Online Supplementary Information for “Using a Probabilistic Model to Assist Merging of Large-scale Administrative Records.”[*]

Ted Enamorado[†] Benjamin Fifield[‡] Kosuke Imai[§]

October 17, 2018

This supplementary appendix presents the theoretical and empirical details, which due to the space constraints are omitted from the main text of our paper.

S1 The Jaro-Winkler String Distance

As discussed in Section 2, we use the Jaro-Winkler string distance (Jaro, 1989; Winkler, 1990), which is a commonly used measure in the literature (e.g., Cohen, Ravikumar and Fienberg, 2003; Yancey, 2005).[1] The Jaro-Winkler string distance between strings s_1 and s_2, which ranges from 0 to 1, is defined as,

$$D(s_1, s_2) \;=\; 1 - \{J(s_1, s_2) + \ell \cdot w \cdot (1 - J(s_1, s_2))\}$$

where

$$J(s_1, s_2) \;=\; \begin{cases} 0 & \text{if } m = 0 \\ \frac{1}{3}\left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m - t/2}{m}\right) & \text{otherwise} \end{cases}$$

where $|s|$ represents that length of string s, m is the number of characters in common between the two strings, t is the number of transpositions between the common characters, $\ell \in [0, 4]$ is the number of consecutive characters in common at the beginning of the two strings, and $w \in [0, 0.25]$ is

[*] The proposed methodology is implemented through an open-source R package, fastLink: Fast Probabilistic Record Linkage, which is freely available for download at the Comprehensive R Archive Network (CRAN; `https://CRAN.R-project.org/package=fastLink`).

[†] Ph.D. Candidate, Department of Politics, Princeton University, Princeton NJ 08544. Email: `tede@princeton.edu`

[‡] Ph.D. Candidate, Department of Politics, Princeton University, Princeton NJ 08544. Email: `bfifield@princeton.edu`, URL: http://www.benfifield.com

[§] Professor of Government and of Statistics, Institute for Quantitative Social Science, Harvard University. 1737 Cambridge Street, MA 02138. Phone: 617–384–6778, Email: `imai@harvard.edu`, URL: `https://imai.fas.harvard.edu`

[1] Another reason for its popularity is the existence of efficient implementations of the Jaro-Winkler string distance in several programming languages including C, C++, and Java.

the weight given to ℓ. For example, if we consider two last names, $s_1 = \texttt{Smith}$ and $s_2 = \texttt{Martinez}$, we have that $m = 3$ (the letters: `m`, `i`, `t`), $|s_1| = 5$, and $|s_2| = 8$, and $t = 2$. If we set $\ell = 4$ and $w = 0.1$, as often done in practice (see e.g., Winkler, 1990; Cohen, Ravikumar and Fienberg, 2003), then the Jaro-Winkler distance for these two strings equals 0.55.

S2 The EM Algorithm

Following Winkler (1988), we apply the expectation and maximization (EM) algorithm, which is an iterative procedure, to estimate the model parameters (Dempster, Laird and Rubin, 1977). Under the modeling assumptions described in Section 2.2, the complete-data likelihood function is given by,

$$\mathcal{L}_{com}(\lambda, \boldsymbol{\pi} \mid \boldsymbol{\gamma}, \boldsymbol{\delta}) \;\propto\; \prod_{i=1}^{N_{\mathcal{A}}}\prod_{j=1}^{N_{\mathcal{B}}}\prod_{m=0}^{1}\left\{\lambda^m(1-\lambda)^{1-m}\prod_{k=1}^{K}\left(\prod_{\ell=0}^{L_k-1}\pi_{km\ell}^{\mathbf{1}\{\gamma_k(i,j)=\ell\}}\right)^{1-\delta_k(i,j)}\right\}^{\mathbf{1}\{M_{ij}=m\}}$$

Given this complete-data likelihood function, the E-step is given by equation (5) where the probability of being a true match is computed for each pair given the current values of model parameters. Using this match probability, the M-step can be implemented as follows,

$$\lambda \;=\; \frac{1}{N_{\mathcal{A}}N_{\mathcal{B}}}\sum_{i=1}^{N_{\mathcal{A}}}\sum_{j=1}^{N_{\mathcal{B}}}\xi_{ij} \tag{S1}$$

$$\pi_{km\ell} \;=\; \frac{\sum_{i=1}^{N_{\mathcal{A}}}\sum_{j=1}^{N_{\mathcal{B}}}\mathbf{1}\{\gamma_k(i,j)=l)\}(1-\delta_k(i,j))\xi_{ij}^m(1-\xi_{ij})^{1-m}}{\sum_{i=1}^{N_{\mathcal{A}}}\sum_{j=1}^{N_{\mathcal{B}}}(1-\delta_k(i,j))\xi_{ij}^m(1-\xi_{ij})^{1-m}} \tag{S2}$$

Then with a suitable set of starting values, we repeat the E-step and M-step until convergence.

When setting the starting values for the model parameters, we impose inequality constraints based on the following two ideas: (1) the set of matches is strictly smaller than the set on non-matches $\lambda \ll 1 - \lambda$, and (2) for binary comparisons, we have $\pi_{k10} \ll \pi_{k11}$ and $\pi_{k01} \ll \pi_{k00}$ for each k (Jaro, 1989; Winkler, 1993; Sadinle and Fienberg, 2013). The latter implies that agreement (disagreement) is more likely among matches (non-matches). In simulation and empirical studies, we find that this constraint dramatically improves the performance of the model by avoiding converging to a local maximum.

S3 Additional Results for Computationally Efficient Implementation

In this section, we provide the details of runtime comparison shown in Section 3.3.

S3.1 Runtime Comparison Breakdown

Section 3 presents our runtime comparisons against two other open-source implementations of the probabilistic record linkage model. Here, we examine the sources of the computational speedups we find in those comparisons. We break down the record linkage process into three stages. The

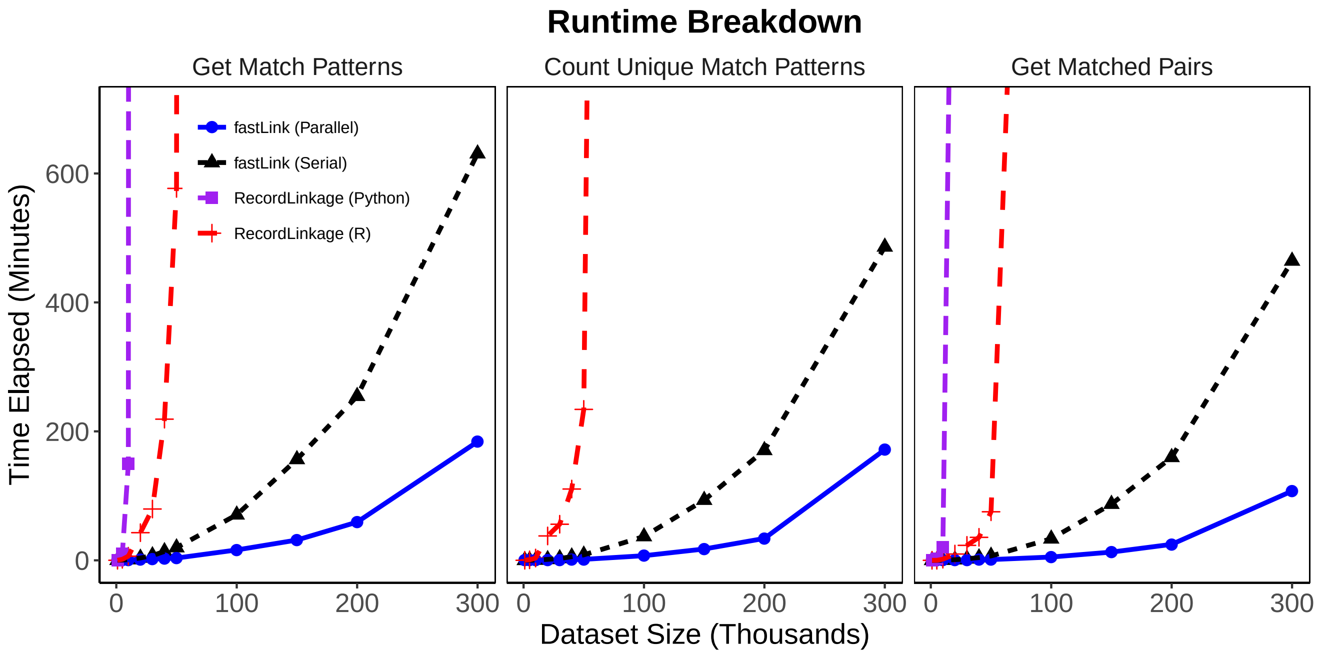

Figure S1: Runtime Comparison Breakdown for Open-source Probabilistic Record Linkage Implementations. The left plot shows the amount of time spent counting the matches between pairs for each matching variable, the middle plot shows the amount of time spent tabulating the unique observed match patterns, and the right plot gives the amount of time spent identifying and returning all pairs identified as matches after the model has been estimated. In all segments, fastLink outperforms the other open-source implementations.

left-hand plot of Figure S1 compares the amount of time spent counting the matches between pairs for each matching variable, the middle plot compares the amount of time spent tabulating the unique observed match patterns across all matching variables, while the last plot shows the time spent returning the matched pairs after estimating the parameters of the Fellegi-Sunter model. For the Python package RecordLinkage, all calculations are done on the cross-product of the two data sets being examined, so there is no stage to benchmark against where all unique match patterns are counted. In all three stages, fastLink dramatically outperforms the other open-source implementations of the Fellegi-Sunter model. Furthermore, parallelization improves the runtime of fastLink even further.

Note that we were only able to run the Python package RecordLinkage for datasets of at most 20,000 observations. Even after increasing the available RAM to 128 GB and running the algorithm for over 24 hours, we would hit memory limits that kept us from running larger matches. Furthermore, for the R package RecordLinkage, we were only able to run their algorithm for datasets of at most 40,000 observations. After running their algorithm for over 24 hours on a dataset of 50,000 observations, we ran into numerical underflow issues that prevented us from completing the match. In order to extrapolate runtime to larger data sets for the sake of presentation, we fit an exponential regression model where the runtime for the completed merges for each package was regressed onto the size of the datasets being merged. We fit this model separately for each software package and each stage of the merge process that we benchmarked. Predictions from these regression models are included in each plot of Figure S1.

S3.2 Parallelization and Random Sampling

Under the proposed probabilistic modeling approach, a vast majority of the computational burden is due to the enumeration of agreement patterns. In fact, the actual computation time of implementing the E and M steps, once hashing is done, is fast even for large data sets. Therefore, for further computational efficiency, we parallelize the enumeration of agreement patterns. Specifically, we take a divide-and-conquer approach by partitioning the two data sets, $\mathcal{A}$ and $\mathcal{B}$, into equally-sized subsets such that $\mathcal{A} = \{\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_{M_\mathcal{A}}\}$ and $\mathcal{B} = \{\mathcal{B}_1, \mathcal{B}_2, \ldots, \mathcal{B}_{M_\mathcal{B}}\}$. Then, using the OpenMP architecture, we count the agreement patterns for each partition pair $\{\mathcal{A}_i, \mathcal{B}_j\}$ in parallel using the hash function given in equation (15). As explained above, we utilize sparse matrix objects to efficiently store agreement patterns for all pairwise comparisons. Finally, the entire pattern-counting procedure is implemented in C++ for additional performance gains. Taken together, our approach provides simple parallelization of the pattern-counting procedure and efficient memory usage so that our linkage procedure can be applied to arbitrarily large problems.

Another advantage of the probabilistic modeling approach is the use of random sampling. Since the number of parameters, i.e., λ and $\boldsymbol{\pi}$, is relatively small, we can efficiently estimate these parameters using a small random subset. For example, in our empirical application presented in Section 4.2, we find that 800,000 observations, which is 5% of the original 16 million observations in the California voter file, is sufficient to obtain identical parameter estimates (see Appendix K for simulation studies we use to guide this choice of sample size). Once we obtain the parameter estimates, then we can compute the match probabilities for every agreement pattern found in the entire data sets in parallel. In this way, we are able to scale the model to massive data sets as illustrated in our empirical applications.

S4 Relaxing the Conditional Independence Assumption

In this appendix, we present two ways to relax the conditional independence assumption while keeping our scalable implementation. The first strategy is to combine the information in pairs of variables so that their correlation can be accounted for. This is easy to implement but is not generalizable. The second strategy is to follow the literature and use log-linear models, which can accommodate more general patterns of correlations across variables (see e.g., Winkler, 1989, 1993; Thibaudeau, 1993; Larsen and Rubin, 2001, and references therein).

