record linkage space are supervised approaches, particularly decision trees and support vector machines, though also neural networks (see, for example, Elfeky et al., 2003; Elmagarmid et al., 2007; Feigenbaum, 2016; Goeken, Huynh, Lenius, & Vick, 2011; Wilson, 2011).[9] A supervised algorithm is one that requires a set of training data—in this case, a set of potential pairs of records that have already been identified as matches or non-matches that the model can use to learn how to identify a match.

Creating training data for record linkage is a major area of research. Record linkage is an unusual classification problem in that the vast majority of randomly selected pairs of records are nonmatches. If pairs are randomly sampled for manual coding to create training data, as would be typical practice for supervised methods, the size of the sample needed to accurately capture what defines a match could be prohibitively large. Feigenbaum tested a variety of supervised models, including probit, random forest, and support vector machines, and also used cross-validation to assess how much training data were necessary for model accuracy. For his match, he determined that training on about 10% of possible pairs is optimal (Feigenbaum, 2016).

Any attempt to create training data from obvious matches and nonmatches runs the risk of not including the most difficult-to-classify pairs. Theoretically these are the pairs that would teach the classifier the most about how to handle similarly ambiguous pairs, reducing the size of the potential match class and the necessary manual review. Active learning methods are applicable here. A learner algorithm identifies the record pairs that a classifier is currently not able to classify and that will thus strengthen the classifier as quickly as possible. Sarawagi and Bhamidipaty explored active learning in deduplication and developed an approach that required fewer than 100 pairs, iteratively selected, to be hand-coded for peak accuracy in their dataset, compared with over 7,000 pairs needed when the training data were selected as a

[8] This paper does not discuss blocking approaches in detail, although there is a body of research and innovation in this space just as in the other parts of record linkage. Blocking methods have particular implications for the scalability and computational requirements of a match. For accessible surveys of research on blocking and other indexing techniques see Christen, 2012 and Steorts et al., 2014.

[9] The ChoiceMaker record linkage software, one of the few record linkage implementations often referenced in the human services space that does not use traditional probabilistic or deterministic matching, uses a supervised maximum entropy model (Borthwick, Buechi, & Goldberg 2003).

random sample (2002). The authors note the value of a "covering and challenging" training set that can "bring out the subtlety of the deduplication function" (p.270).

Unsupervised classification algorithms do not require training data, mitigating this cost. An example of unsupervised classification is traditional probabilistic record linkage with parameters defined by the expectation maximization algorithm. Elfeky, Verykios, and Elmagarid provide one example of another unsupervised approach (k-means clustering), but found it was generally outperformed by a supervised model (Elfeky et al., 2003). However, they also experimented, with some success, with a hybrid approach. This approach used an unsupervised method to identify obvious matches and nonmatches and then used these for training data for a supervised classifier (see Christen, 2007 for a similar approach).

Collective Matching and Graph Approaches

A more recent body of research shifts away from looking at record linkage as a strictly pairwise challenge. Instead, collective approaches think of the universe of records across the datasets to be linked as nodes in a graph (i.e., a network). Some of those nodes represent the same individual and should be connected; isolated clusters within the graph represent unique individuals. These methods often use pairwise similarity measures to identify potential connections between any two records. However, decisions about which matches are ultimately accepted, and which records are identified as a single entity, are informed not only by the similarity of a given pair but also by things like the number of overall records that share that level of similarity. These approaches are particularly robust for deduplication or matching multiple datasets at the same time, since they consider a group of like records collectively, rather than looking only at pairs.

Steorts, Hall, & Fienberg (2016) proposed a linkage structure that goes beyond clustering the existing records. Instead, records are matched to inferred latent individuals, unobserved entities that only show up in observed records with some margin of data error. Records are thus only indirectly matched to other records when two records are linked to the same latent entity. As with other collective matching approaches, this method simultaneously addresses both deduplication (within one file) and record linkage (across more than one file). It also promotes transitivity in the output of a linkage, the property that if record pairs A-B and B-C are identified as matches, then A and C should match as well.

Sadinle took a similar approach and compared the results to traditional probabilistic matching. He found that probabilistic methods were significantly faster but collective approaches had better performance, particularly in cases where there was low overlap between the two datasets being matched (Sadinle, 2017).

The biggest limitation to these types of match methodologies is that they are extremely computationally intensive. In general, while collective matching seems to be a promising area of research, it is still under development and largely theoretical. Efficiencies and methods to scale for implementation are current research topics (Mamun et al., 2016).

Commercial Applications

In practice, we found very little evidence that the classification and graph theory methods described above have reached mainstream application. In addition to our interviews, we also conducted a series of web searches and explored documentation for various commercial products. Although there are plenty of products that promise "proprietary match algorithms," what we were able to find across products suggests frequent and extensive application of probabilistic record linkage, with various product-specific applications—such as special comparison methods—customized to certain data types and imputation or standardization tools for common fields. A notable exception is the ChoiceMaker tool, which implements a supervised classifier (Borthwick et al., 2003).

