RESEARCH REPORT SERIES
(Statistics #2024-05)

A Review of Modern Multinomial-Derived and Partition-Based Record-Linkage Methods

Yves Thibaudeau

Center for Statistical Research & Methodology
Research and Methodology Directorate
U.S. Census Bureau
Washington, D.C. 20233

Report Issued: August 15, 2024

Disclaimer: This report is released to inform interested parties of research and to encourage discussion. The views expressed are those of the authors and not those of the U.S. Census Bureau.

A Review of Modern Multinomial-Derived and Partition-Based Record-Linkage Methods

Yves Thibaudeau, CSRM, U.S. Census Bureau

Abstract

Fellegi and Sunter (1969) introduced the first theory of record linkage. Their work was interpreted and applied in many situations. But an infrastructure to support generalizing the F-S theory is not available until 40 years later. Sadinle and Fienberg (2013) discuss partitioning as a possible infrastructure for record linkage. Partitioning makes clear the limitations of F-S and paves the way to generalizing it. Partitioning also fits naturally in the Bayesian paradigm. We review the advent and instances of partitioning in the recent literature in particular in the context of linking applications connected to Dirichlet-Multinomial processes. We suggest directions for the future.

1. Introduction

Automated record linkage goes back to early computing. As computational power expanded record linkage became a discipline in itself. Benjamin Tepping (1968) at the Census Bureau pioneered computer matching by laying out probabilistic principles for record linkage. One principle is weighting comparison fields on records being linked according to their discrimination in identifying record pairs characterizing single entities -persons, businesses. Several computer matching systems have been built at the Census Bureau on the same principles since then. They include the Statistical Research Division Record Linkage software (Winkler 1999) used for unduplication (FensterMaker Mule 2003) and "BigMatch" (Yancey 2004). Since these systems were designed, several decades ago in some cases, several developments have occurred.

1. Machine learning, as an instance of artificial intelligence, is omnipresent in classification and record linkage.
2. Whereas traditional methods for record linkage were not flexible to adapt and change on their own during record linkage applications, Bayesian methods allow for recursive processes mimicking the human brain.
3. A third dimension of change will not be explored in depth in this paper but its impact is overwhelming, it is the evolution of the computing technology. In the interest of concision we focus on the previous two items.

We seek to unify the proliferation of ideas to present a unified theory of record linkage. We focus on record-linkage methods classified as “unsupervised learning” under the machine learning taxonomy (see Shyam and Singh 2021). Conceptually these methods can also be described as “latent learning” since knowledge on the observations is accumulated by reference to unobserved latent structures. One unifying concept is “partitioning”, formally introduced in the record-linkage arena by Sadinle and Fienberg (2013). Partitioning ensures that the latent associations retrieved by the record linkage remain logical and do not conflict with the probabilistic assumptions implicit in a given application. The review centers on modern record-linkage methods and how efficiently they adapt to complex partition assignments.

In this context we review record linkage methods in two broad categories: static methods and dynamic methods. A linkage rule is static if its parameters do not change during its application and leads to a static partition assignment. A dynamic record-linkage rule allows for a parameter updating probabilistic mechanism as the underlying partition assignment is also being revised to characterize various record-linkage scenarios. Dynamic record linkage is usually Bayesian in nature, the Bayesian paradigm being a natural environment for sequentially revising parameters and estimates, and making sequential decisions. The review gives an historical context in section 2. Section 3 reviews the principles of multinomial derived record linkage and introduces partitioning in the context of static record linkage. The basic principles extend to section 4 where we review instances of dynamic record linkage. For the sake of completeness section 5 discusses supervised learning, which can lead to both competing and/or improved methodology relative to that presented earlier.

2. Linking and Weighing Information

2.1 Automation

Automated record linkage originates in studies involving linking of person records such as the Canada Vital Statistical Index (Marshal 1947). Birth and death certificates were matched within and between Canadian provinces. The identity of each person and information on births and deaths were represented on punchcards. These were painstaking efforts not supported by a mathematically well defined methodology.

Newcombe, Kennedy, Axford and James (1959) designed a system to automate the classication of vital records. The Oxford study (Acheson 1963) was another ambitious project for its times and also attempted to automate the linkage of medical records. Technically the work of Newcombe et al. (1959) stands out. These authors discuss weighing comparison fields using log-odds and making probabilistic computations. Exact computations in real time were out of reach and these authors propose empirical approximations to weigh the comparison fields in a pair of records with the purpose of linking them. Newcombe (1967) discusses automating record linkage as well as information retrieval but stops short of proposing specific probabilistic axioms. Tepping (1968) appears to be first to introduce a more rigorous probabilistic framework for record linkage informally, paving the way for the likelihood-based framework of Fellegi Sunter, by positing matching probabilities conditional on the observed data and proposing to estimate these probabilities through sampling. Various peripheral, but important, aspects of computer matching were introduced over the years. For instance, in the context of list of individuals, string comparisons involving name were -and still are- the object of attention. See Newcombe, Fair and Lalonde (1992).

2.2 Statistically Principled Record Linkage

Fellegi and Sunter (1969) introduce for the first time an explicit statistical theory for weighting record comparison and linking records. The original Fellegi-Sunter theory and methodology remains widely cited and is the starting point of several other development that we explore in the paper. We emphasize that a careful examination of the fine print of the F-S theory reveals its own limitations, which we suggest are not sufficiently emphasized in existing work. Most of this review is devoted to reviewing more recent research motivated in good part by a desire to overcome the limitations of F-S.

3. Static Record Linkage

3. 1 Notation

We assume independent multinomial trials each associated to a fixed-size cluster of records and characterized by a categorical outcome defined by a comparison pattern involving comparison fields of each record and by the unique matching status specific to the cluster.

Let $o_{m\boldsymbol{c}}$ be a count of clusters classified by the outcomes of the trials. $m \in \boldsymbol{M}$ indexes the matching status and $\boldsymbol{c} \in \boldsymbol{C}$ indexes the comparison patterns of each outcome. The matching status m resolves any ambiguity on the coincidence or non-coincidence of the entities characterizing the records in the cluster. The comparison pattern $\boldsymbol{c}$ associated to a cluster is a vector whose element c_i indexes the state of the comparison over the comparison field i, where $i = 1, \dots, F$. Let $\boldsymbol{o} = [o_{mc}]_{m \in \boldsymbol{M}\ \boldsymbol{c} \in \boldsymbol{C}}$ be the vector of all such counts.

Fellegi and Sunter (1969) investigate a simple case where each cluster includes 2 records, which either characterize the same ($m = 2$, a “match”) or distinct ($m = 1$, a “nonmatch” status) entities, so $\boldsymbol{M} = \{1, 2\}$. The comparison over each comparison field is either agree ($c_i = 2$) or disagree ($c_i = 1$), so $\boldsymbol{C} = \bigotimes^F \{1, 2\}$. Then $o_{1\boldsymbol{c}}$ is the count of record pairs exhibiting a match status of "nonmatch" and comparison pattern $\boldsymbol{c}$. Likewise $o_{2\boldsymbol{c}}$ is the count of record pairs exhibiting a match status of "match" and comparison pattern $\boldsymbol{c}$.