S4.1 Combining the Information in Pairs of Variables

This strategy is better explained with a simple example. Consider the following contingency table of binary agreements.

First name	Last name	Date of birth	count
1	1	1	n_1
1	1	0	n_2
1	0	1	n_3
1	0	0	n_4
0	1	1	n_5
0	1	0	n_6
0	0	1	n_7
0	0	0	n_8

We combine two variables, e.g., first name and a date of birth, so that their correlation can be accounted for. This yields the following contingency table of three variables.

First name − Date of birth	Last name	count
1−1	1	n_1
1−0	1	n_2
1−1	0	n_3
1−0	0	n_4
0−1	1	n_5
0−0	1	n_6
0−1	0	n_7
0−0	0	n_8

For the new combined variable, we can have four possible values. We may order them as: 3 = 1−1, 2 = 0−1, 1 = 1−0, and 0 = 0−0. We can then use the EM algorithm as in the standard case. Fortunately, this modification has little impact on the scalability of our implementation. The reason is that fastLink constructs the whole agreement vector for each pair through its hashing implementation, and hence this gives sufficient information for combining pairs of variables.

S4.2 Log-linear Modeling

The previous approach is simple, but cannot accommodate a more general pattern of interaction. It is also difficult to apply when variables have missing values and measurement error. We discuss an alternative approach based on log-linear models that has been developed in the literature.

The observed data likelihood function of the Fellegi-Sunter model *without* the conditional independence assumption is given by,

$$\mathcal{L}_{obs}(\lambda, \boldsymbol{\theta} \mid \boldsymbol{\delta}, \boldsymbol{\gamma}) \;=\; \prod_{i=1}^{N_{\mathcal{A}}} \prod_{j=1}^{N_{\mathcal{B}}} \left\{ \sum_{m=0}^{1} \lambda^m (1-\lambda)^{1-m} \boldsymbol{\pi}_m(i,j;\boldsymbol{\theta}_m) \right\}$$

where $\boldsymbol{\pi}_m(i,j) = \Pr(\gamma(i,j) \mid M_{ij} = m, \boldsymbol{\theta}_m)$ for $m \in \{0,1\}$ and $\boldsymbol{\theta}_m$ represents a vector of model parameters. The corresponding complete data log-likelihood function is,

$$\log \mathcal{L}_{com}(\lambda, \boldsymbol{\theta} \mid \boldsymbol{\delta}, \boldsymbol{\gamma}) \;=\; \sum_{i=1}^{N_{\mathcal{A}}} \sum_{j=1}^{N_{\mathcal{B}}} \{ M_{ij} \log(\lambda) + (1 - M_{ij}) \log(1-\lambda) +$$

$$M_{ij} \log(\boldsymbol{\pi}_1(i,j;\boldsymbol{\theta}_1)) + (1 - M_{ij}) \log(\boldsymbol{\pi}_0(i,j;\boldsymbol{\theta}_0))\}$$

As in the case of the conditional independence assumption, the parameters will be estimated via the EM algorithm, where the E-step takes the following form,

$$\xi_{ij} \;=\; \frac{\lambda \boldsymbol{\pi}_1(i,j;\boldsymbol{\theta}_1)}{\lambda \boldsymbol{\pi}_1(i,j;\boldsymbol{\theta}_1) + (1-\lambda)\boldsymbol{\pi}_0(i,j;\boldsymbol{\theta}_0)}$$

The M-Step then is as follows,

$$\begin{aligned}
\lambda &= \frac{1}{N_{\mathcal{A}} N_{\mathcal{B}}} \sum_{i=1}^{N_{\mathcal{A}}} \sum_{j=1}^{N_{\mathcal{B}}} \xi_{ij} \\
\boldsymbol{\theta}_m &= \underset{\boldsymbol{\theta}_m^*}{\operatorname{argmax}} \sum_{i=1}^{N_{\mathcal{A}}} \sum_{j=1}^{N_{\mathcal{B}}} \xi_{ij}^m (1-\xi_{ij})^{1-m} \log(\boldsymbol{\pi}_m(i,j;\boldsymbol{\theta}_m^*))
\end{aligned}$$

As noted by Larsen and Rubin (2001), Murray (2016), and others, this second M-step corresponds to the optimization of the weighted log-likelihood function for the log-linear model with a contingency table. The log-linear model allows for the inclusion of interaction terms and can be fitted via the iterative proportional fitting procedure or the quasi-poisson regression.

S4.3 Simulation Results

We next evaluate the accuracy of FDR and FNR estimates under the log-linear modeling approach. Figure S2 presents the results. We compare the performance of the proposed methodology across different levels of measurement error. In particular, the log-linear model we use includes all two-way interactions across fields for the set of non-matches, while no interactions are added for the set of matches. In addition, we consider two datasets of equal size (100,000 records each), three levels of measurement error: 20%, 50%, and 80%, and non-differential measurement error with dependence across string-valued linkage fields (first name, last name, and street name) as the error structure. Missing values drawn completely at random are added at a 6% rate in all linkage fields other than year of birth. See Section I.2 for a detailed description of how the simulation datasets with measurement error are constructed.

We find that the true FDR is low and its estimate is accurate across different degrees of overlap and measurement error (see the top panel of Figure S2). As shown in the bottom panel of the figure, the true FNR is greater especially when the amount of measurement error is large. However, its estimate remains relatively accurate even when the overlap is as small as 20%. This represents an improvement over the performance of the conditional independence model shown in Figure 2. This suggest that the log-linear model helps improve the estimation of error rates even when the overlap of data sets is limited and the amount of measurement error is large.

In a separate simulation appendix, we present all the results for the 135 simulation scenarios in which measurement error is the main source of noise. We consider the Fellegi-Sunter models with and without the conditional independence assumptions. As summarized in Appendix I below, once the conditional independence assumption is relaxed, the Fellegi-Sunter model yields more precise estimates even when the datasets being linked have unequal sizes.

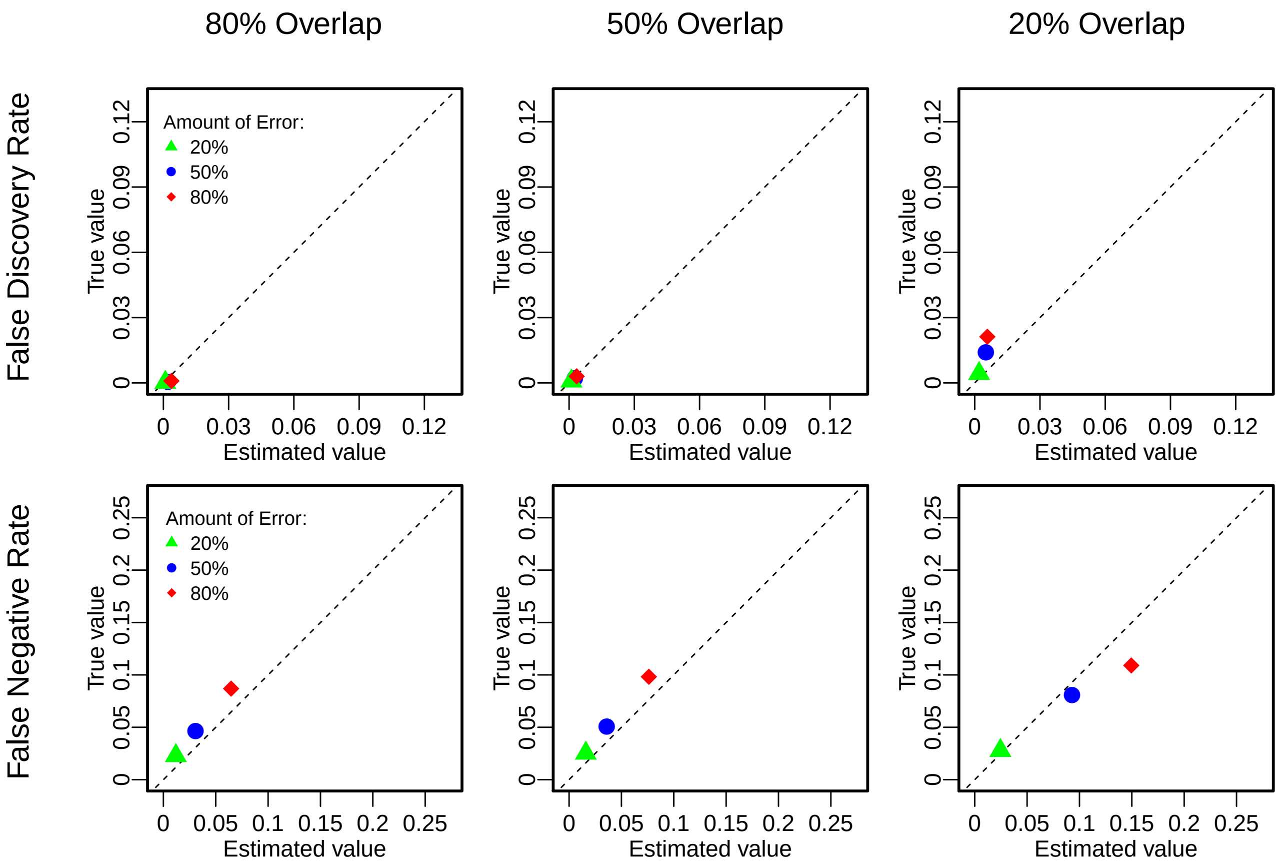

Figure S2: Accuracy of FDR and FNR Estimates under Measurement Error for the Fellegi-Sunter model allowing for dependences across linkage fields. The top panel compares the estimated FDR (x-axis) with its true value (y-axis) whereas the bottom panel compares the estimated FNR against its true value. We consider two datasets of equal size (100,000 records each), non-differential measurement error with dependence across string-valued linkage fields (first name, last name, and street name) as the error structure, and three levels of measurement error: 20% (green triangles), 50% (blue circles), 80% (red diamonds). In addition, we added a 6% share of missing values drawn completely at random in all linkage fields other than year of birth. Overall, we can see that the FDR and FNR estimates are accurate across different levels of overlap, and that the FNR tends to be larger as the amount of measurement error increases.

S5 Enforcing One-to-One Merge

Although we generally recommend that researchers use all matched observations by weighting them according to the match probability, there exist several procedures that can be used to enforce one-to-one merge if researchers wish to do so. First, researchers may choose the record with the greatest match probability where a tie is broken with random sampling. This procedure is simple but may introduce bias if several observations have a relatively large value of match probability. Alternatively, researchers may randomly sample a record with the selection probability proportional to the match probability. This procedure eliminates the aforementioned bias.

However, these two procedures may end up matching multiple records in $\mathcal{A}$ with an identical record in the data set $\mathcal{B}$, which may not be desirable in some cases. To avoid this problem, matching can be done without replacement but this significantly increases computational burden.

For moderate or small data sets, many researchers use the method proposed by Jaro (1989), which resolves the problem of duplicate matches by maximizing the sum of the matching probabilities for matched pairs. Jaro (1989) shows that this can be formulated as the linear sum assignment problem, for which an efficient algorithm exists. In addition, Winkler (1994) proposes a similar approach that generalizes and further improves the method of Jaro (1989). However, as noted by Winkler (1994), we find that these algorithms are prohibitively slow for large data sets. Thus, we reduce the number of assignments by focusing on those pairs whose probabilities are greater than a threshold c, which could be set, for example, to 0.85.

Specifically, suppose that there are D_c pairs with $\xi_{ij} \geq c$ and let $\mathcal{A}_c$ and $\mathcal{B}_c$ represent the set of unique observations in the data set $\mathcal{A}$ and $\mathcal{B}$, respectively, that belong to at least one of these pairs. Then, the assignment problem can be written as,

$$\begin{aligned} \text{maximize} \sum_{i \in \mathcal{A}_c} \sum_{j \in \mathcal{B}_c} \xi_{ij} M_{ij} \quad \text{subject to} \quad \sum_{i \in \mathcal{A}_c} M_{ij} &\leq 1 \text{ for each } j \in \mathcal{B}_c, \\ \text{and} \quad \sum_{j \in \mathcal{B}_c} M_{ij} &\leq 1 \text{ for each } i \in \mathcal{A}_c \end{aligned} \tag{S3}$$

where $M_{ij} \in \{0, 1\}$ for all (i, j). To turn this into the linear sum assignment problem, we must have a one-to-one match, i.e., $|\mathcal{A}| = |\mathcal{B}|$, $\sum_{i \in \mathcal{A}} M_{ij} = 1$ and $\sum_{j \in \mathcal{B}} M_{ij} = 1$. Following Jaro (1989), we add artificial observations to the smaller of the two data sets with zero probabilities for all of their potential matches, i.e., $\xi_{ij} = 0$, such that these constraints are satisfied.

