Using a Probabilistic Model to Assist Merging of Large-scale Administrative Records*

Ted Enamorado† Benjamin Fifield‡ Kosuke Imai§

Forthcoming in *American Political Science Review*

Abstract

Since most social science research relies upon multiple data sources, merging data sets is an essential part of researchers' workflow. Unfortunately, a unique identifier that unambiguously links records is often unavailable, and data may contain missing and inaccurate information. These problems are severe especially when merging large-scale administrative records. We develop a fast and scalable algorithm to implement a canonical probabilistic model of record linkage that has many advantages over deterministic methods frequently used by social scientists. The proposed methodology efficiently handles millions of observations while accounting for missing data and measurement error, incorporating auxiliary information, and adjusting for uncertainty about merging in post-merge analyses. We conduct comprehensive simulation studies to evaluate the performance of our algorithm in realistic scenarios. We also apply our methodology to merging campaign contribution records, survey data, and nationwide voter files. An open-source software package is available for implementing the proposed methodology.

Key Words: EM algorithm, false discovery rate, false negative rate, missing data, mixture model, record linkage

*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`). We thank Bruce Willsie of L2 and Steffen Weiss of YouGov for data and technical assistance, Jake Bowers, Seth Hill, Johan Lim, Marc Ratkovic, Mauricio Sadinle, five anonymous reviewers, and audiences at the 2017 Annual Meeting of the American Political Science Association, Columbia University (Political Science), Fifth Asian Political Methodology Meeting, Gakusyuin University (Law), Hong Kong University of Science and Technology, the Institute for Quantitative Social Science (IQSS) at Harvard University, the Quantitative Social Science (QSS) colloquium at Princeton University, Universidad de Chile (Economics), Universidad del Desarrollo, Chile (Government), the 2017 Summer Meeting of the Society for Political Methodology, the Center for Statistics and the Social Sciences (CSSS) at the University of Washington for useful comments and suggestions.

†Ph.D. Candidate, Department of Politics, Princeton University, Princeton NJ 08544. Email: `tede@princeton.edu`, URL: `http://www.tedenamorado.com`

‡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 Introduction

As the amount and diversity of available data sets rapidly increase, social scientists often harness multiple data sources to answer substantive questions. Indeed, merging data sets, in particular large-scale administrative records, is an essential part of cutting-edge empirical research in many disciplines (e.g., Jutte, Roos and Browne, 2011; Ansolabehere and Hersh, 2012; Einav and Levin, 2014). Data merging can be consequential. For example, the American National Election Studies (ANES) and Cooperative Congressional Election Study (CCES) validate self-reported turnout by merging their survey data with a nationwide voter file where only the matched respondents are treated as registered voters. While Ansolabehere and Hersh (2012) advocate the use of such a validation procedure, Berent, Krosnick and Lupia (2016) argue that the discrepancy between self-reported and validated turnout is due to the failure of the merge procedure rather than social desirability and non-response bias.

Merging data sets is straightforward if there exists a unique identifier that unambiguously links records from different data sets. Unfortunately, such a unique identifier is often unavailable. Under these circumstances, some researchers have used a deterministic algorithm to automate the merge process (e.g., Bolsen, Ferraro and Miranda, 2014; Figlio and Guryan, 2014; Meredith and Morse, 2014; Adena et al., 2015; Giraud-Carrier et al., 2015; Ansolabehere and Hersh, 2017; Berent, Krosnick and Lupia, 2016; Cesarini et al., 2016; Hill, 2017) while others have relied upon a proprietary algorithm (e.g., Ansolabehere and Hersh, 2012; Richman, Chattha and Earnest, 2014; Figlio and Guryan, 2014; Hill and Huber, 2017; Hersh, 2015; Engbom and Moser, 2017). However, these methods are not robust to measurement error (e.g., misspelling) and missing data, which are common to social science data. Furthermore, deterministic merge methods cannot quantify the uncertainty of the merging procedure and instead typically rely on arbitrary thresholds to determine the degree of similarity sufficient for matches.[1] This means that post-merge data analyses fail to account for the uncertainty of the merging procedure, yielding a bias due to measurement error. These methodological challenges are amplified especially when merging large-scale administrative records.

[1]These thresholds are highly dependent on data. For example, Ansolabehere and Hersh (2017) find that using 3 fields with exact matches as the threshold works well for the Texas voter file, but the same threshold may not work for other data. In contrast, probabilistic methods can automatically weight observations.