3.2 Likelihood

In general let $\boldsymbol{p} = [p_{mc}]_{m \in \boldsymbol{M}\ c \in \boldsymbol{C}}$ be a vector of in dividual probabilities where p_{mc} is the probability a random cluster of k records exhibit matching status $m \in \boldsymbol{M}$ and comparison pattern $\boldsymbol{c}$. The nominal constraints on $\boldsymbol{p}$ as a probability vector are

$$0 < p_{mc} < 1 \qquad m \in \boldsymbol{M} \quad \boldsymbol{c} \in \boldsymbol{C} \tag{1}$$

$$\sum_{m \in \boldsymbol{M}} \sum_{\boldsymbol{c} \in \boldsymbol{C}} p_{mc} = 1 \tag{2}$$

The general likelihood is

$$L(\boldsymbol{p}|\boldsymbol{o}) = \prod_{m \in \boldsymbol{M}} \prod_{\boldsymbol{c} \in \boldsymbol{C}} (p_{mc})^{o_{mc}} \tag{3}$$

The product in (3) assumes the outcomes are independent multinomial trials leading to the count vector $\boldsymbol{o}$. An added complication in unsupervised learning is $\boldsymbol{o}$ is not observable since m, the status of each cluster, is not. The observable data is the vector $\boldsymbol{o}_+ = [o_{+c}]_{\boldsymbol{c} \in \boldsymbol{C}}$ where

$$o_{+\boldsymbol{c}} = \sum_{m \in \boldsymbol{M}} o_{mc} \tag{4}$$

$o_{+\boldsymbol{c}}$ is the count of pairs of record exhibiting comparison pattern of $\boldsymbol{c}$. Let $\boldsymbol{o}_+$ be the vector of the aggregated counts in (4).

In the case of two files being matched (Fellegi-Sunter), $\boldsymbol{o}_+$ has 2^F components representing the counts for all possible outcomes of the pair comparisons over the F comparison fields. At the same time $\boldsymbol{p}$, the parameter of the likelihood in (3), has $2^F - 1$ independent components or degrees of freedom. In general when the degrees of freedom exceed the number of observations the model ($\boldsymbol{p}$) cannot be estimated.

3.3 Partitions

3.3.1 Partition for One-to-One Matching

We explore in more depth partions for one-to-one matching. This is generalized in several directions later. Sadinle (2013) introduces partitioning for the sets of clusters and their associated status to guarantee the validity of the likelihood in (3) for clusters of arbitrary size.

We first present a simple example of partitioning. Consider the situation where we have two files $f_A = [r_1, r_2\,]$, $f_B = [s_1, s_2\,]$. Each file has two records. We are interested in linking f_A and f_B and so each prospective cluster is a pair in the cartesian product $f_A \times f_B$ consisting of one record from f_A and one record from f_B. After linking the two files we want to infer the status for each of the following 4 pairs: (r_1, s_1), (r_1, s_2), (r_2, s_1), (r_2, s_2). A priori we are willing to consider all the pairs in $f_A \times f_B$ as potential matches. Suppose we are confident f_A and f_B are free of duplicates and we wants to exclude from consideration overlapping match assignment. That is we want to enforce one-to-one matching. Without loss of generality we let the partition representing the status of the pair (r_i, s_j) be $\{\{r_i, s_j\}\}$ if the status is a "match" and $\{\{r_i\}, \{s_j\}\}$ if the status is a "non-match". The assignment of a partition to each pair in the Cartesian product leads to a partition assignment. Enforcing one-to-one matching we have 7 permissible partition assignments. We give an instance of a full assignment

$$\mathcal{A}_1 = \{\{\{r_1, s_1\}\}, \{\{r_2, s_2\}\}, \{\{r_1\}, \{s_2\}\}, \{\{r_2\}, \{s_1\}\}\}$$

$\mathcal{A}_1$ characterizes a possible outcome for two independent pair-based multinomial trials each invoving F observed field comparisons and an unobserved matching status. $\mathcal{A}_1$ indicates both (r_1, s_1) and (r_2, s_2) have a status of "match" and along with the associated field comparisons they are the possible outcomes of two independent multinomial trials given the likelihood in (3) where $\boldsymbol{M} = \{1, 2\}$ and $\boldsymbol{C} = \bigotimes^F \{1, 2\}$ as defined in section 3.1. These two partitions drive the rest of the partition assignment $\mathcal{A}_1$. That is the partitions for(r_1, s_2) and (r_2, s_1) in $\mathcal{A}_1$ can be logically derived under the one-to-one rule for partition assignments. Accordingly we can aleviate the notation and define unambigously

$$\mathcal{A}_1 = \{\{r_1, s_1\}, \{r_2, s_2\}\}$$

Likewise we define the remaining partition assignments:

$$\mathcal{A}_2 = \{\{r_1, s_2\}, \{r_2, s_1\}\}$$
$$\mathcal{A}_3 = \{\{r_1, s_1\}, \{r_2\}, \{s_2\}\}$$
$$\mathcal{A}_4 = \{\{r_1, s_2\}, \{r_2\}, \{s_1\}\}$$
$$\mathcal{A}_5 = \{\{r_2, s_1\}, \{r_1\}, \{s_2\}\}$$
$$\mathcal{A}_6 = \{\{r_2, s_2\}, \{r_1\}, \{s_1\}\}$$
$$\mathcal{A}_7 = \{\{r_1\}, \{r_2\}, \{s_1\}, \{s_2\}\}$$

Each partition assignment assigns a distinct partition to each pair in the cartesian product and unambiguously define its matching status. In the frequentist perspective (Fisher 1973 p. 171) that prevailed when record-linkage was developed, a record-linkage methodology focused on identifying the most likely partition assignment. The increasingly prevalent Bayesian approach (Bayes approach, Lehman 1983 p. 237) focuses on minimizing loss functions (Lehman 1983, p. 5) which may lead to identifying multiple desirable partition assignments . We first review the frequentist approach along with fundamentals and review recent record-linkage methods devolving from the Bayesian approach in the later sections.

Note $\mathcal{A}_1$ and $\mathcal{A}_2$ assign a maximum number of match statuses, 2 in this case, to the pairs in $f_A \times f_B$, which defines "complete bipartite matching" between f_A and f_B. Although rare in practice, complete bipartite matching is easy to visualize and the number of partition assignments associated to it is easy to compute. Doing so gives us an idea of the total number of permissible partition assignments.