Privacy-Preserving Record Linkage Methods

Privacy-preserving record linkage methods attempt to create linkage approaches that do not require direct access to identifiers. Thus, these methods may be uniquely suited to situations in which data governance or privacy restrictions make access to the raw identifiers impossible.

Much of the current research and theory in privacy-preserving record linkage is concerned with broader privacy and cryptography standards. There are myriad ways of defining acceptable levels of exposure, including many questions of what linkage or analytic results are acceptable for each party to see and whether data can be held by a third party as a way of preserving privacy (see Hall & Fienberg, 2010 and Vatsalan, Christen, & Verykios, 2013 for a broader overview). We focus primarily on the linkage process itself to understand the ways in which privacy-preserving record linkage approaches are or are not compatible with the types of linking methodologies discussed so far in this paper.

In general, privacy-preserving record linkage methods find ways to obscure identifying or sensitive information while still permitting comparison of records from different sources. One of the most common approaches is to use secure hash encodings. Hashed values of fields preserve the uniqueness of the underlying fields (i.e., the same string will always have the same hashed value), but once a value has been hashed, it is not possible to recover the underlying value. Hashed values can be shared and compared deterministically, but because changing even one character in the input value completely changes the hash, it is not possible to use them for any partial comparison. Hash approaches can be adapted for string comparison, however, if, for example, words are broken up into sets of n-grams (smaller sequences of *n* items—characters, syllables, etc., depending on the context), which are then hashed for comparison. Bloom filters operate similarly to hashes but with benefits in space and efficient use of memory (Hall & Fienberg, 2010).

DISCUSSION

The known limitations of probabilistic record linkage methods—i.e., the scenarios under which the generalized mathematical solutions fail to hold—include "lack of sufficient variables for matching, sampling or lack of overlap in lists, and extreme variations in the messiness of the data" (Winkler, 1999, p.6). These are among the quality challenges most prevalent in administrative data. In practice, these limitations in the datasets are generally addressed by use of deterministic rules to structure probabilistic matches together with extensive clerical review.

Our review of recent research and innovations in record linkage theory suggest that developing methods could ultimately increase accuracy by providing opportunities to expand the range of data elements used in the links or increase the efficiency of those links.

Classification approaches borrowed from machine learning do not retain the same assumptions as the probabilistic record linkage approach. For example, they do not assume that the data elements being compared are independent of one another. In practical terms, record linkage implementations often ignore some violations in the requirement of independence without an obvious cost in accuracy; for example, we generally suspect that data entry error is concentrated in certain records, so that if one field has a mistake another field is more likely to have a mistake too. Removing this

requirement altogether opens the door to incorporating many more data points. Nicknames and family member's names both correlate with an individual's name, but also provide potentially significant distinguishing data for a match. Even complex logic, like comparing whether birth date in one dataset falls before death date in another, becomes more feasible with this record linkage approach. This kind of model is potentially more robust for including things like alternate names, relationships, and addresses, all of which were possible new inputs identified by interviewees. Wilson writes of the application of a neural network with more detailed features on a sample of genealogical records, "To see improvements this consistent and this dramatic indicates that those still using traditional record linkage need to take notice and use more powerful features and better classifiers to improve accuracy" (2011, p. 14).

As is usually the case when adding more and more complex predictors to a model, these techniques will have trade-offs in transparency and interpretability. Nor do classification approaches by definition address scalability concerns, since these methods oftentimes still identify a class of potential matches requiring clerical review.

Furthermore, a quality match with a supervised classifier requires good training data. While smart applications of active learning can reduce this burden, a number of unknowns about high-quality training data remain. How often does a training set need to be refreshed? How extensible across new or different datasets could training data or the model derived from that data be? How can the accuracy of training data be understood given the lack of a clear truth and the potential for inconsistencies across coders in creating training sets? For people or organizations looking to regularly link a variety of datasets, these are important and unanswered questions.

Finally, while the literature contains a wide array of examples of different classifiers applied to record linkage and different approaches to developing training data, there are few commercial implementations and little advice to help an analyst select the appropriate method.

While still very much under development, collective matching approaches to record linkage may one day present opportunities to approach versions and match frequency in a new way. While these approaches are very resource intensive at first, they can be adapted for ideal incremental matching: adding new records to an existing match without rerunning the entire process, while allowing for changes in the old match results in response to new information. To the extent that a collective approach uses an underlying list of pairwise similarities, new pairs can be added to an existing list and then all pairs input into the clustering algorithm (Gruenheid, Dong, & Srivastava, 2014).