S6 Incorporating Auxiliary Information

An advantage of the probabilistic model we introduce is that we can incorporate auxiliary information. This point has not been emphasized enough in the literature. We first describe how to adjust for the fact that some names are more common than others. We then consider how to incorporate aggregate information about migration.

S6.1 Incorporating Name Frequencies

Some names are more common than others and as such they are more likely to contribute to false matches. Unfortunately, it is difficult to incorporate this information directly into the model and estimation without significantly increasing computational burden. Instead, inspired by the previous work (Newcombe et al., 1959; Newcombe and Kennedy, 1962; Fellegi and Sunter, 1969; Winkler, 2000), we make an ex-post adjustment to the match probabilities. Unlike Winkler (2000), however, we do not assume that the frequency of a name in the set of true matches is equal to the frequency of the same name in a smaller data set. Finally, we note that if many informative variables such as date of birth and addresses are available, these name adjustments may make little difference (Yancey, 2000).

Let $f_p^{\mathcal{A}}$ and $f_p^{\mathcal{B}}$ represent the frequency of each name $p = 1, \ldots, P$ for the data set $\mathcal{A}$ and the data set $\mathcal{B}$, respectively, where P is the total number of unique names that appear in the two data sets. These frequencies can be obtained from the Census data. When such data are not available, one may also use the sample frequencies in the data sets $\mathcal{A}$ and $\mathcal{B}$ as the estimates such that $\sum_{p=1}^{P} \hat{f}_p^{\mathcal{A}} = N_{\mathcal{A}}$ and $\sum_{p=1}^{P} \hat{f}_p^{\mathcal{B}} = N_{\mathcal{B}}$. Following Winkler (2000), we assume that the conditional probability of agreement in the name field given the match status is identical across different names p, i.e.,

$$\Pr(\gamma_{\text{name}}(i,j) = 1 \mid \text{name}_i = \text{name}_j = p, M_{ij} = m) \; = \; \Pr(\gamma_{\text{name}}(i,j) = 1 \mid \text{name}_i = \text{name}_j, M_{ij} = m)$$

for $m = 0, 1$ and $p = 1, 2, \ldots, P$. The assumption may be violated, for example, if certain names are more likely to be spelled incorrectly (even after conditioning on the true match status). Appendix F.3.1 shows how to make adjustments to the probability under this assumption.

S6.2 Incorporating Aggregate Migration Rates

One of the challenges researchers face when merging administrative records over time is the identification of those individuals who move from one address to another. The fact that the addresses of these movers differ between the two data sets reduces their match probabilities, leading to a greater chance of a false negative. However, aggregate migration rates are often available from other data sources. For example, in one of our applications, we use the annual migration data for the United States are available from the Internal Revenue Service based on individual tax returns (see `https://www.irs.gov/uac/soi-tax-stats-migration-data`). Here, we show how to incorporate such aggregate migration rates into our model. Unlike Steorts (2015), our prior distribution is based on auxiliary data rather than the data used for merging.

Suppose that we wish to merge two data sets $\mathcal{A}$ and $\mathcal{B}$ for the same state but measured at different points of time. We further assume that other data sources give us the number of migrants out of and into the state as well as the number of in-state movers during this time period. Then, we form the prior mean of the probability of a match, i.e., $\lambda = \Pr(M_{ij} = 1)$, as,

$$\lambda^{\text{prior}} \;=\; \frac{\#\text{ of non-movers} + \#\text{ of within-state movers}}{N_{\mathcal{A}} \times N_{\mathcal{B}}} \tag{S4}$$

Additionally, we formulate the prior mean of the probability that a matched pair has different addresses, i.e., $\pi_{\text{adr},1,0} = \Pr(\gamma_{\text{adr}}(i,j) = 0 \mid M_{ij} = 1)$, as,

$$\pi_{\text{adr},1,0}^{\text{prior}} \;=\; \frac{\#\text{ of in-state movers}}{\#\text{ of in-state movers} + \#\text{ of non-movers}} \tag{S5}$$

We recommend that users specify a binary match for the address field when incorporating prior information on in-state movers. This avoids the unrealistic assumption that a partial match on address for a pair in the matched set is as likely as an exact match on address, i.e., $\Pr(\gamma_{\text{adr}}(i,j) = \ell \mid M_{ij} = 1) = \Pr(\gamma_{\text{adr}}(i,j) = \ell' \mid M_{ij} = 1)$ for all $\ell, \ell' \neq 0$.

We use the above prior mean for the conjugate prior distributions on λ, i.e., $\text{Beta}(a_\lambda, b_\lambda)$, and on $\pi_{\text{adr},1,0}$, i.e., $\text{Beta}(a_{\text{adr}}, b_{\text{adr}})$, while leaving the priors for the other parameters improper. To specify these two hyperparameters of each prior distribution, we require researchers to attach the weight to the prior information relative to the data and specify the prior means given in equations (S4) and (S5). Let w_λ and w_{adr} denote these weights for λ and $\pi_{\text{adr},1,0}$, which range from zero to one. For example, if we specify $w_\lambda = w_{\text{adr}} = 0.9$, then the resulting estimates of

λ and $\pi_{\text{adr},1,0}$ will be approximately equal to the weighted averages of their corresponding ML estimates and the prior mean where the former accounts for 10% and the latter makes up 90%. Appendix F.3.2 describes the details of the EM algorithm that integrates this prior information.

S6.3 Details of Incorporating Auxiliary Information

In this appendix, we describe the details of incorporating two types of auxiliary information: name frequencies and aggregate migration rates.

S6.3.1 Name Frequencies

We show how to make an ex-post adjustment to the match probability given in equation (4) for any given pair whose records have an identical name. Under this assumption described in Section F.1, we have,

$$\begin{aligned} & \Pr(\gamma_{\text{name}}(i,j)=1, \text{name}_i=\text{name}_j=p \mid M_{ij}=m) \\ = \;& \Pr(\gamma_{\text{name}}(i,j)=1 \mid \text{name}_i=\text{name}_j, M_{ij}=m) \times \Pr(\text{name}_i=\text{name}_j=p \mid M_{ij}=m) \end{aligned} \tag{S6}$$

for $p=1,2,\ldots,P$ and $m=0,1$ where $\text{name}_i=p$ ($\text{name}_j=p$) implies that observation i in the data set $\mathcal{A}$ (observation j in the data set $\mathcal{B}$) has name p.

Thus, we have,

$$\begin{aligned} & \Pr(M_{ij}=1 \mid \boldsymbol{\delta}(i,j), \boldsymbol{\gamma}(i,j), \text{name}_i=\text{name}_j=p) \\ = \;& \frac{\lambda \cdot \Pr(\boldsymbol{\gamma}(i,j), \boldsymbol{\delta}(i,j), \text{name}_i=\text{name}_j=p \mid M_{ij}=1)}{\sum_{m=0}^{1} \lambda^m (1-\lambda)^{1-m} \cdot \Pr(\boldsymbol{\gamma}(i,j), \boldsymbol{\delta}(i,j), \text{name}_i=\text{name}_j=p \mid M_{ij}=m)} \end{aligned} \tag{S7}$$

where the conditional independence assumption given in equation (S6) implies,

$$\begin{aligned} & \Pr(\boldsymbol{\gamma}(i,j), \boldsymbol{\delta}(i,j), \text{name}_i=\text{name}_j=p \mid M_{ij}=m) \\ = \;& \Pr(\text{name}_i=\text{name}_j=p \mid M_{ij}=m) \times \prod_{k=1}^{K} \left(\prod_{\ell=0}^{L_k-1} \pi_{km\ell}^{\mathbf{1}\{\gamma_k(i,j)=\ell\}} \right)^{1-\delta_k(i,j)} \end{aligned} \tag{S8}$$

for $m=0,1$ and $p=1,2,\ldots,P$. By Bayes' rule, the adjustment factors are estimated as,

$$\begin{aligned} & \Pr(\text{name}_i=\text{name}_j=p \mid M_{ij}=m) \\ = \;& \frac{\sum_{i'=1}^{N_{\mathcal{A}}} \sum_{j'=1}^{N_{\mathcal{B}}} \xi_{i'j'}^m (1-\xi_{i'j'})^{1-m} \mathbf{1}\{\text{name}_{i'}=\text{name}_{j'}=p\}}{\sum_{i'=1}^{N_{\mathcal{A}}} \sum_{j'=1}^{N_{\mathcal{B}}} \mathbf{1}\{\text{name}_{i'}=\text{name}_{j'}=p\}} \times \frac{f_p^{\mathcal{A}} f_p^{\mathcal{B}}}{N_{\mathcal{A}} \times N_{\mathcal{B}}} \times \frac{1}{\lambda^m (1-\lambda)^{1-m}} \end{aligned} \tag{S9}$$

If we substitute equations (S8) and (S9) into equation (S7), we obtain,

$$\begin{aligned} & \Pr(M_{ij}=1 \mid \boldsymbol{\delta}(i,j), \boldsymbol{\gamma}(i,j), \text{name}_i=\text{name}_j=p) \\ = \;& \frac{\left[\sum_{i'=1}^{N_{\mathcal{A}}} \sum_{j'=1}^{N_{\mathcal{B}}} \xi_{i'j'} \mathbf{1}\{\text{name}_{i'}=\text{name}_{j'}=p\} \right] \cdot \left(\prod_{\ell=0}^{L_k-1} \pi_{k1\ell}^{\mathbf{1}\{\gamma_k(i,j)=\ell\}} \right)^{1-\delta_k(i,j)}}{\sum_{m=0}^{1} \left[\sum_{i'=1}^{N_{\mathcal{A}}} \sum_{j'=1}^{N_{\mathcal{B}}} \xi_{i'j'}^m (1-\xi_{i'j'})^{1-m} \mathbf{1}\{\text{name}_{i'}=\text{name}_{j'}=p\} \right] \cdot \left(\prod_{\ell=0}^{L_k-1} \pi_{km\ell}^{\mathbf{1}\{\gamma_k(i,j)=\ell\}} \right)^{1-\delta_k(i,j)}} \end{aligned}$$

where for our ex-post adjustment we use the maximum likelihood estimates $\hat{\pi}_{km\ell}$ and $\hat{\xi}_{ij}$.

S6.3.2 Aggregate Migration Rates

We derive the hyperparameters of prior distribution by specifying the prior means and the weights researchers attach to the prior mean relative to the ML estimates. We begin with the derivation of hyperparameters for the prior distribution of λ. Recall that with our conjugate prior Beta(a_λ, b_λ), the M-step can be written as,

$$\lambda \;=\; \frac{1}{b_\lambda - a_\lambda + N_\mathcal{A} N_\mathcal{B}} \left(a_\lambda - 1 + \sum_{i=1}^{N_\mathcal{A}} \sum_{j=1}^{N_\mathcal{B}} \xi_{ij} \right) \tag{S10}$$

We rewrite this expression as a weighted average of the ML estimate and the prior mean,

$$\lambda \;=\; \frac{1}{\frac{b_\lambda - a_\lambda}{N_\mathcal{A} N_\mathcal{B}} + 1} \times \underbrace{\frac{\sum_{i=1}^{N_\mathcal{A}} \sum_{j=1}^{N_\mathcal{B}} \xi_{ij}}{N_\mathcal{A} N_\mathcal{B}}}_{\text{ML estimate}} + \frac{\frac{(a_\lambda - 1)(a_\lambda + b_\lambda)}{a_\lambda N_\mathcal{A} N_\mathcal{B}}}{\frac{b_\lambda - a_\lambda}{N_\mathcal{A} N_\mathcal{B}} + 1} \times \underbrace{\frac{a_\lambda}{a_\lambda + b_\lambda}}_{\text{Prior mean}}$$