3.3.2 Partitions (cont'd)

Defining the logical configuarions of independent clusters generated by (3) through partition assignment can be expanded in several directions. For the purpose of the review it is useful at first to only increase the size of the clusters and maintaining one-to-one matching for any pair of files. Consider the triplet (r_i, s_j, t_k). In the context of independent multinomial trials generating such triplets based on the likelihood in (3) we have $\boldsymbol{M} = \{1, 2, 3, 4, 5\}$ and $\boldsymbol{C} = \bigotimes^F \{1, 2, 3, 4, 5\}$. That is status $m \in \boldsymbol{M}$ takes 5 values associated to the following partitions

$$\{\{r_i, s_j, t_k\}\}, \{\{r_i, s_j\}, \{t_k\}\}, \{\{r_i, t_k\}, \{s_j\}\}, \{\{s_j, t_k\}, \{r_i\}\}, \{\{r_i\}, \{s_j\}, \{t_k\}\}$$

That is (r_i, s_j, t_k) represents 3 records characterizing the same entity ($m = 5$), or 2 records characterizing the same entity, but the other record characterizes another ($m = 2,3,4$), or each record characterizing a different entity ($m = 1$). Each indicator c_i indexing 3-way comparison over field i is similarly defined.

We can extend the previous example to 3 files $f_A = [r_1, r_2]$, $f_B = [s_1, s_2]$, $f_C = [t_1, t_2]$. Under one-to-one-to-one matching any permissible partition assignment describing the matching status of the 8 triples in the cartesian product $f_A \times f_B \times f_C$ can be assumed to be a logical consequence of the outcome of two independent multinomial trials under the likelihood in (3), each comprising a set of three way comparisons over the comparison fields and a 3-way matching status as just described.

In this case we identify 82 distinct partition assignments and there are 4 matched-record bijections. For the sake of clarity we write down one such partition assignment, which defines a complete tripartite matching:

$$\mathcal{A}_0 = \{\{r_1, s_1, t_1\}, \{r_2, s_2, t_2\}\}$$

Again, the explicit partitions in $\mathcal{A}_0$, unambiguously ascribe partitions the the all the triples in $f_A \times f_B \times f_C$. It is tempting to attempt reconstructing the partition assignments for triplets through pair partitioning. For instance the triplet (r_i, s_j, t_k) can be decomposed into to the 3 pairs (r_i, s_j), (r_i, t_k), (s_j, t_k). The problem is in general outcomes from independent multinomial trials involving record triples and their statuses cannot be mapped to a set of independent pair-based outcomes of multinomials trials. We will further discuss this issue in subsequent sections.

In general, when linking u files of v records each and assuming one-to-one matching between the files in any subset of two files, there are $(v!)^{(u-1)}$ complete multipartite matchings. This expression should terrify anyone who comtemplates enumerating all the partition assignments when u and/or v are large. For instance, for $v = 1000$ and $u = 2$, a trivial situation, there are no personal computers commonly available on the market that can keep all the partition assignments in memory. In later sections we review methodologies designed to finesse this terrifying number.

3.4 Parameter Estimation and Matching Decisions

3.4.1 Maximizing The likelihood

In static record linkage and classical statistical estimation maximizing the likelihood is the key to making optimal decisions. When some of the classifying variables are not observed the EM algorithm is often used to maximize the likelihood conditional on the observations and to identify the parameters of the likelihood.

To exemplify the context and simplify the discussion we center on the case $\boldsymbol{M} = \{1, 2\}$ and $\boldsymbol{C} = \bigotimes^F \{1, 2\}$ throughout and point to higher dimensional generalizations. Winkler (1988) treats the unobserved matching status as a "latent" class and recast the model in (3) as a latent class model (Haberman, 1979) so that the tools already developed for these models can be adapted and exploited. As such Winkler adapt and fully expand an instance of the EM algorithm due to Haberman (1979). Winkler gives credits for the first implementation of the EM algorithm to Kelly (1986). Jaro (1988) also develop a prototype.

The EM algorithm (Dempster Laird Rubin, 1977) produces a sequence of simulated complete observations $\langle \boldsymbol{o}^{(n)} \rangle$, $n = 1, \ldots$ along with an associated sequence of direct maximum likelihood estimates $\langle \boldsymbol{p}^{(n)} \rangle$, $n = 1, \ldots$ based on the recursive formula were p is the MLE on the basis of the artificial ovservations. (4) continues to hold for all n by setting.

$$o_{\boldsymbol{mc}}^{(n+1)} = \left(\frac{p_{\boldsymbol{mc}}^{(n)}}{\sum_{m \in \boldsymbol{M}} p_{\boldsymbol{mc}}^{(n)}} \right) o_{+\boldsymbol{c}} \qquad (5)$$

$\boldsymbol{p}^{(n)}$ is obtained by maximizing on the basis of $\boldsymbol{o}^{(n)}$**.** In general this can be done by treating (3) as a poisson generalized linear model. The sequence is designed to converge to $\widehat{\boldsymbol{p}}$, that is the point where (1) is maximum while $\boldsymbol{o}^{(n)}$ remain subject to (4).

The sequence $\langle \boldsymbol{p}^{(n)} \rangle$ converges to $\widehat{\boldsymbol{p}}$, the MLE of $\boldsymbol{p}$, or a local, maximum, when is $\boldsymbol{p}$ is estimable. In general $\boldsymbol{p}$ is estimable for some but not all values of $\boldsymbol{o}_+$. Fellegi and Sunter (1969) derive conditions for estimability for the case $F = 3$ and an explicit expression for $\widehat{\boldsymbol{p}}$ when it is estimable.

3.4.2 Maximizing the likelihood over the Partition Assignments

Under the frequentist approach the overarching goal is to identify the most likely partition assignment $\mathcal{A}_*$ thereby making a decision on the statuses of all the pairs. $\mathcal{A}_*$ maximizes the likelihood in (3) for $\boldsymbol{p} = \widehat{\boldsymbol{p}}$. This can be done by brute force, successively evaluating the likelihood for each partition assignment. In that context it is useful to think of the partition assignment and the unobserved pair statuses as unknown parameters, rather than unobserved random variables. Note this still falls short of a full Bayesian framework.

When the files f_A, f_B grow large the cartesian product $f_A \times f_B$ grows quadratically. But, as we saw in 3.3.2, the number of partition assignment grows exponentially. The brute force approach -evaluating all the partition assignments- just does not scale up. It turns out that for pair-based independent multinomial trials there is a well-known shortcut.

3.4.3 Fellegi-Sunter Rule

The original Fellegi Sunter Rules is valid when classifying pairs of records derived from independent multinomial outcomes with likelihood (3) into matches and non-matches. It is an application of the Neyman-Pearson Lemma. The Fellegi-Sunter rule ranks the cells of the table, cross-classified by $\boldsymbol{c} \in \boldsymbol{C}$, according to the ratio

$$\Upsilon_{\boldsymbol{c}} = \log\left(\frac{\hat{p}_{2\boldsymbol{c}}}{\hat{p}_{1\boldsymbol{c}}}\right) \tag{6}$$

The original model proposed by Fellegi Sunter enforces conditional independence constraints on $\boldsymbol{p}$ leading to a factorization. In the log scale it means can be expressed as a sum of weights and so Fellegi and Sunter recuperate the proposals of Newcombe (1967) and Tepping (1968). Under the conditional independence constraints the log-ratio in (6) decomposes in a sum of "weights" in the log scale.

$$\Upsilon_{\boldsymbol{c}} = \Lambda_{\boldsymbol{c}} + B \tag{7}$$

$$\Lambda_{\boldsymbol{c}} = \sum_{i=1}^{F} W_i(c_i) \tag{8}$$

where

$$W_i(c_i) = \begin{cases} \log\left(\frac{1-\mu_i}{1-\upsilon_i}\right) & c_i = 1 \\ \log\left(\frac{\mu_i}{\upsilon_i}\right) & c_i = 2 \end{cases} \tag{9}$$

μ_i is the conditional probability of the pair of records exhibiting an agreement on field i if $m = 2$ (status is a true match) and υ_i is the conditional probability of the pair exhibiting an agreement on field i if $m = 2$ (status is a true nonmatch). B is the baseline:

$$B = \log\left(\frac{\rho}{1-\rho}\right) \tag{10}$$

ρ is the marginal probability of a true match ($m = 2$). The actual value of ρ is inconsequential when applying the Fellegi-Sunter decision rule since it does not change the ranking of the $\Lambda_{\boldsymbol{c}}$'s between the $\boldsymbol{c}$'s.