Beyond introducing data on family members (e.g., parents, children, spouses) as separate data points, collective record linkage methods also represent promising opportunities to use these data more broadly. For example, comparing who an individual shares a household with across multiple data sources could help confirm that the records are the same (Bhattacharya & Getoor, 2007).

In the long term, as they reach mainstream implementation, collective methods also represent a promising solution for analysts who seek to merge more than two datasets at a time. In the short term, Sadinle and Fienberg present an expansion of the Felligi-Sunter theorem applicable across more than two datasets (Sadinle & Fienberg, 2013).

While privacy-preserving record linkage does not speak to the methodological challenges identified by our interviewees, it potentially addresses restrictions on access to personal identifiers. These restrictions serve as one set of barriers to long-term growth of integrated administrative data for research and program management. Our respondents already cited datasets (particularly wage data) where they faced these restrictions and had to share identifiers with an external party and rely on some else's link methodology. Depending on the level of information that is allowed to be shared between parties, it is possible to apply various record linkage approaches to data in hashed values or Bloom filters to identify pairs with high levels of similarity.

In practice, however, privacy-preserving record linkage approaches reduce the amount of detail and variation in the data before a match. These approaches allow for some level of matching records across datasets where governance considerations prevent matching of identifiers. However, in terms of allowing for maximum nuance in the match, even the best privacy-preserving approaches are handicapped relative to record linkage with full identifiers. Brown, Randall, Ferrante, Semmens, and Boyd show that it is possible to use privacy-preserved datasets without major loss of accuracy in record linkage (2017), but there are a lack of benchmarks by which to truly compare the performance of various record linkage options for different purposes.[10]

RECOMMENDATIONS

This report outlines a number of potentially fruitful alternatives to traditional record linkage to address particular concerns for analysts linking human services datasets. However, options for application are severely limited. Collective matching techniques are largely still theoretical. While a number of researchers have demonstrated promising applications of classifiers, there have been few comparisons across different approaches, little is known about how to create effective training data, and commercialization is extremely limited.

There is no clear methodological path forward for practitioners of human services record linkage. We believe an active community of practice dedicated to record linkage methodology specific to this use case is urgently needed to provide this guidance and document best practices for practitioners. We provide more information about our rationale and detail on this recommendation in a companion statement (Wiegand & Goerge, 2019a).

[10] See Wiegand & Goerge (2019a) for more on the challenges of comparing performance across methods.

REFERENCES

Bhattacharya, I., & Getoor, L. (2007). Collective entity resolution in relational data. *ACM Transactions on Knowledge Discovery from Data, 1*(1), 5-es. https://doi.org/10.1145/1217299.1217304

Borthwick, A., Buechi, M., & Goldberg, A. (2003). *Key concepts in the Choicemaker 2 Record Matching System*. Retrieved from http://dc-pubs.dbs.uni-leipzig.de/files/Borthwick2003KeyConceptsintheChoiceMaker.pdf

Brown, A. P., Randall, S. M., Ferrante, A. M., Semmens, J. B., & Boyd, J. H. (2017). Estimating parameters for probabilistic linkage of privacy-preserved datasets. *BMC Medical Research Methodology, 17*(1), 95. https://doi.org /10.1186/s12874-017-0370-0

Christen, P. (2007). A two-step classification approach to unsupervised record linkage. *Conferences in Research and Practice in Information Technology, 70*, 107–116. Retrieved from http://crpit.com/confpapers/CRPITV70Christen.pdf

Christen, P. (2012). A survey of indexing techniques for scalable record linkage and deduplication. *IEEE Transactions on Knowledge and Data Engineering, 24*(9), 1537–1555. https://doi.org/10.1109/TKDE.2011.127

Commission on Evidence-Based Policymaking. (2017). *The promise of evidence-based policymaking: Report of the Commission on Evidence-Based Policymaking*. Retrieved from https://www.cep.gov/content/dam/cep/report/cep-final-report.pdf

Elfeky, M. G., Verykios, V. S., Elmagarmid, A. K., Ghanem, T. M., Huwait, A. R., Verykios, V., & Elmagarrnid, A. K. (2003). *Record linkage: A machine learning approach, a toolbox, and a digital government web service*. Retrieved from http://docs.lib.purdue.edu/cstech/1573

Elmagarmid, A. K., Ipeirotis, P. G., & Verykios, V. S. (2007). Duplicate record detection: A survey. *IEEE Transactions on Knowledge and Data Engineering, 19*(1), 1–16. Retrieved from https://www.cs.purdue.edu/homes/ake/pub/TKDE-0240-0605-1.pdf

Feigenbaum, J. J. (2016). *Automated census record linking: a machine learning approach*. Retrieved from https://open.bu.edu/handle/2144/27526

Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association, 64*(328), 1183–1210. https://doi.org/10.1080/01621459.1969.10501049