While the coefficients for the ML estimate and the prior mean in this equation do not add up exactly to 1, we show below that their sum is approximately 1, enabling us to interpret them as weights. Some algebraic manipulation shows,

$$\begin{aligned} \frac{1}{\frac{b_\lambda - a_\lambda}{N_\mathcal{A} N_\mathcal{B}} + 1} + \frac{\frac{(a_\lambda - 1)(a_\lambda + b_\lambda)}{a_\lambda N_\mathcal{A} N_\mathcal{B}}}{\frac{b_\lambda - a_\lambda}{N_\mathcal{A} N_\mathcal{B}} + 1} &= \frac{N_\mathcal{A} N_\mathcal{B}}{b_\lambda - a_\lambda + N_\mathcal{A} N_\mathcal{B}} + \frac{a_\lambda - 1}{a_\lambda} \frac{a_\lambda + b_\lambda}{b_\lambda - a_\lambda + N_\mathcal{A} N_\mathcal{B}} \\ &\approx \frac{N_\mathcal{A} N_\mathcal{B} + a_\lambda + b_\lambda}{b_\lambda - a_\lambda + N_\mathcal{A} N_\mathcal{B}} \end{aligned}$$

The approximation follows because for large data sets, we typically have $a_\lambda \gg 1$ so that $(a_\lambda - 1)/a_\lambda \approx 1$ (see equation (S10)). In addition, since λ is a small number and at most $\frac{\min(N_\mathcal{A}, N_\mathcal{B})}{N_\mathcal{A} N_\mathcal{B}}$, a_λ is negligible compared to b_λ. Hence, the sum of the weights effectively reduces to 1. This leads to the following two equalities,

$$\lambda^{\text{prior}} \;=\; \frac{a_\lambda}{a_\lambda + b_\lambda} \quad \text{and} \quad \frac{w_\lambda}{1 - w_\lambda} \;=\; \frac{(a_\lambda - 1)(a_\lambda + b_\lambda)}{a_\lambda N_\mathcal{A} N_\mathcal{B}}.$$

Solving these equations yields,

$$a_\lambda \;=\; \frac{w_\lambda}{1 - w_\lambda} \lambda^{\text{prior}} N_\mathcal{A} N_\mathcal{B} + 1 \quad \text{and} \quad b_\lambda \;=\; \frac{(1 - \lambda^{\text{prior}}) a_\lambda}{\lambda^{\text{prior}}}.$$

We determine the values of hyperparameters for the prior distribution of $\pi_{\text{adr},1,0}$ in the same way. First, recall that the M-Step that incorporates prior information is,

$$\tilde{\pi}_{\text{adr},1,0} \;=\; \frac{\sum_{i=1}^{N_\mathcal{A}} \sum_{j=1}^{N_\mathcal{B}} \mathbf{1}\{\gamma_k(i,j) = l)\} (1 - \delta_k(i,j)) \xi_{ij}^m (1 - \xi_{ij})^{1-m} + (a_{\text{adr}} - 1)}{\sum_{i=1}^{N_\mathcal{A}} \sum_{j=1}^{N_\mathcal{B}} (1 - \delta_k(i,j)) \xi_{ij}^m (1 - \xi_{ij})^{1-m} + (a_{\text{adr}} - 1) + (b_{\text{adr}} - 1)}.$$

We reexpress this equation as the following weighted average of the ML estimate and the prior mean,

$$\tilde{\pi}_{\text{adr},1,0} \;=\; \frac{1}{1 + \frac{(a_{\text{adr}} - 1) + (b_{\text{adr}} - 1)}{\lambda^{\text{prior}} N_\mathcal{A} N_\mathcal{B}}} \underbrace{\hat{\pi}_{\text{adr},1,0}}_{\text{ML estimate}} + \frac{\frac{(a_{\text{adr}} - 1)(a_{\text{adr}} + b_{\text{adr}})}{a_{\text{adr}} \lambda^{\text{prior}} N_\mathcal{A} N_\mathcal{B}}}{1 + \frac{(a_{\text{adr}} - 1) + (b_{\text{adr}} - 1)}{\lambda^{\text{prior}} N_\mathcal{A} N_\mathcal{B}}} \underbrace{\frac{a_{\text{adr}}}{a_{\text{adr}} + b_{\text{adr}}}}_{\text{Prior mean}}$$

where we use prior information about λ by replacing the term $\sum_{i=1}^{N_{\mathcal{A}}}\sum_{j=1}^{N_{\mathcal{B}}}(1-\delta_k(i,j))\xi_{ij}^m(1-\xi_{ij})^{1-m}$, which is equal to the expected number of matches, with $\lambda^{\text{prior}}N_{\mathcal{A}}N_{\mathcal{B}}$. Then, we can show that the sum of the coefficients is approximately equal to 1,

$$\begin{aligned}
& \frac{1}{1+\frac{(a_{\text{adr}}-1)+(b_{\text{adr}}-1)}{\lambda^{\text{prior}}N_{\mathcal{A}}N_{\mathcal{B}}}} + \frac{\frac{(a_{\text{adr}}-1)(a_{\text{adr}}+b_{\text{adr}})}{a_{\text{adr}}\lambda^{\text{prior}}N_{\mathcal{A}}N_{\mathcal{B}}}}{1+\frac{(a_{\text{adr}}-1)+(b_{\text{adr}}-1)}{\lambda^{\text{prior}}N_{\mathcal{A}}N_{\mathcal{B}}}} \\
= & \frac{\lambda^{\text{prior}}N_{\mathcal{A}}N_{\mathcal{B}}}{\lambda^{\text{prior}}N_{\mathcal{A}}N_{\mathcal{B}}+(a_{\text{adr}}-1)+(b_{\text{adr}}-1)} + \frac{a_{\text{adr}}-1}{a_{\text{adr}}}\frac{(a_{\text{adr}}-1)+(b_{\text{adr}}-1)}{\lambda^{\text{prior}}N_{\mathcal{A}}N_{\mathcal{B}}+(a_{\text{adr}}-1)+(b_{\text{adr}}-1)} \\
\approx & \ 1
\end{aligned}$$

where the approximation follows from the fact that for large data sets we have $a_{\text{adr}} \gg 1$. Finally, we obtain the hyperparameters of the prior distribution by solving the following equations,

$$\pi_{\text{adr},1,0}^{\text{prior}} = \frac{a_{\text{adr}}}{a_{\text{adr}}+b_{\text{adr}}}, \quad \text{and} \quad \frac{w_{\text{adr}}}{1-w_{\text{adr}}} = \frac{(a_{\text{adr}}-1)(a_{\text{adr}}+b_{\text{adr}})}{a_{\text{adr}}\lambda^{\text{prior}}N_{\mathcal{A}}N_{\mathcal{B}}}$$

The result is given by,

$$a_{\text{adr}} = \frac{w_{\text{adr}}}{1-w_{\text{adr}}}\pi_{\text{adr},1,0}^{\text{prior}}\lambda^{\text{prior}}N_{\mathcal{A}}N_{\mathcal{B}}+1, \quad \text{and} \quad b_{\text{adr}} = \frac{(1-\pi_{\text{adr},1,0}^{\text{prior}})a_{\text{adr}}}{\pi_{\text{adr},1,0}^{\text{prior}}}.$$

S7 The Properties of the Weighted Maximum Likelihood Estimator

We show that the expected value of the weighted log-likelihood function given in equation (13) is equal to the expected value of the log-likelihood function of the original model defined in equation (12),

$$\begin{aligned}
& \mathbb{E}\left[\int \xi_{ij}^* \log P_\theta(Y_i \mid Z_i^*, \mathbf{X}_i)\ dZ_i\right] \\
= & \ \mathbb{E}\left[\int P(Z_i^* \mid \boldsymbol{\gamma}, \boldsymbol{\delta})\left\{\int \log P_\theta(Y_i \mid Z_i^*, \mathbf{X}_i)\ P(Y_i \mid Z_i^*, \mathbf{X}_i, \boldsymbol{\gamma}, \boldsymbol{\delta})\ dY_i\right\} dZ_i^*\right] \\
= & \ \mathbb{E}\left[\int\int \log P_\theta(Y_i \mid Z_i^*, \mathbf{X}_i)\ \frac{P(Y_i \mid Z_i^*, \mathbf{X}_i, \boldsymbol{\gamma}, \boldsymbol{\delta})P(\mathbf{X}_i \mid Z_i^*, \boldsymbol{\gamma}, \boldsymbol{\delta})P(Z_i^* \mid \boldsymbol{\gamma}, \boldsymbol{\delta})}{P(\mathbf{X}_i \mid \boldsymbol{\gamma}, \boldsymbol{\delta})}\ dY_i\ dZ_i^*\right] \\
= & \ \mathbb{E}\left[\int\int \log P_\theta(Y_i \mid Z_i^*, \mathbf{X}_i)\ P(Y_i, Z_i^* \mid \boldsymbol{\gamma}, \boldsymbol{\delta}, \mathbf{X}_i)\ dY_i\ dZ_i^*\right] \\
= & \ \mathbb{E}\left[\mathbb{E}\{\log P_\theta(Y_i \mid Z_i^*, \mathbf{X}_i) \mid \boldsymbol{\gamma}, \boldsymbol{\delta}, \mathbf{X}_i\}\right] \ = \ \mathbb{E}\{\log P_\theta(Y_i \mid Z_i^*, \mathbf{X}_i)\}
\end{aligned}$$

where the second equality follows from equations (7) and (10). Under mild regularity conditions, we can show that the weighted ML estimator is asymptotically normal,

$$\sqrt{N_{\mathcal{A}}}(\hat{\theta}-\theta_0) \ \rightsquigarrow \ \mathcal{N}(0,\ \Omega^{-1}\Delta\Omega^{-1})$$

where

$$\begin{aligned}
\Omega &= -\mathbb{E}\left[\left(\frac{\partial^2}{\partial\theta\partial\theta^\top}\sum_{j=1}^{N_\mathcal{B}}\xi_{ij}^*\log P_\theta(Y_i \mid Z_i^* = Z_j, \mathbf{X}_i)\right)_{\theta_0}\right] \\
\Delta &= \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\sum_{j=1}^{N_\mathcal{B}}\xi_{ij}^*\log P_\theta(Y_i \mid Z_i^* = Z_j, \mathbf{X}_i)\right)_{\theta_0}\left(\frac{\partial}{\partial\theta}\sum_{j=1}^{N_\mathcal{B}}\xi_{ij}^*\log P_\theta(Y_i \mid Z_i^* = Z_j, \mathbf{X}_i)\right)_{\theta_0}^\top\right]
\end{aligned}$$

and θ_0 is the true value of θ.

S8 Comparison with Alternative Probabilistic Methods

Although many social scientists use deterministic methods, the probabilistic modeling has been a predominant approach in the statistics literature ever since the publication of Fellegi and Sunter (1969). Recently, some researchers have proposed alternative probabilistic models, which consider data merging as the problem of forming a bipartite graph (e.g., Steorts, 2015; Steorts, Hall and Fienberg, 2016; Sadinle, 2017). One advantage of this approach is that the one-to-one match restriction can be imposed as part of the model rather than as a post-match processing step.

However, their major drawback is the lack of scalability. In particular, they are implemented using Markov chain Monte Carlo methods and cannot merge large administrative records, which our methodology is designed to handle. For this reason, we do not examine these alternative probabilistic models in Section 3.3. Indeed, while fastLink took under 10 seconds on a single core machine to merge two data sets of 1,000 observations as shown in Figure 3, the bipartite graph implementation proposed by Sadinle (2017) took 109 seconds to run 1,000 iterations of the Gibbs sampler with a 100-iteration burn-in period. Merging the same data sets under the same settings for the bipartite graph model proposed by Steorts (2015) and implemented in the R package `blink` took even longer — over 7,800 seconds or 130 minutes. Furthermore, merging two data sets of 5,000 observations with `blink` took over 16 hours to run, which is more than 1,800 times longer than fastLink took to complete the same merge on a single core. For any data set larger than 1,000 observations, the Sadinle (2017) implementation ran into memory errors, while `blink` had not completed any merge larger than 5,000 observations after running for more than a day.

For the sake of completeness, however, this appendix compares the accuracy of fastLink with that of these alternative probabilistic methods using the small-scale validation data sets analyzed by the original authors (Steorts, 2015; Sadinle, 2017). To summarize the results, we find that the accuracy of our algorithm is comparable to that of these state-of-art methodologies.

We begin with the method proposed by Steorts (2015) who uses, as a validation data set, the RLdata500 data set of only 500 records, which is part of the RecordLinkage package in R. The validation data set contains five linkage variables (first and last name, day, month, and year of birth) where noise is added to one randomly selected variable for any given record. The goal of this validation study is to identify 50 records that are known to be duplicates by matching one observation against another within this data set. Steorts (2015) reports that the FNR and FDR of her methodology are estimated to be 0.02 and 0.04, respectively.

Procedure	FNR	FDR
Empirical Bayes	0.020	0.040
fastLink	0.003	0.000

Table S1: Error Rates obtained from Deduplication of the RLdata500 Data Set. Empirical Bayes results are obtained from Steorts (2015), while fastLink represents the estimates obtained from our implementation of the Fellegi-Sunter.