The decision maker determines a cutoff for $\Upsilon_{\boldsymbol{c}}$ in (6) associated to a partition assignment assigning a match status to all the pairs in a cell of $\boldsymbol{o}_+$ where $\Upsilon_{\boldsymbol{c}}$ is greater than the cutoff and a non-match status to all the other pairs.

When valid, the Fellegi-Sunter rule is optimal in the sense it guarantees that the partition assignment with value of Υ_c equal to or immediately above the cutoff has the lowest expected recall error among all the partition assignments whose expected precision is equal or smaller. Selecting a partition assignment according to Felleg-Sunter guarantees the most complete recall of matches among all the partition assignment yielding a smaller or equal expected precision -proportion of falsely declared matches.

Note the actual maximum of the likelihood in (3), and the theoretically optimal decision is defined by setting the cutoff at $W_i(2) = 1$. Then the probability of a match in the cell just above the cutoff is at or just above the estimated marginal probability $\hat{\rho}$. The associated partition could be termed the "absolute optimal partition assignment". However the associated decision may entail a level of risk -either precision or recall errors- that is too high or too low for the decision maker in a specific situation and a different Fellegi-Sunter cutoff can be chosen.

3.4.4 Limitations of Fellegi-Sunter

The F-S methodology suggests an approach to make decisions baised on pairwise comparisons between records associated to independent multinomial pair-based outcomes. In general Fellegi-Sunter does not straightforwardly extend to making decisions based on multi-way comparisons involving clusters of 3 or more records, which may surface when unduplicating a file or linking multiple files.

For example consider linking 3 files assuming one-to-one matching between the files in any subset of two files. The triple of records (r_i, s_j, t_k) spawns 3 pairs of records, (r_i, s_j), (r_i, t_k), (s_j, t_k), whose matching status we can attempt to predict by applying the FS and setting a cutoff for Υ_c. A cutoff predicting (r_i, s_j), (r_i, t_k) are matches, but not (s_j, t_k), is logically untenable and reflect fact that in general triple-based multinomial trials cannot be constructed by piecing together independent pair-based multinomial trials. In general the dependencies between the pair-based outcomes within clusters of more than 2 records means applying the Fellegi-Sunter rule on the pairs defined by the clusters to infer their multi-way match status can be biased and lead to logical impossibilities and is not supported by the F-S theory.

3.4.5 Enforcing One-to-One Matching

We have restricted the partitition assignment to one-to-one matching. The Fellegi-Sunter continues to hold if matching is not one-to-one matching and nominally involves the entire cartesian product. If desired, one-to-one matching in the absence of defined partition assignments can be enforced using a linear sum assignment routine (Koopmans and Beckman 1957), preserving polynomial-order scalability.

3.4.6 Extending Pairs to Multiples

Sadinle and Fienberg (2013) propose a way to extend static record-linkage from pair-based to cluster-based. The multi-way status of clusters of records is adjucated based on the likelihood in (3). For the sake of the discussion we focus on the 3-file situation described above and giving rise to the set of 5 partitions in described in section 3.2.2. For the likelihood in (3) we have $\boldsymbol{M} = \{1, 2, 3, 4, 5\}$ and $\boldsymbol{C} = \bigotimes^F \{1, 2, 3, 4, 5\}$.

The latent class model is now extended to 5 classes, one for each partition. If we are going to follow the conditional independence the number the size of the underlying classification grows exponentially in the size of the partition while the number parameters grows linearly. The authors experiment with the EM algorighm to identify the 5 clusters associated to the 5 matching status and compute a local maximum. Identifying 5 classes may prove to be too tall an order for the EM algorithm in some situations (Galindo Garre Vermunt 2006). The authors have introduced elements of supervised learning to achieve this goal which is tantamount to introducing prior information and the method is no longer strictly likelihood based. The authors also suggest generalizing the Fellegi Sunter rule to avoid explicitly evaluating every partition assignment. They generalize the concept of optimality in the Fellegi-Sunter sense to decisions on

clusters. But the decisions associated to this approach may leave out “undeclared clusters”. The higher the dimensionality of the cluster the more prevalent this can become.

3.5 Shortcomings of Static Modeling

The first shortcoming is we have remained in the static latent class modeling environment of unsupervised learning. In this environment it is all or nothing. As reported by Winkler (1994) the EM algorithm may fail to find a legitimate estimate, that is the sequence $p_{mc}^{(n)}$ converges on the boundary in terms of the probabilities. The offshoot is the probability of a true match is zero in half or more of the cells which goes against the premises of the model. At the same time this implies

$$\log\big(W_i(c_i)\big) \to -\infty$$

as attempts are made to maximize the likelihood in (3). Galindo Gare and Vermunt (2006) discuss this pathological behavior in the wider context of latent class models as instances of convergence to "boundary values". That is one or more components of $\boldsymbol{p}$ converge to 0 or 1, outside the parameter space. As the number of partitions and associated latent classes increases, so does the number of parameters to be estimated and the possibility of boundary values. It is fair to say Sadinle and Fienberg (2013) hit the limits of the EM algorithm. To finesse those limits they introduce element of supervised learning which we discuss in later sections. More generally egularization or penalized likelihood maximization has gained in popularity among practicians facing issues of non-estimability. Thibshirani (1996) opened a realm of new possibilities when he introduced penalized likelihood maximization via the LASSO. Chen et al. (2018) discuss instances of this approach applied to latent class models and record linkage. They use LASSO to tutor the EM algorithm and estimate a classification of cognitive diagnosis. More research is needed in the general context of record linkage.

A second limitation of the static approach is it provides only snapshots of the process, namely a snapshot of the process if $\boldsymbol{p} = \widehat{\boldsymbol{p}}$. While $\widehat{\boldsymbol{p}}$ is the most likely value, it falls short of describing a set of possibly useful partition assignments.

A third limitation is the difficulty to estimate the record linkage error. The efforts we described in that direction have failed to provide a satisfactory solution except in specific cases. A note bellow discuss two solutions addressing this problem.

Faced with these shortcomings new methodologies have been proposed. Representative proposals are reviewed in the following sections.