Getoor, L., & Machanavajjhala, A. (2013). *Entity resolution for big data*. Retrieved from http://users.umiacs.umd.edu/~getoor/Tutorials/ER_KDD2013.pdf

Goeken, R., Huynh, L., Lenius, T., & Vick, R. (2011). New methods of census record linking. *Historical Methods, 44*(1), 7–14. https://doi.org/10.1080/01615440.2010.517152

Gruenheid, A., Dong, X. L., & Srivastava, D. (2014). Incremental record linkage. *Proceedings of the VLDB Endowment, 7*(9), 697–708. https://doi.org/10.14778/2732939.2732943

Gu, L., Baxter, R., Vickers, D., & Rainsford, C. (2003). Record linkage: Current practice and future directions. *CSIRO Mathematical and Information Sciences Technical Report, 3*(83).

Hall, R., & Fienberg, S. E. (2010). Privacy-preserving record linkage. *Privacy in Statistical Databases: UNESCO Chair in Data Privacy International Conference Proceedings*, 269–283. https://doi.org/10.1007/978-3-642-15838-4_24

Herzog, T. N., Scheuren, F. J., & Winkler, W. E. (2007). *Data quality and record linkage techniques*. New York, NY: Springer.

Hettiarachchi, G. P., Hettiarachchi, N. N., Hettiarachchi, D. S., & Ebisuya, A. (2014). Next generation data classification and linkage: Role of probabilistic models and artificial intelligence. *Proceedings of the 4th IEEE Global Humanitarian Technology Conference, GHTC 2014*, 569–576. https://doi.org/10.1109/GHTC.2014.6970340

Hotz, V. J., Goerge, R. M., Balzekas, J., and Margolin, F. (Eds.). (1998). *Administrative data for policy-relevant research: Assessment of current utility and recommendations for development*. Chicago: Northwestern University and the University of Chicago Joint Center for Poverty Research.

Jaro, M. A. (1989). Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. *Journal of the American Statistical Association, 84*(406), 414–420. https://doi.org/10.1080/01621459.1989.10478785

Mamun, A.-A., Aseltine, R., & Rajasekaran, S. (2016). Efficient record linkage algorithms using complete linkage clustering. *PLOS ONE, 11*(4), e0154446. https://doi.org/10.1371/journal.pone.0154446

Newcombe, H. B., Kennedy, J. M., Axford, S. J., & James, A. P. (1959). Automatic linkage of vital records. *Science, 130*(3381), 954–959. https://doi.org/10.1126/SCIENCE.130.3381.954

Sadinle, M. (2017). Bayesian estimation of bipartite matchings for record linkage. *Journal of the American Statistical Association, 112*(518), 600–612. https://doi.org/10.1080/01621459.2016.1148612

Sadinle, M., & Fienberg, S. E. (2013). A generalized Fellegi-Sunter framework for multiple record linkage with application to homicide record systems. *Journal of the American Statistical Association, 108*(502), 385–397. https://doi.org/10.1080/01621459.2012.757231

Sarawagi, S., & Bhamidipaty, A. (2002). Interactive deduplication using active learning. *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 269–278.

Steorts, R., Hall, R. & Fienberg, S. (2016). A Bayesian approach to graphical record linkage and deduplication. *Journal of the American Statistical Association, 111*(516): 1660-1672.

Steorts, R. C., Ventura, S. L., Sadinle, M., & Fienberg, S. E. (2014). A comparison of blocking methods for record linkage. In J. Domingo-Ferrer (ed.), *Privacy in statistical databases,* 253–268. https://doi.org/10.1007/978-3-319-11257-2_20

Vatsalan, D., Christen, P., & Verykios, V. S. (2013). A taxonomy of privacy-preserving record linkage techniques. *Information Systems, 38*(6), 946–969. https://doi.org/10.1016/J.IS.2012.11.005

Wiegand, E. R., & Goerge, R. M. (2019a). *Recommendations for ensuring the quality of linked human services data sources*. Washington, DC: Family Self-Sufficiency and Stability Research Consortium.

Wiegand, E. R., & Goerge, R. M. (2019b). *Using and linking administrative datasets for family self-sufficiency research*. Washington, DC: Family Self-Sufficiency and Stability Research Consortium.

Wilson, D. R. (2011). Beyond probabilistic record linkage: Using neural networks and complex features to improve genealogical record linkage. *Proceedings of the International Joint Conference on Neural Networks*, 9–14. https://doi.org/10.1109/IJCNN.2011.6033192

Winkler, W. E. (1999). *The state of record linkage and current research problems*. Retrieved from https://www.census.gov/srd/papers/pdf/rr99-04.pdf

Winkler, W. E. (2002). *Methods for record linkage and Bayesian networks.* Retrieved from https://www.census.gov/srd/papers/pdf/rrs2002-05.pdf