When applying our algorithm, we use three categories for the string valued variables (first and last names), i.e., exact (or nearly identical) match, partial match, and disagreement, based on the Jaro-Winkler string distance with 0.94 and 0.88 as the cutpoints following the recommendation of Winkler (1990). For the numeric valued fields (day, month, and year of birth), we use a binary comparison, based on exact matches. As shown in Table S1, we find the estimated FNR and FDR of fastLink to be exactly zero, outperforming the empirical Bayes method proposed by Steorts (2015) in this particular validation study.

Next, we compare the accuracy of fastLink with that of the method proposed by Sadinle (2017). We use the synthetic validation data sets analyzed in the original article, each of which contains 500 records with four linkage fields (first name, last name, age, and occupation). The degree of overlap is varied from 10%, 50%, to 100%. In addition, for each pair of data sets being merged, noise is added to either 1, 2, or 3 linkage fields. The author created 100 pairs of datasets per overlap and number of fields with noise combination, which lead to a total of 900 record linkage results (see the original paper for the details of this simulation study). Four agreement levels are used for the string valued fields (first and last name) based on a renormalized (between 0 and 1) version of Levenshtein edit distance with the following cutoffs {0, 0.25, 0.50}, where zero (one) means that two strings are identical (different). Finally, binary comparisons based on exact matching are applied to age and occupation.[2]

Sadinle (2017) and later McVeigh and Murray (2017), using these synthetic data sets, find that as the degree of overlap between datasets decreases, the performance of the Fellegi-Sunter model deteriorates. Moreover, both studies show that even when the overlap between datasets is equal to 50% the Fellegi-Sunter model performs poorly in terms of precision (share of true matches out of all the declared matches) and recall (share of recovered true matches out of all the true matches), see Figure 2 in Sadinle (2017) and Figure 4 in McVeigh and Murray (2017).

Figure S3 shows the results of our analysis when we apply fastLink followed by a one-to-one assignment restriction (blue boxplots labelled as fastLink), and compare them with the results of the bipartite graph model proposed by Sadinle (2017) (red boxplots labelled as BetaRL). Each row represents the degree of overlap, with the top row showing the results for the datasets with a 100% overlap, the middle row those with a 50% overlap, and the bottom row those with a small overlap (only 10% of the records). Note that within each row, we present the three levels of measurement error i.e., one, two, and three fields with a typographical error added (corresponding

[2] Both age and occupation are categorical variables with 8 mutually exclusive categories each. For more details see Sadinle (2017).

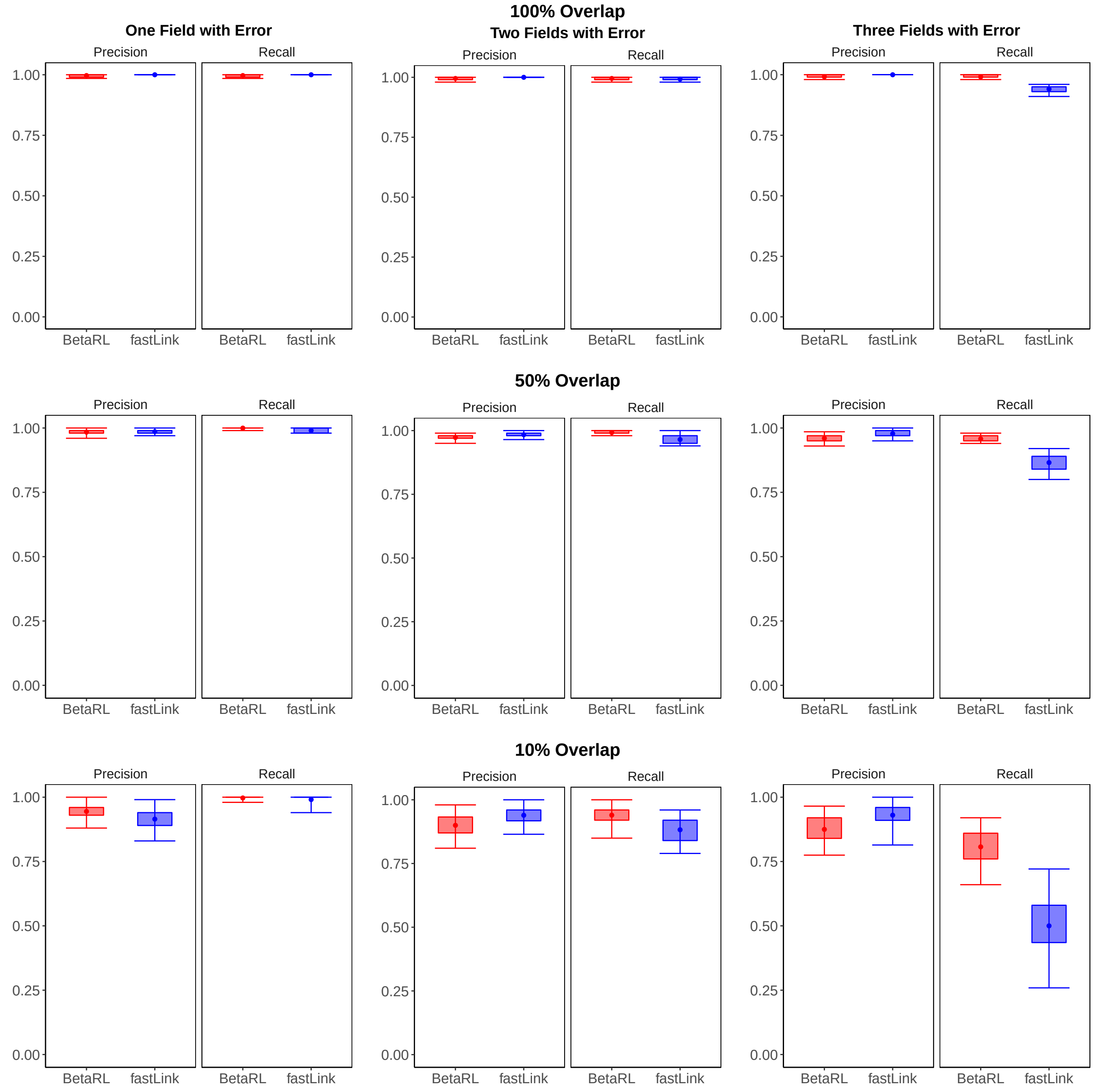

Figure S3: Precision and Recall across Different Levels of Overlap and Number of Fields with Error based on Sadinle (2017)'s validation datasets. The red boxplot (labelled `BetaRL`) is the implementation of the bipartite graph model proposed by Sadinle (2017) and the blue boxplot is the `fastLink` implementation using Levenshtein string distance measures. Precision and recall rates are high when the overlap between datasets is large (50% and 100%).

to each column).

Sadinle (2017) and McVeigh and Murray (2017) find that for a 50% overlap, to keep a high level of recall one must sacrifice precision – they find a precision rate of around 70% for a recall close to 100%. In contrast, we find that when the overlap is larger or equal to 50% the precision and recall of fastLink are extremely high (both close to 100%). The only exception is when there are three errors per linkage field and the overlap is 50% or less (right-hand plot in the second row). While fastLink performs as well as BetaRL in terms of precision, recall is worse for fastLink compared to BetaRL. While Sadinle (2017) and McVeigh and Murray (2017) find that the performance of the Fellegi-Sunter model worsens dramatically when the amount of overlap is 10% regardless of the amount of measurement error, we find that fastLink does not perform well only when the amount of measurement error is large (i.e., 3 out of 4 fields with measurement error).

Why can fastLink dramatically improve the performance of the Fellegi-Sunter model compared to the simulation results obtained by Sadinle (2017) and McVeigh and Murray (2017)? A key explanation is that throughout estimation fastLink tries to enforce the inequality constraints among the different values of $\boldsymbol{\pi}_{kml}$ as described in Section B. We use a rejection sampling to obtain new starting values for the parameter estimates that violate the inequality constraints by sampling from a Dirichlet distribution until they satisfy those constraints. For example, for binary comparisons, the inequality constraints are given by $\pi_{k10} < \pi_{k11}$ and $\pi_{k01} < \pi_{k00}$ for each k. Suppose that for some k', we have $\pi_{k'11} < \pi_{k'10}$. Then, we sample (π_{k10}, π_{k11}) from Dirichlet(α_0, α_1) with $\alpha_0 < \alpha_1$ (e.g., $\alpha_0 = 1$, $\alpha_1 = 5$) until we obtain $\pi_{k'10} \ll \pi_{k'11}$. This helps to prevent the EM algorithm from converging to a local maximum. Finally, we note that for fastLink, moving from the Levenshtein to Jaro-Winkler improves precision and recall in cases where the amount of overlap is small and the data is noisy.

In sum, the analysis of this appendix shows that the scalability of fastLink is much greater than the state-of-the-art probabilistic record linkage methods while their accuracy is about the same.

S9 The Details of Simulation Setups

In this section, we describe the details of simulation setups for a total of 270 simulation studies we conducted. Let $\mathcal{A}$ and $\mathcal{B}$ denote the smaller and larger dataset of two data sets to be merged, respectively. We first create the larger data set $\mathcal{B}$ by randomly selecting 100,000 records out of our pool of 341,160 voters, i.e., $N_{\mathcal{B}} = 100,000$.

- **Size balance:** We consider three size balances by setting $N_{\mathcal{A}} \in \{1000, 10000, 100000\}$. This creates the size balance of 1:100, 1:10, and 1:1.

- **Degree of overlap**: We consider the 20%, 50%, and 80% overlap as a fraction of the smaller data set $\mathcal{A}$, i.e., $\rho = 0.2, 0.5, 0.8$. We first obtain $\rho N_{\mathcal{A}}$ records at random from the data set $\mathcal{B}$ and then select $(1-\rho)N_{\mathcal{A}}$ observations at random from the pool of 241,160 observations that are not part of the data set $\mathcal{B}$

Once we have these datasets with different levels of overlap and size, we then create two types of simulations:

- Different **missing data** mechanisms across multiple levels of severity in the amount of missing values.

- Different **measurement error** structures across multiple levels of severity in the amount of typographical errors.

We now describe each of these simulation studies.

S9.1 Simulations with Missing Data

We consider five missing data mechanisms and three missing data proportions, 5%, 10%, and 15%, i.e., $\omega_m \in \{0.05, 0.1, 0.15\}$. We introduce missing data into the following five variables, first name, middle initial, last name, house number, and street name. The missing data mechanisms are:

1. **Missing completely at random (MCAR):** For each of the aforementioned five linkage variables, we independently and randomly select $\omega_m N_{\mathcal{A}}$ and $\omega_m N_{\mathcal{B}}$ observations for the data sets $\mathcal{A}$ and $\mathcal{B}$, respectively, and recode their values as missing.

2. **Missing at Random (MAR) with independence across the linkage variables:** We make the missing probability of each variable dependent on year of birth. For each of the linkage variables in the first data set $\mathcal{A}$, we first compute the quantiles of the year of birth variable. Among the observations whose quantile is greater than or equal to 0.2, we shuffle these quantile values. In contrast, those records whose quantile values are less than 0.2, we leave them unchanged. This induces a moderate amount of correlation between age and the probability of missing. Finally, we set the probability of missing to be proportional to the resulting quantile. Using these probabilities, we independently and randomly select $\omega_m N_{\mathcal{A}}$ observations to have missing values for each variable. We repeat the same procedure for the second data set, $\mathcal{B}$.

3. **Missing not at Random (MNAR) with independence across the linkage variables:** The only difference between MNAR and MAR is that once we shuffle the quantiles of year of birth we multiply each quantile by 0.3 for all individuals that had voted in the 2004 Presidential election. This makes those who voted less likely to have missing values.

4. **MAR and MNAR with dependence across the linkage variables:** The only difference between these settings and their independence counterpart is that for each linkage field (other than year of birth), we use the missing probabilities that are proportional to the same set of quantiles (without shuffling them). This makes the same group of observations likely to have missing values in multiple covariates.

In addition, for the following variables: first name, last name, and street name, we added three types of typographical errors. They are transpositions, deletions, and insertions. For transpositions, we swap the position of randomly selected two consecutive characters, e.g., John → Jonh. For deletions, we remove randomly selected one character, e.g., John → Jon. For replacements, we replace a randomly chosen character with another randomly selected but different character, e.g., John → Jobn. For each of the aforementioned linkage variables, we randomly select 6% of their observations, and then split the chosen observations in three groups of equal size so that each group is subject to one of the three aforementioned typographical errors. Thus, under these simulations, we vary the degree and structure of the missing values while keeping those of measurement error constant.