4. Dynamic Record Linkage

In general for record linkage operations involvoving complex partitions the likelihood in (3) is customized for a given application and becomes difficult to tract algebraically. Deriving the full set of partition assignments a priori may not be feasible computationally.

An alternative is to establish tractable rules for the partition assignment and to generate a stochastic sequence $\langle \boldsymbol{p}^{(n)} \rangle$ and an associated sequence $\langle \boldsymbol{o}^{(n)} \rangle$ bound by those rules. That is the sequence will follow a course governed by the target likelihood but, rather than converging to $\widehat{\boldsymbol{p}}$, a static estimate, the sequence extends as long as necessary to become an approximation of a "posterior distribution" through its frequency history. Formally the posterior distribution is obtained by conjugating the “prior distribution” (DeGroot p. 139) in accordance to the likelihood in (3) (DeGroot p. 159). This is often not possible to resolve mathematically and constructing a sequence $\langle \boldsymbol{p}^{(n)} \rangle$ provides a valid numerical approximation. The advantage of an approximate posterior distribution is we have a full probabilistic characterization of the linkage process across the possible values of $\boldsymbol{p}$, rather than a single estimate $\widehat{\boldsymbol{p}}$, which we hope is ideal but tells us only part of the story. We call this approach “dynamic record linkage” in the sense the underlying sequences $\langle \boldsymbol{p}^{(n)} \rangle$ and $\langle \boldsymbol{o}^{(n)} \rangle$ are in flux as long as necessary, as opposed to reaching a stationary point as in likelihood maximization.

4.1 Dynamic Unduplication

Unduplication is the process of linking a file to itself. Attempting to unduplicate a file often reveals clusters of records of arbitray sizes potentially characterizing the same entity. An instance of a dynamic process Sadinle (2014) proposes an approach borrowing statistically from the Fellegi-Sunter theory adapted to a dynamic environment in the context of the cartesian product of one file with itself. The author uses the log ratio $\Lambda_{\boldsymbol{c}}$ in (8), which is a function of $\boldsymbol{p}$, to construct dynamic sets of entities whose association to the records of a file are sequentially updated according to a new or existing partition. Given that R records are still subject to unduplication the pivotal proposition connecting the predicted duplicate status of a record to the most recent value of $\boldsymbol{p}$ in the sequence $\langle \boldsymbol{p}^{(n)} \rangle$ is

$$P(\omega \sim E) \propto \begin{cases} \prod_{\boldsymbol{c} \in \boldsymbol{\Gamma}(\omega,\ E)} e^{\Lambda_{\boldsymbol{c}}} & E \text{ is labeled} \\ \left(R - \sum_{E \text{ is labeled}} I_E \right)^{-1} & E \text{ is not labeled} \end{cases}$$

We use the notation $\omega \sim E$ to mean record ω is a characterization of entity E. The probability that a new record ω is a characterization of E is proportional to the products of the odd ratios for all the comparison patterns to a record already identified as a characterization, $\boldsymbol{c} \in \boldsymbol{\Gamma}(\omega,\ E)$. This means $\Lambda_{\boldsymbol{c}}$ is in constant flux as its value updated after each decision involving a new record. This implies many more comparisons than in the static paradigm and also much more information on the plausibility of the matches. A new value for $\boldsymbol{p}$ is generated from the updated posterior distribution to extend the sequence $\langle \boldsymbol{p}^{(n)} \rangle$ after a duplicate status is assigned to a record. The full cycle reassigning all the records is repeated until the posterior distribution of $\boldsymbol{p}$ stabilizes.

4.2 Tancredi and Liseo

In the context of capture-recapture -matching two independent listing of the same population reference here- Tancredi and Bruneo (2011) propose a largely non-parametric approach rooted in combinatorics. In that case the simple partition of pairs is retained but the authors allow for complex comparison patterns making the FS classification rule unapplicable.

Tancredi and Liseo (2011) introduce a non-parametric Bayesian set up derived from the "hit or miss" model of Copas and Hilton (1990) to predict number of links between 2 snapshots of a single population taken a different times (capture-recapture sampling). The hit or miss model is intrinsically overparameterized and the authors make simplifying assumption involving complex combinatorics. This approach greatly broadens the restrictive assumptions of Fellegi-Sunter. Tancredi and Brunero no longer models only the agreement-disagreement patterns but the configurations of the records themselves to probabilistically accounts for multiple latent links simultaneously. The inference is supported by generating a dynamic Markov Chain Monte-Carlo sampler which implicitly defines a predictive distribution on the links themselves. The limitations come from the complexities underlying the MCMC which involve extensive numeric combinatorial computations that are sure to explode when generalizing the method to many populations.

4.3 Smered and d-blink

Several authors build on the method of Tancredi and Brunero, borrowing from small area estimation (Gosh 2020) in building a hierarchical model tuned for improved performance. Steorts, Hall and Fienberg (2016), Steorts (2015), and Marchant, Steorts, Kaplan, Rubinstein and Elazar (2019), describe an entire class of MCMC sampling implementations whose common thread is they eschew pairwise record com-

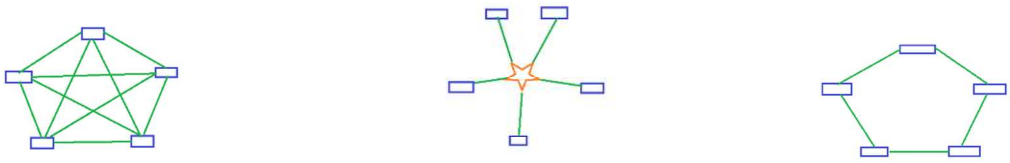
parisons to probabilistically characterize links between records. Instead unobservable "latent entities" broker the links between disjoint records thereby identifying clusters of records. Inclusion of an observed record to a specific cluster is predicated by the degree of similarity between the attributes of the latent entity and the attributes of the observed record.

The comparison between record and entity attributes allows for natural distortions. We focus on "d-blink" the more mature implementation of Marchant et al. (2021). An immediate advantage of this approach is relative computational frugality. In particular in situations where cluster sizes can grow large and clusters. In such situations the approach of Sandinle potentially involves many pairwise comparisons to ascertain the matching status of a cluster. d-blink prescribes making comparisons only to the latent entities, who serve as representatives for their respective clusters. Taken to an extreme, the number of pairwise comparisons grows quadratically as cluster size grows, while the number of comparisons transacted by a single latent entity grows linearly. However, if only small clusters (2 or 3 records) are predominant in a specific situation, then the computational advantages of d-blink will be less impor-
tant.

For purpose of comparison Figure 1 illustrates 3 types of matching structures for 5 records. The first structure (Sadinle Fienberg 2013) entails carrying each pair comparisons. The second structure (d-blink, Marchant et al. 2021) is based on “brokering” -the broker is represented by a star- and involves substantially fewer comparisons as cluster size grows. The third structure reflects the approach of Thibaudeau (1992) and was motivated chiefly by the computational frugality of the era. Only two comparisons are needed to ascertain inclusion to the structure. This entails discarding the information on all the other comparisons. It is justified by setting a high Fellegi-Sunter cutoff for inclusion and robust clerical support to review comparisons marginally under the cutoff.