S9.2 Simulations with Measurement Error

We use a correlation structure similar to the one used for the missing data to construct five measurement error data structures with three different levels, 20%, 50%, and 80%, i.e., $\omega_e \in \{0.20, 0.50, 0.80\}$. Out of the six variables we use to merge data sets, the following three string-valued variables are the subject of measurement error: first name, last name, and street name. We add three types of typographical errors – transpositions, deletions, and replacements – as described above in Section I.1. Each type of typographical error occurs at an equal rate i.e., $\omega_e/3$. The measurement error structures are:

1. **Classical measurement error:** For each of the aforementioned three linkage variables, we independently and randomly select $\omega_e N_{\mathcal{A}}$ and $\omega_e N_{\mathcal{B}}$ observations for the data sets $\mathcal{A}$ and $\mathcal{B}$, respectively.

2. **Non-differential measurement error with independence across the three linkage variables:** Similar to the missing data case, we make the measurement error probability of each variable dependent on year of birth. Again, for each dataset, we first compute the quantiles of the year of birth variable. Among the observations whose quantile is greater than or equal to 0.2, we shuffle these quantile values. In contrast, those records whose quantile values are less than 0.2, we leave them unchanged. We set the probability of measurement error to be proportional to the resulting quantile. Using these probabilities, we independently and randomly select $\omega_e N_{\mathcal{A}}$ and $\omega_e N_{\mathcal{B}}$ observations to be perturbed for each variable.

3. **Differential measurement error with independence across the linkage variables:** The only difference between non-differential and differential measurement error is that once we shuffle the quantiles of year of birth we multiply each quantile by 0.3 for all individuals that had voted in the 2004 Presidential election. This makes those who voted less likely to have measurement error.

4. **Non-differential and differential measurement error with dependence across the linkage variables:** The only difference between these settings and their independence counterparts is that for each linkage field (other than year of birth), we use the measurement error probabilities that are proportional to the same set of quantiles (without shuffling them so that the propensity of measurement error is correlated across linkage variables). This makes the same group of observations likely to have measurement error in the three string-valued covariates.

Throughout these simulations, we use MCAR with the missingness rate of 6% while varying the structure and amount of measurement error as specified above.

Figure S4 presents the results regarding the accuracy of FDR and FNR estimates in the top and bottom panels, respectively, for merging two data sets of 100,000 records each. We compare

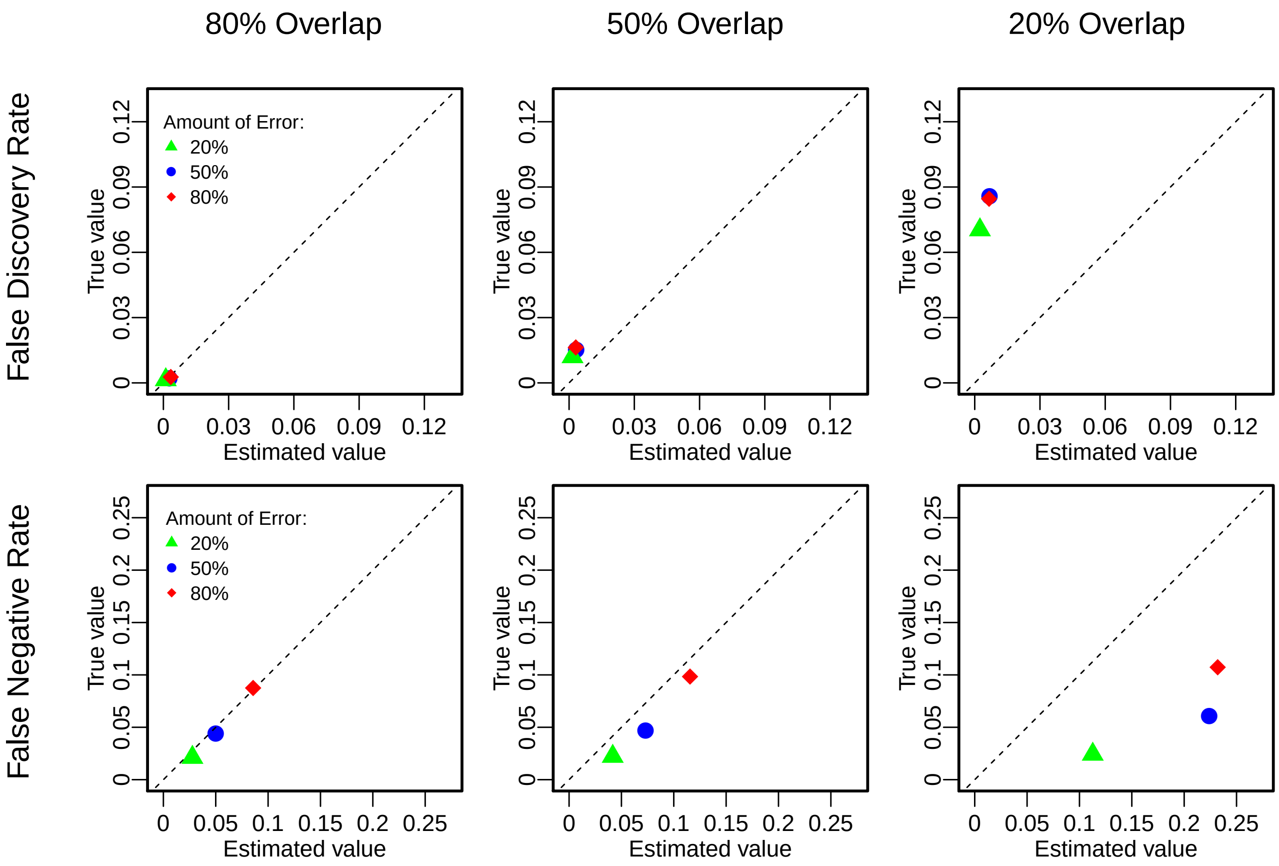

Figure S4: Magnitude of FDR and FNR and the Accuracy of their Estimates under Dependent and Non-differential Measurement Error. The top panel compares the estimated FDR (x-axis) with its true value (y-axis) whereas the bottom panel compares the estimated FNR against its true value. We consider two datasets of equal size (100,000 records each), non-differential measurement error with dependence across string-valued linkage fields (first name, last name, and street name) as the error structure, and three levels of measurement error: 20% (green triangles), 50% (blue circles), 80% (red diamonds). In addition, we added a 6% share of missing values drawn completely at random in all linkage fields other than year of birth. The FDR and FNR estimates are accurate when the overlap is sufficiently large i.e., larger or equal to 50%.

the performance of the proposed methodology across different amounts of measurement error per string-valued linkage field (first name, last name, and street name): 20% (green triangles), 50% (blue circles), and 80% (red diamonds). The results are based on the simulation setting with dependent and non-differential measurement error. This allows us to compare the results in Figure S4 with those presented in Figure S2, which are obtained without the conditional independence assumption. The columns of both panels represent different degrees of overlap between the data sets being merged.

We find that as the amount of overlap between data sets increases, the level of precision in terms of both FDR and FNR increases and their estimates become more accurate. However, when the amount of overlap is small (10 %), the performance is severely affected. For example, while the model estimates the FDR to be less than 1% (regardless of the amount of measurement error), the true FDR is greater than 6%. This result is consistent with the results presented in Section 3.

Due to space constraints, we present the complete set of results for all the 270 simulations in a separate simulation appendix. Overall, we find that the Fellegi-Sunter model under the conditional independence assumption does not yield reliable estimates of the FDR and FNR when the amount of overlap between data sets is small, regardless of the amount of missing data and measurement error. However, when the conditional independence assumption is relaxed as done in Appendix D, the Fellegi-Sunter model yields more precise estimates as shown in Figure S2. These patterns are consistently found across simulations even when two datasets have unequal sizes.

S10 Additional Empirical Results with Different Thresholds

S10.1 Using 0.75 as the Threshold

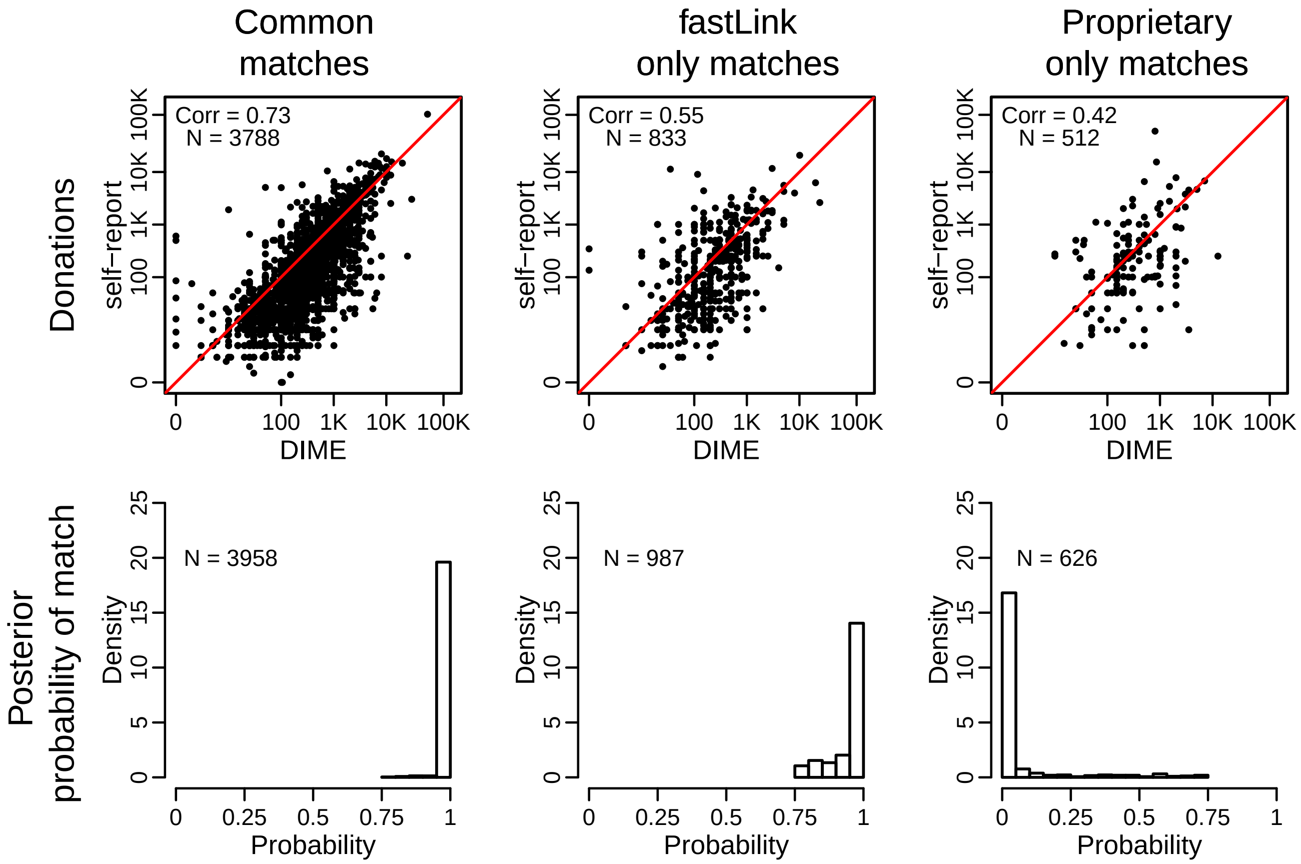

Figure S5: Comparison of fastLink and the Proprietary Method with the Threshold of 0.75. The top panel compares the self-reported donations (y-axis) by matched CCES respondents with their donation amount recorded in the DIME data (x-axis) for the three different groups of observations: those declared as matches by both fastLink and the proprietary method (left), those identified by fastLink only (middle), and those matched by the proprietary method only (right). The bottom panel presents histograms for the match probability for each group. For fastLink, we use one-to-one match with the threshold of 0.75.

S10.2 Using 0.95 as the Threshold

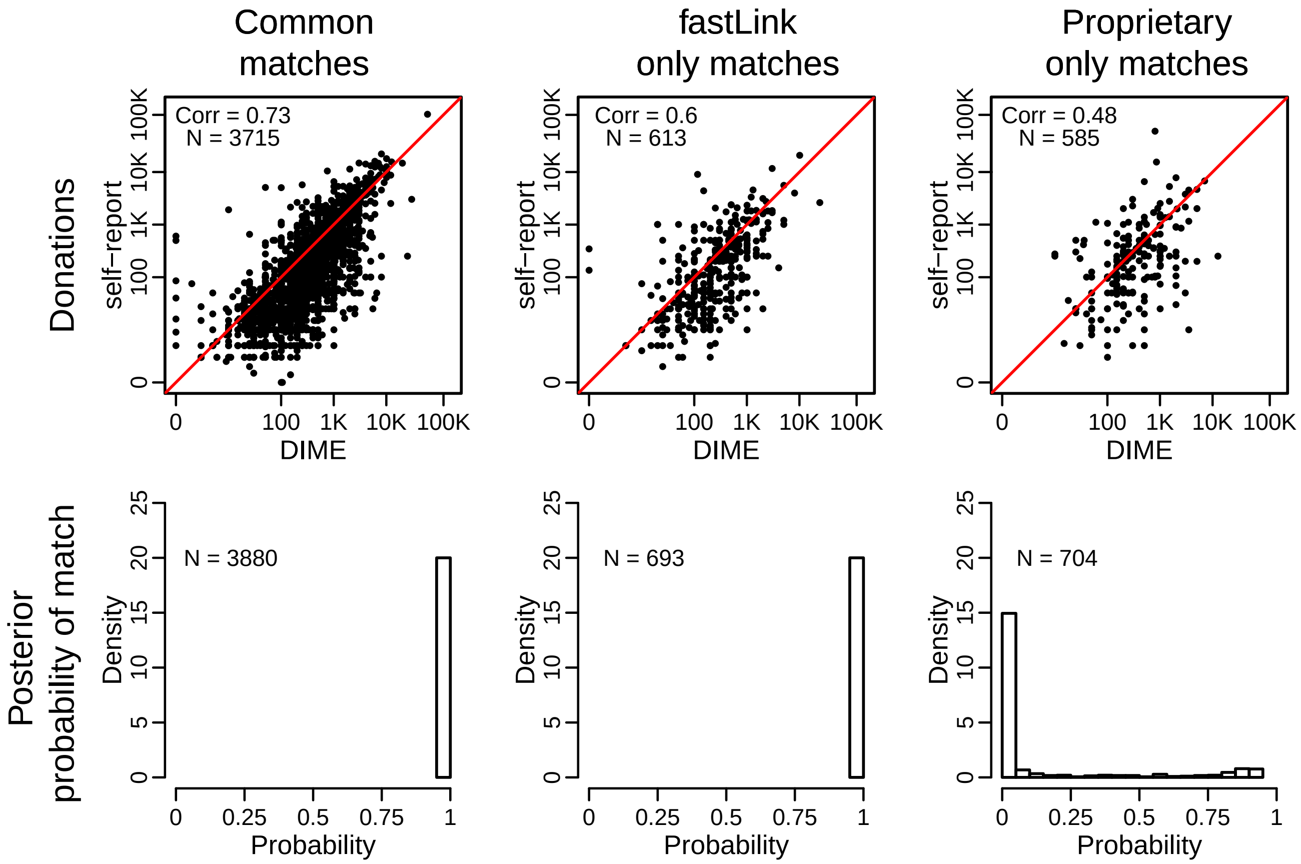

Figure S6: Comparison of fastLink and the Proprietary Method with the Threshold of 0.95. The top panel compares the self-reported donations (y-axis) by matched CCES respondents with their donation amount recorded in the DIME data (x-axis) for the three different groups of observations: those declared as matches by both fastLink and the proprietary method (left), those identified by fastLink only (middle), and those matched by the proprietary method only (right). The bottom panel presents histograms for the match probability for each group. For fastLink, we use one-to-one match with the threshold of 0.95.

S10.3 One-to-many Match: Weighted Analysis

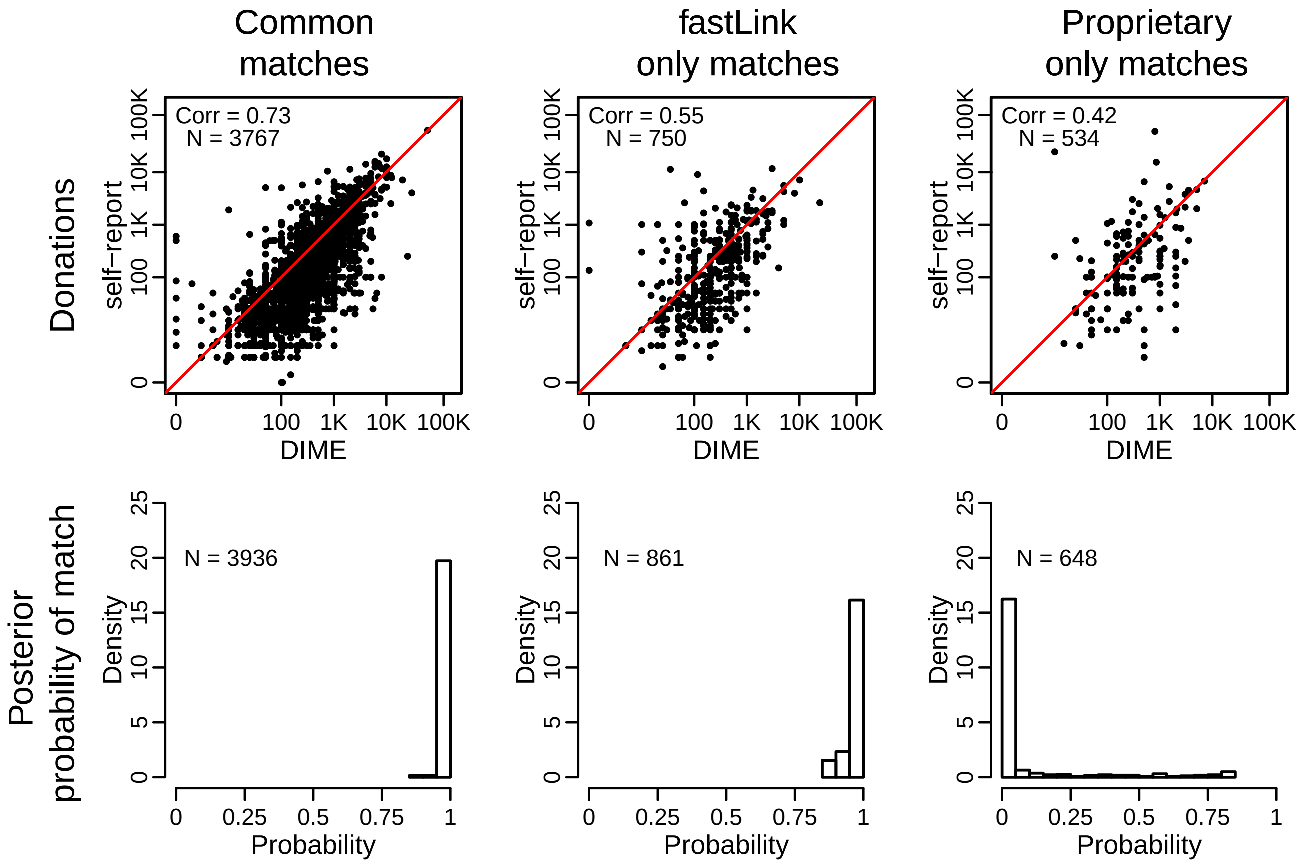

Figure S7: **Comparison of fastLink and the Proprietary Method with the Threshold of 0.85 using a Weighted Average Analysis for a One-to-many merge. The top panel compares the self-reported donations (y-axis) by matched CCES respondents with their donation amount recorded in the DIME data (x-axis) for the three different groups of observations: those declared as matches by both fastLink and the proprietary method (left), those identified by fastLink only (middle), and those matched by the proprietary method only (right). The bottom panel presents histograms for the match probability for each group. For fastLink, if a CCES respondent was matched to many DIME observations, we average all the matched contributions weighting each case by their corresponding probability of being a match.**

S11 Simulation Results for Random Sampling

Figure S8 shows that fitting fastLink to merely 5% of the sample will yield parameter estimates essentially identical to those based on the full sample.

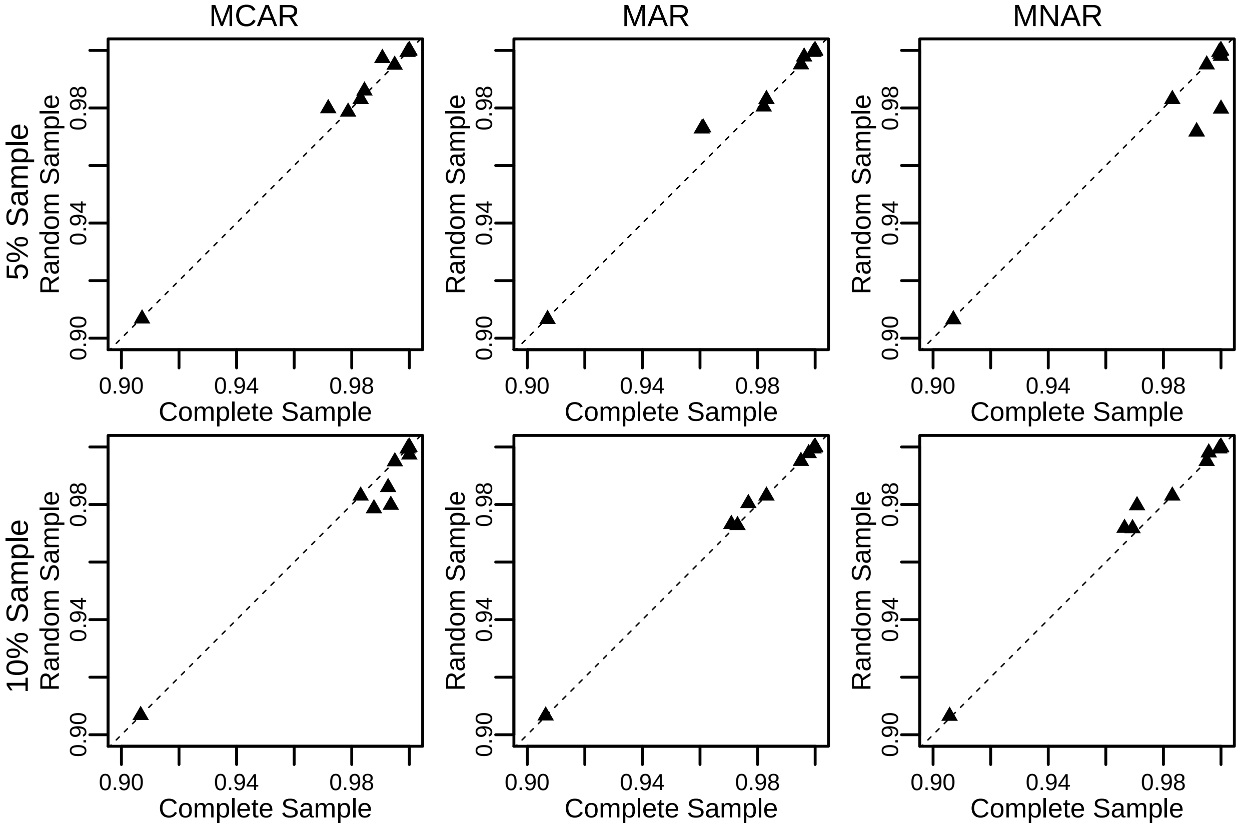

Figure S8: Parameter Estimates from Random Samples Compared Against the Parameter Estimates from the Full Dataset. The top panel compares parameter estimates from running fastLink on a full simulated dataset of size 100,000 (x-axis) against a 5% random sample from that same dataset (y-axis) under three different missing data mechanisms. The bottom panel compares parameter estimates from running fastLink on the same full simulated dataset against a 10% random sample from that same dataset. For all exercises except for the 5% random sample under MNAR, the parameter estimates from the random sample approximate the full-sample parameter estimates very closely.

S12 Simulation Results for k-means as a Blocking Method

As noted in Section 3, in our simulations the year of birth is observed for all records. Thus, we can use clustering methods such as k-means to group similar observations. One advantage of making group level comparison (blocking) is that the amount of overlap increases. In this simulation setting with four clusters, the ratio of true matches to all possible pairs is approximately 6×10^{-6}. This is more than three times as large as the corresponding ratio for no blocking and is comparable to the case of overlap of 50%.