Figure 1. 3 Cluster structures for Matching:

1 All pairwise comparisons (Sadinle Fienberg 2013)
2 Brokering (d-blink 2021)
3 Circular (Thibaudeau 1992)



The central assumption of d-blink is statistical independence between links involving a common latent entity conditional on the the attributes of the entity itself. So, at any point during the sequence, conditional on the information retrieved from a common latent entity, there is nothing to learn about the association of a record to a cluster by examining the other records in the same cluster. This assumption stems from the authors' application of the hit or miss model (Copas Hilton 1990).

4.4 General Partitions in Dynamic Record Linkage

4.4.1 Multifile Partitions

Aleshin-Guendel and Sadinle (2023) extend the partitioning approach of Sadinle and Fienberg (2013) to “multifile partitions”. Although Sadinle and Fienberg (2013) discuss linking multiple files their methodology is confined to one-to-one matching between any two files, precluding record duplication within any of the single files. Aleshin-Guendel and Sadinle design a MCMC sampler powered by a generalized multinomial-Dirichlet likelihood/posterior distribution, that is a generalization of (3). The methodology partitions each file to support the identification of duplicate records within the file leading to a d-blink

styled matching structure with a representative or “broker” for each cluster identified in a single file. Then the brokers are linked and matched one-to-one between filesto produce a hierarchichical matching structure and a hierarchical partition assignment. The MCMC continuously upgrades the partition assignment, which needs not be of fixed size. Again not every possible partition assignment can be visited but under regularity conditions the MCMC will revisit the most plausible partition assignments and avoid the unproductive ones with a high probability.

The complexity of the partition assignments means trying to identify the “most likely assignment”, as is easily done for the Fellegi-Sunter model, becomes impractical and in good part meaningless. Accordingly these authors take advantage of the Bayesian approach and propose loss functions tailored for this situation. The loss functions allow for discounting partition assignments whose linkage decisions are ambiguous.

4.4.2 Random Partitions

Given the basic architecture of d-blink Marchant et al. (2023) introduce a random partition infrastructure. It enable d-blink to upgrade partition assignments based on “learned experience” on a rolling basis. A probability distribution over a set of partition assignment is continuously upgraded on the basis of better knowledge and new partition assignments are drawn, minimizing the probability of the process languishing in the exploration of unproductive partition assignments.

5. Supervised and Semi-Supervised Learning

5.1 Training Data

One way to finesse the difficulties associated to the parameter estimation and the classification of a large number of partitions when relying on a non-identifiable model is to use "training data" (labelled data). This approach, when possible, infuses a new life into static record linkage. In the situations presented in the paper this means $\boldsymbol{o} = [o_{mc}]_{m \in \boldsymbol{M}^* \; c \in \boldsymbol{C}}$ is observed for a subset of clusters $\boldsymbol{M} \subset \boldsymbol{M}^*$ in (3). In the Fellegi-Sunter set-up this means we are left with a logistic regression on a subset of pairs of known status. Several other regression techniques are also possible, in particular tree-based regressions, such as random forest. Pita et al. (2017) reviews various supervised learning techniques. These authors report that logistic regression, which is the natural extension of Fellegi-Sunter in supervised learning, appears to perform well.

5.2 Supervised vs Unsupervised Learning

The supervised learning arena allows for many new approaches. Central to all of them is the availability and quality of training data. Supervised training application are static by nature, although new training data can be sampled and the learning process repeated through the linkage. Conceptually the parameters of the model are estimated from the training data and do not change during execution.

Supervised learning enable much creativity and new methodologies. However requirements that a representative training data may be unrealistic. The data available may only partly reflect the data to be linked. Christen (2007, 2008) proposes an approach to circumvent this issue.

5.3 Two-Step Approach

Christen (2007, 2008) proposes a nuanced supervised learning approach using "seed data" from the data sets subject to matching to recreate the paradigm of supervised learning in the absence of bonafide training data. The seed data are created by selecting a subset of the data and computing several distance measures relative to the worse and best case scenarios to identify a synthetic set of “truth data” in lieu of observed training data.

5.4 Semi-Supervised Learning Extensions of Fellegi-Sunter

Larsen and Rubin (2001) enrich the Fellegi-Sunter rule by introducing one-step of semi-supervised learning in the EM algorithm

$$\Omega_{mc}^{(n+1)} = o_{mc}^{(n+1)} + q_{mc}$$

Then $\Omega_{mc}^{(n+1)}$ is substituted for o_{mc} in the maximization of the likelihood in (3) (M-step) to obtain $p_{mc}^{(n+1)}$ for the next E-step. o_{+c} is the counts of the match patterns for which m is unknown while q_{mc} is the count of record pairs of with matching status m exhibiting matching pattern $\boldsymbol{c}$ that were verified through the clerical operations. This approach stabilizes the EM algorithm by allowing the estimation of less restrictive model -relaxing the conditional independence assumptions- and in of itself introducing observed information on the status. The drawback is the cost of the clerical review.

6. Discussion

6.1 Dynamic Record Linkage

Explicit partitioning as introduced by Sandinle and Fienberg has enabled the development of a rigorous theory and principled applications of complex record linkage that were lacking since the advent of Fellegi-Sunter. The implicit challenge, probabilistically characterizing all the possible partition assignments, remains monumental. At this time the multi-file approach of Aleshin-Guendel (2023) and the latest iteration of d-blink (Marchant et al. 2023) offer the latest advances to do this. Aleshin-Guendel take advantage of file boundaries to reduce computations by casting the problem as one-to-one matching problem between any two files, gaining substantial efficiency relative to merging the files together and attempting to unduplicating them from scratch. The gains in efficiency increase as the number of files grows.

d-blink relies on a unit level model reminiscent of small-area estimation (Gosh 2020) making it applicable very generally. Nominally the range of the partition assignments is overly broad since there are no boundary between files. But by using blocking and other prescriptions on partitioning the problem is brought down to a more manageable size. Perhaps the most hopeful features in dynamic record-linkage at this time is introduced in the latest version of d-blink (Marchant et al. 2023). That is random partitioning. The unit level model of d-blink allow for defining separate process shaping the partition assignment (Pitman 1995), which also account for the success history of the previous assignments earlier in the process. This approach is promising and should be further exploited.

Despite these very encouraging advances, genuinely large applications (linking databases of millions or billions of records) remain beyond the scope of dynamic Bayesian record-linkage at the time of this writing. Historically -i.e. per the history of Gibbs sampling- these methods have measured up to the challenges sooner or later. We are hopeful the ingenuity of the professionals active in this area combined with technological advances will lead to overcoming the obstacles to provide comprehensive probabilistic characterizations of large record linkage processes. There remain tools available in the Bayesian arsenal,

such as large-sample approximations (Tierney Kass Kadane 1989, Thibaudeau 1988) that can speed up the process.

6.2 Static Record Linkage

Bayesian record linkage gives a rich probabilistic characterization of record-linkage processes. Very importantly the Bayesian approach naturally integrates measurement of risks and errors, which are problemetic to extract from non-Bayesian record linkage, static in nature. However there remains a demand for snapshots and simple estimates as offered by static record linkage. In particular the high scability of Fellegi-Sunter is attractive, even as one must be aware Fellegi-Sunter nominally has important limitations.

Sadinle and Fienberg (2013) fully generalize the theory of Fellegi-Sunter contrived to one-to-one matching of two files to multi-file matching based on their partition system. The difficulty is the likelihood in (3) becomes analytically difficult to handle as the dimension of the clusters grows and the number of parameters explodes. Generalization of the Fellegi-Sunter rule to summarize the risks associated to the exponentially growing number of partition assignments also become onerous and difficult to formulate as cluster dimension increases. Solving these problems will be key to apply static record linkage methods on large problems in the future.

A static alternative to the approach of Sadinle and Fienberg is to assign complex partitions by sequentially applying Fellegi-Sunter on pair-based outcomes and making conditional decisions on matching status at each stage. This approach is displayed in the sidebar and is based on marginal tables and subtables of the full classification. This approach becomes more complex when several files are being linked but it gives hope of retaining the high scalability of Fellegi-Sunter in general multi-file linking.

6.3 Conclusion

The importance of Bayesian record linkage supported by sophisticated partitioning schemes will continue to grow. We presented two examples of supervised learning. There are many more applicable to record linkage and classification in general. For an empirical evaluation see Soru and Ngonga Ngomo (2014). The Bayesian approach is the unifier we were looking for. It allows for integrating most supervised-learning precesses into unsupervised-learning processes of the type we focused on in the paper by eliciting appropriate prior distributions. Therefore the two learning approaches do not need to be exclusive of each other. This supports investing in more research to generalize and optimize the methods in the paper. The payoff is likely to be substantial and not to be limited to only record linkage but to machine learning in general.

Using Fellegi-Sunter to Link 3 Files

Linking 3 files is decomposed into 4 instances of linking 2 files, each instances using Fellegi-Sunter. Step 1 assign initial partitions. Steps 2, 3 and 4 assign the final partitions.

We link the files A, B and C with records r_i, s_j and t_k.

1. Link A to B using F-S.

$$f_A \times f_B \rightarrow \left\{\{r_i\},\{s_j\}\right\}, \left\{\{r_i, s_j\}\right\}$$

2. Join r_i and s_j in each initial partition $\left\{\{r_i, s_j\}\right\}$ to construct $A \otimes B$, a file of double records $[r_i, s_j]$.

Similarly construct a new file $C \otimes C$ by doubling up the t_k's: $[t_k, t_k]$.

Link $A \otimes B$ to $C \otimes C$ using F-S.

$$f_{A\otimes B} \times f_{C\otimes C} \rightarrow \begin{cases} \left\{\{r_i, s_j\},\{t_k\}\right\} \\ \left\{\{r_i, s_j, t_k\}\right\} \end{cases}$$

3. Link the r_i's in the initial partitions $\left\{\{r_i\},\{s_j\}\right\}$ to C using F-S.

$$f'_A \times f_C \rightarrow \begin{cases} \left\{\{r_i\},\{s_j\},\{t_k\}\right\} \\ \left\{\{r_i, t_k\},\{s_j\}\right\} \end{cases}$$

4. Link the s_j's in the initial partitions $\left\{\{r_i\},\{s_j\}\right\}$ to C using F-S.

$$f'_B \times f_C \rightarrow \begin{cases} \left\{\{r_i\},\{s_j\},\{t_k\}\right\} \\ \left\{\{s_j, t_k\},\{r_i\}\right\} \end{cases}$$

If both $\left\{\{r_i, t_k\},\{s_j\}\right\}$ and $\left\{\{s_j, t_k\},\{r_i\}\right\}$ are assigned to the same i, j, k there is a conflict and the tie is broken based on the strengths of the F-S scores of the last record comparisons.

Notes:

Error Estimation under Fellegi-Sunter

Resnick and Asher (2019) and Binette and Steorts (2022) derive expressions in terms of p_{mc} in (2) allowing to quantify the error conditional on the decisions $\boldsymbol{C}(\mu)$ (declared matches) and $\boldsymbol{C}(\upsilon)$ (declared nonmatches). Let α and β be the type I and type II errors. We posit:

$$\alpha = \left(\frac{1}{o_{++}}\right) \sum_{\boldsymbol{c} \in \boldsymbol{C}(\mu)} \left(1 - \frac{\hat{p}_{2\boldsymbol{c}}}{\sum_{m \in \boldsymbol{M}} \hat{p}_{m\boldsymbol{c}}}\right) o_{+\boldsymbol{c}} \tag{11}$$

$$\beta = \left(\frac{1}{o_{++}}\right) \sum_{\boldsymbol{c} \in \boldsymbol{C}(\upsilon)} \left(\frac{\hat{p}_{2\boldsymbol{c}}}{\sum_{m \in \boldsymbol{M}} \hat{p}_{m\boldsymbol{c}}}\right) o_{+\boldsymbol{c}} \tag{12}$$

The problem is the formula in (11) and (12) is they implicitely embed an estimate of the proportion of matches, ρ, known to be biased -inflated. Resnick and Asher propose substituting an estimate computed from the classification based on Fellegi Sunter. This implies replacing $\hat{p}_{2\boldsymbol{c}}$ with

$$p^*_{2\boldsymbol{c}} = \left(\frac{\sum_{\boldsymbol{c} \in \boldsymbol{C}(\mu)} o_{+\boldsymbol{c}}}{\sum_{\boldsymbol{c} \in \boldsymbol{C}} o_{+\boldsymbol{c}} \sum_{\boldsymbol{c} \in \boldsymbol{C}} \hat{p}_{2\boldsymbol{c}}}\right) \hat{p}_{2\boldsymbol{c}} \tag{13}$$

In (11) and (12). The authors report satisfactory results. In effect (13) aligns the estimate of the proportion of matches on the decision of the analyst when setting the Fellegi-Sunter cutoff.

Belin and Rubin (1995) propose fitting a separate model, a gaussian mixture, to estimate the error rates of the record linkage.

Disclaimer

Any views expressed on statistical methodology, technical or operational issues are those of the author and not those of the U.S. Census Bureau.

References

Aleshin-Guendel, S., Sadinle, M. (2023). Multifile Partitioning for Record Linkage and Duplicate Detection. *Journal of the American Statistical Association*, 118(543), 1786-1795. https://doi.org/10.1080/01621459.2021.2013242

Acheson D. (1963). Oxford Record Linkage Study. *British Medical Journal*, 2(5252), 310.

Belin T., Rubin D. (1995). A Method for Calibrating False Match Rates in Record Linkage. *Journal of the American Statistical Association*, 90(430), 694-707.