We apply fastLink to each block separately. As shown in the right most column of Figure S9, blocking significantly improves the accuracy for the FDR and FNR estimates as well as their true values although the bias is not completely eliminated. The potential of k-means as a blocking strategy is something we leave for further research; the point here is to show that blocking can help to improve the precision of our estimates in cases where the amount of overlap between datasets is small.

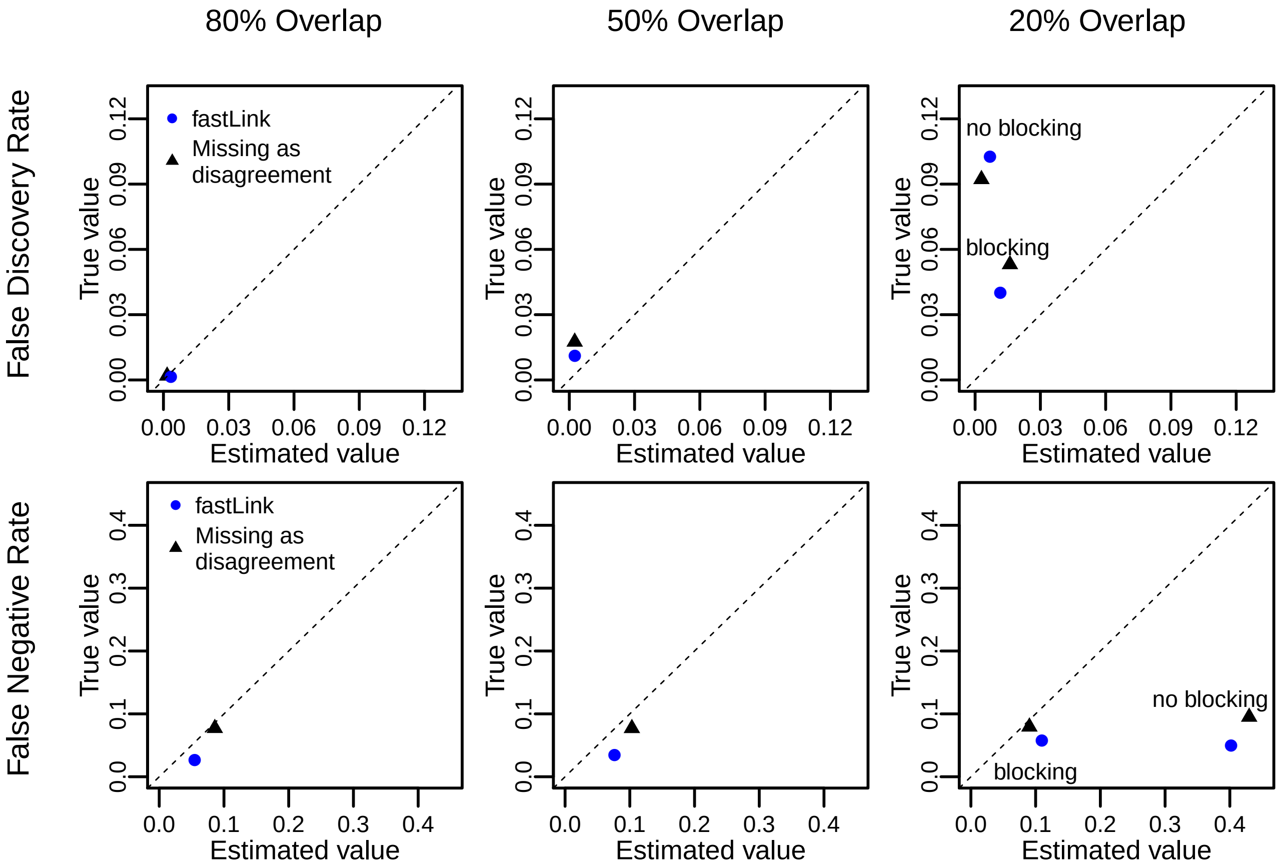

Figure S9: Accuracy of FDR and FNR Estimates. The top panel compares the estimated FDR (x-axis) with its true value (y-axis) whereas the bottom panel compares the estimated FNR against its true value. We consider the medium amount of missing data generated under MAR as a missingness mechanism and add measurement error to some linkage fields. The blue solid circles represent the estimates based on fastLink whereas the black solid triangles represent the estimates obtained by treating missing data as disagreements. The FDR and FNR estimates are accurate when the overlap is high. In addition, fastLink gives lower FDR and FNR than the same algorithm that treats missing values as a disagreement. Note that in cases where the overlap is small (20%), blocking via k-means improves the precision of our estimates.

S13 Quality of Matches for Merging Election Survey Data with Political Contribution Data

To further examine the quality of the matches, Table S2 presents the five most frequent agreement patterns separately for the matches identified by both methods, fastLink only, and the proprietary method only. Most agreement patterns identified by both methods have identical (or nearly identical) values in almost all linkage variables, yielding match probabilities close to 1. In contrast, the matches identified by the proprietary method alone has several fields of disagreements and as a result fastLink assigns those patterns low match probabilities. We note that this does not necessarily invalidate the proprietary method because their matches may be based on other information to which we do not have access. In addition, since the contribution data are compiled by donors themselves, the quality of matches for fastLink can be further improved if we code common nicknames as similar to their corresponding first names. Nevertheless, the overall results indicate that fastLink produces matches whose quality is better or at least as good as the proprietary method.

Name			Address				
First	Middle	Last	House	Street	Zip	Counts	Prob.
Common matches							
identical	NA	identical	identical	identical	identical	1356	1.00
identical	identical	identical	identical	different	identical	1324	1.00
identical	identical	identical	identical	identical	identical	266	1.00
identical	identical	identical	identical	different	identical	208	1.00
identical	NA	identical	NA	NA	identical	118	0.99
fastLink only matches							
identical	NA	identical	identical	different	identical	112	0.99
different	NA	different	identical	identical	identical	109	0.95
identical	NA	identical	NA	NA	identical	98	0.99
identical	NA	identical	identical	identical	identical	68	1.00
different	NA	identical	different	identical	identical	55	0.98
Proprietary method only matches							
identical	NA	identical	different	different	identical	27	0.59
different	NA	identical	NA	NA	identical	19	0.04
different	NA	different	different	identical	identical	19	0.01
different	NA	different	identical	different	identical	14	0.01
identical	NA	different	NA	NA	identical	13	0.01

Table S2: Five Most Frequent Agreement Patterns for Matches Identified by Both Methods, fastLink Only, and the Proprietary Method Only. For a given agreement pattern, the number of matches and the match probability (according to fastLink) are presented in the "Counts" and "Prob." columns, respectively. For fastLink we use one-to-one match with the threshold of 0.85.

S14 Comparing Age-Windowing Blocking to k-Means Blocking for Merging Two Voter Files

In Section 4.2, we merge two nationwide voter files from 2014 and 2015 to track movers within and across the United States. An important part of the workflow for conducting this merge is the choice of blocking strategy, where within each state-gender pair we run the k-means algorithm on first name to create blocks of around 250,000 voters each.

Here, we compare the k-means blocking strategy to a commonly used blocking strategy known as age-windowing. In age windowing, a block is defined by the set of observations whose ages are all within plus or minus some range of years, frequently chosen to be a year. In the exercise here, we have repeated the within-state merge for Florida blocking on gender and using an age window of ±1 year. This means that the match is first run for all women in Florida aged 18-20, then all women aged 19-21, and proceeding upwards by 1 year at a time. The process is then repeated for all men in Florida.

The results, shown in Table S3, suggest that the two blocking strategies yield very similar results. The match count, match rate, and FDR across different match probability thresholds are all extremely close for k-means and age windowing, and while FNR is consistently higher for k-means versus age windowing, it never goes above 0.3% in any of the thresholds.

	k-Means			Age Windows			
	75%	85%	95%	75%	85%	95%	Exact
Match count (millions)	9.76	9.75	9.75	9.73	9.73	9.73	7.31
Match Rate (%)	92.45%	92.39%	92.37%	92.21%	92.21%	92.21%	69.23%
FDR (%)	0.02%	0%	0%	0.01%	0.01%	0.01%	
FNR (%)	0.19%	0.26%	0.28%	0.01%	0.01%	0.02%	

Table S3: Comparison of blocking strategies when merging the 2014 Florida voter file to the 2015 Florida voter file. This table compares the match count, match rate, False Discovery Rate (FDR), and False Negative Rate (FNR) for a merge of the 2014 Florida voter file to the 2015 Florida voter file when using k-means blocking on first name versus age windowing blocking. Each column denotes a different threshold for declaring a match, from most lenient to strictest, while the final column gives the match count and match rate under an exact match.

References

Cohen, W. W., P. Ravikumar and S. Fienberg. 2003. "A Comparison of String Distance Metrics for Name-Matching Tasks." In International Joint Conference on Artificial Intelligence (IJCAI) 18.

Dempster, Arthur P., Nan M. Laird and Donald B. Rubin. 1977. "Maximum Likelihood from Incomplete Data Via the EM Algorithm (with Discussion)." *Journal of the Royal Statistical Society, Series B, Methodological* 39:1–37.

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

Jaro, Matthew. 1989. "Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida." *Journal of the American Statistical Association.* 84:414–420.

Larsen, Michael D. and Donald B. Rubin. 2001. "Iterative Automated Record Linkage Using Mixture Models." *Journal of the American Statistical Association* 96:32–41.

McVeigh, Brendan S. and Jared S. Murray. 2017. Practical Bayesian Inference for Record Linkage. Technical Report. Carnegie Mellon University.

Murray, Jared S. 2016. "Probabilistic Record Linkage and Deduplication after Indexing, Blocking, and Filtering." *Journal of Privacy and Confidentiality* 7:3–24.

Newcombe, H. B. and J. M. Kennedy. 1962. "Record Linkage: Making Maximum Use of the Discriminating Power of Identifying Information." *Communications of Association for Computing Machinery* 5:563–67.

Newcombe, H. B., J. M. Kennedy, S. J. Axford and A. P. James. 1959. "Automatic Linkage of Vital Records." *Science.* 130:954–959.

Sadinle, Mauricio. 2017. "Bayesian Estimation of Bipartite Matchings for Record Linkage." *Journal of the American Statistical Association.*

Sadinle, Mauricio and Stephen Fienberg. 2013. "A Generalized Fellegi-Sunter Framework for Multiple Record Linkage With Application to Homicide Record Systems." *Journal of the American Statistical Association.* 108:385–397.

Steorts, Rebecca C. 2015. "Entity Resolution with Empirically Motivated Priors." *Bayesian Analysis.* 10:849–875.

Steorts, Rebecca C., Rob Hall and Stephen E. Fienberg. 2016. "A Bayesian Approach to Graphical Record Linkage and Deduplication." *Journal of the American Statistical Association* 111:1660–1672.

Thibaudeau, Yves. 1993. “The Discrimination Power of Dependency Structures in Record Linkage.” *Survey Methodology.*

Winkler, William E. 1988. Using the EM Algorithm for Weight Computation in the FellegiSunter Model of Record Linkage. In *Proceedings of the Section on Survey Research Methods, American Statistical Association.* pp. 667–671.

Winkler, William E. 1989. Near Automatic Weight Computation in the Fellegi-Sunter Model of Record Linkage. Technical Report. Proceedings of the Census Bureau Annual Research Conference.
URL: *https://www.researchgate.net/publication/243778219_Near_Automatic_Weight_Computation_in_the_Fellegi-Sunter_Model_of_Record_Linkage*

Winkler, William E. 1990. “String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage.” Proceedings of the Section on Survey Research Methods. American Statistical Association.
URL: *https://www.iser.essex.ac.uk/research/publications/501361*

Winkler, William E. 1993. “Improved Decision Rules in the Fellegi-Sunter Model of Record Linkage.” In Proceedings of Survey Research Methods Section, American Statistical Association.
URL: *http://ww2.amstat.org/sections/srms/Proceedings/papers/1993_042.pdf*

Winkler, William E. 1994. Advanced Methods for Record Linkage. Technical Report. Proceedings of the Section on Survey Research Methods, American Statistical Association.

Winkler, William E. 2000. Using the EM Algorithm for Weight Computation in the Felligi-Sunter Model of Record Linkage. Technical Report No. RR2000/05. Statistical Research Division, Methodology and Standards Directorate, U.S. Bureau of the Census.

Yancey, Willian. 2000. Frequency-Dependent Probability Measures for Record Linkage. In *Proceedings of the Secion on Survet Research Methods.* American Statistical Association pp. 752–757.

Yancey, Willian. 2005. “Evaluating String Comparator Performance for Record Linkage.” Research Report Series. Statistical Research Division U.S. Census Bureau.