Chen Y., Li X., Liu J., Ying Z. (2017). Regularized Latent Class Analysis with Application in Cognitive Diagnosis. Psychometrika, 82, 660-682. https://doi.org/10.1007/s11336-016-9545-6

Christen, P. (2008). Automatic Record Linkage Using Seeded Nearest Neighbour and Support Vector Machine Classification. Proceedings of the ACM 14th International Conference on Knowledge Discovery and Data Mining Las Vegas Nevada. Association for Computing Machinery. 151-159

Christen, P. (2007). A Two-Step Classification Approach to Unsupervised Record Linkage. Proceedings

of the 6th Australasian Data Mining Conference Gold Coast, Australia. Australian Computer Society. 111-120.

DeGroot, M. (2004). Optimal Statistical Decisions. Wiley Classics Library Edition.

Dempster, A., Laird, N., Rubin, D. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. JRSSB, 39(1), 1-38. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

Fellegi, I., Sunter, A. (1969). A Theory for Record Linkage. JASA, 64(328), 1183-1210.

Fenstermaker, D., Mule, T. (2003). Further Study of Person Duplication in Census 2000. A.C.E. Revision II Washington DC (USA). U.S. Census Bureau. Memorandum Series PP-30. https://www2.census.gov/programs-surveys/decennial/2000/program-management/5-review/ace/pp-15r.pdf

Fisher R. (1973). Statistical Methods and Scientific Inference Third Edition. Hafner Press.

Galindo Garre, F., Vermunt, G. (2006). Avoiding Boundary Estimates in Latent Class Analysis by Bayesian Posterior Mode Estimation. *Behaviormetrika*, 33(1), 43-59

Gosh, M. (2020). Small Area Estimation: Its Evolution in Five Decades. *Statistics in Transition*, 21(4),1-22. http://dx.doi.org/10.21307/stattrans-2020-022

Haberman, S. (1979). Analysis of Qualitative Data Vol. 2. New York: Academic Press.

Jaro, M. (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.

Koopmans, T., Beckmann, M. (1957). Assignment Problems and the Location of Economic Activities. *Econometrica*, 25(1), 53-65. https://www.jstor.org/stable/1907742

Larsen, M., Rubin, D. (2001). Iterative Automated Record Linkage Using Mixture Models. *Journal of the American Statistical Association*. 96(453), 32-41. https://doi.org/10.1198/016214501750332956

Lehman, E. (1983). Theory of Point Estimation. Wiley.

Marchant, N., Rubinstein, J., Steorts, R. (2023). Bayesian Graphical Entity Resolution Using Exchangeable Partition Priors. *Journal of Survey Statistics and Methodology*.11(3), 569-96. https://doi.org/10.1093/jssam/smac030

Marchant, N., Steorts, R., Kaplan, A., Rubinstein, B., Elazar D. (2021) d-blink: Bayesian End-to-End Distributed Entity Resolution. *Journal of Computational and Graphical Statistics*. 30(2), 406-421. https://doi.org/10.1080/10618600.2020.1825451

Marshal, J. (1947). Canada's National Vital Statistics Index. *Population Studies*. 1(2), 204-211. https://doi.org/10.2307/2172289

Newcombe, H., Fair, M., Lalonde, P. (1992). The Use of Names for Linking Personal Records. *Journal of the American Statistical Association*, 87(420), 1193-1204.

Newcombe, H. (1967). Record Linking: The Design of Efficient Systems for Linking Records into Individual and Family Histories. *Amer. J. Hum. Genet*., 19(3 Pt. 1), 335-359.

Newcombe, H., Kennedy, J., Axford, S., James, A. (1959). Automatic Linkage of Vital Records. *Science*. 130(3381), 954-959. https://doi.org/10.1126/science.130.3381.954

Pita, R., Mendonca, E., Reis, S., Barreto, M., Denaxas, S. (2017). A Machine Learning Trainable Model to Assess the Accuracy of Probabilistic Record Linkage. Bellatreche L. Chakravarty S. eds. Administrative Data Research Network Annual Conference. Springer International.

Pitman, J. (1995). Exchangeable and Partially Exchangeable Random Partitions. *Probability Theory and Related Fields*, 102, 145-158.

,

Resnick, D., Asher, J. (2019). Measurement of Type I and Type II Record Linkage Error. Proceedings of the American Statistical Association Government Statistics Section Joint Statistical Meetings Denver CO USA. https://www.researchgate.net/publication/336796421_Measurement_of_Type_I_and_Type_II_Record_Linkage_Error

Sadinle, M. (2014). Detecting Duplicates in a Homicide Registry Using A Bayesian Partitioning Approach. Annals of Applied Statistics, 8(4), 2404-2434. DOI: 10.1214/14-AOAS779

Sadinle, M., Fienberg, S. (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

Shyam, R., Singh, R. (2021). A Taxonomy of Machine Learning Techniques. *Jour. Adv. Robotics*, 8(3),18 25.

Soru, T., Ngonga Ngomo, A. (2014). A Comparison of Supervised Learning Classifiers for Link Discovery. Proceedings of the 10th International Confefence on Semantic Systems Leipzig Germany, 40-44.

Steorts, R., Hall, R., Fienberg (2016). A Bayesian Approach to Graphical Record Linkage and Deduplication. Journal of the American Statistical Association,111(516), 1660-1672. https://doi.org/10.1080/01621459.2015.1105807

Steorts, R. (2015). Entity Resolution with Empirically Motivated Priors. *Bayesian Analysis*, 10(4), 849-875. DOI: 10.1214/15-BA965SI

Tancredi, A., Liseo, B. (2011). A Hierarchical Bayesian Approach to Record Linkage and Population Size problems. *Annals of Applied Statistics*, 5(2B), 1553-1585. DOI: 10.1214/10-AOAS447

Tepping, B. (1968). A Model For Optimum Linkage of Records. *Journal of the American Statistical Association*, 63(324),1321-1332.

Thibaudeau, Y. (1992). Identifying Discriminatory Models in Record Linkage. Proceedings of the American Statistical Association Section on Survey Research Methods Joint Statistical Meetings Boston MA USA, 835-840. http://www.asasrms.org/Proceedings/papers/1992_141.pdf

Thibaudeau, Y. (1988). Approximating the Moments of a Multimodal Posterior Distribution with The Method of Laplace PhD Dissertation. Carnegie Mellon Pittsburgh PA USA.

Thibshirani R. (1996). Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society B, 58(1), 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Tierney, L., Kass, R., Kadane, J. (1989). Fully Exponential Laplace Approximations to Expectations and Variances of Nonpositive Functions. *Journal of the American Statistical Association* 84(407), 710-716.

Winkler, W. (1988). Using the EM Algorithm for Weight Computation in the Fellegi- Sunter Model of Record Linkage. Proceedings of the American Statistical Association Section on Survey

Research Methods Joint Statistical Meetings New Orleans LA USA, 667-671.

Winkler, W. (1994). Advanced Methods for Record Linkage. 1994; Research Report 94A. U.S. Census Bureau. https://www.census.gov/content/dam/Census/library/working-papers/1994/adrm/rr94-5.pdf

Yancey, W. (1994). BigMatch: A Program for Large-Scale Record Linkage. In: Proceeddings of the American Statistical Association Section on Survey Research Methods Toronto Canada. 4652-4655.