We demonstrate that social scientists should use probabilistic models rather than deterministic methods when merging large data sets. Probabilistic models can quantify the uncertainty inherent in many merge procedures, offering a principled way to calibrate and account for false positives and false negatives. Unfortunately, while there exists a well-known statistics literature on probabilistic record linkage (e.g., Winkler, 2006b; Herzog, Scheuren and Winkler, 2007; Harron, Goldstein and Dibben, 2015), the current open-source implementation does not scale to large data sets commonly used in today's social science research. We address this challenge by developing a fast and scalable implementation of the canonical probabilistic record linkage model originally proposed by Fellegi and Sunter (1969). Together with parallelization, this algorithm, which we call fastLink, can be used to merge data sets with millions of records in a reasonable amount of time using one's laptop computer. Additionally, building on the prior methodological literature (e.g., Lahiri and Larsen, 2005), we show (1) how to incorporate auxiliary information such as population name frequency and migration rates into the merge procedure and (2) how to conduct post-merge analyses while accounting for the uncertainty about the merge process. We describe these methodological developments in Section 2.

In Section 3, we conduct comprehensive simulation studies to evaluate the robustness of fastLink to several factors including the size of data sets, the proportion of true matches, measurement error, and missing data proportion and mechanisms. A total of 270 simulation settings consistently show that fastLink significantly outperforms the deterministic methods. While the proposed methodology produces high quality matches in most situations, the lack of overlap between two data sets often leads to large error rates, suggesting that effective blocking is essential when the expected number of matches is relatively small. Furthermore, fastLink appears to perform at least as well as recently proposed probabilistic approaches (Steorts, 2015; Sadinle, 2017). Importantly, our merge method is faster and scales to larger data sets than these state-of-art methods.

In Section 4, we present two empirical applications. First, we revisit Hill and Huber (2017) who examine the ideological differences between donors and non-donors by merging the CCES data of more than 50,000 survey respondents, with the a campaign contribution database of over 5 million donor records (Bonica, 2013). We find that the matches identified by fastLink are at least as high-quality as those identified by the proprietary method, which was used by the original authors. We also improve the original analysis by incorporating the uncertainty of the merge process in

the post-merge analysis. We show that although the overall conclusion remains unchanged, the magnitude of the estimated effects are substantially smaller.

As the second application, we merge two nationwide voter files of over 160 million voter records each, representing one of the largest data merges ever conducted in social science research.[2] By merging voter files over time, scholars can study the causes and consequences of partisan residential segregation (e.g., Tam Cho, Gimpel and Hui, 2013; Mummolo and Nall, 2016) and political analytics professionals can develop effective micro-targeting strategies (e.g., Hersh, 2015). We show how to incorporate available within-state and across-state migration rates in the merge process. Given the enormous size of the data sets, we propose a two-step procedure where we first conduct a within-state merge for each state followed by across-state merges for every pair of states. The proposed methodology is able to match about 95% of voters, which is about 30 percentage points greater than the exact matching method. Although it is more difficult to find across-state movers, we are able to find 20 times as many such voters than the existing matching method.

Finally, we give concluding remarks in Section 5. We provide an open-source R software package fastLink: Fast Probabilistic Record Linkage, which is freely available at the Comprehensive R Archive Network (CRAN; `https://CRAN.R-project.org/package=fastLink`)for implementing our methodology so that other researchers can effectively merge data sets in their own projects.

2 The Proposed Methodology

In this section, we first introduce the canonical probabilistic model of record linkage originally proposed by Fellegi and Sunter (1969). We describe several improvements we make to this model, including a fast and scalable implementation, the use of auxiliary information to inform parameter estimation, and the incorporation of uncertainty about the merge process in post-merge analyses.

2.1 The Setup

Suppose that we wish to merge two data sets, $\mathcal{A}$ and $\mathcal{B}$, which have sample sizes of $N_{\mathcal{A}}$ and $N_{\mathcal{B}}$, respectively. We use K variables, which are common to both data sets, to conduct the merge.

[2]While Hersh (2015) conducted a large-scale data merge, his relied upon a proprietary algorithm. Others such as Ansolabehere and Hersh (2017) and Tam Cho, Gimpel and Hui (2013) match datasets of several million voters each, but neither of these studies approaches the scale of our applications. Note that the US Census Bureau routinely conducts large-scale data merges for decennial census (Winkler, Yancey and Porter, 2010).

	Name				Address	
	First	Middle	Last	Date of birth	House	Street
Data set $\mathcal{A}$						
1	James	V	Smith	12-12-1927	780	Devereux St.
2	Robert	NA	Martines	01-15-1942	60	16th St.
Data set $\mathcal{B}$						
1	Michael	F	Martinez	02-03-1956	4	16th St.
2	James	D	Smithson	12-12-1927	780	Dvereuux St.
Agreement patterns						
$\mathcal{A}.1 - \mathcal{B}.1$	different	different	different	different	different	different
$\mathcal{A}.1 - \mathcal{B}.2$	identical	different	similar	identical	identical	similar
$\mathcal{A}.2 - \mathcal{B}.1$	different	NA	similar	different	different	different
$\mathcal{A}.2 - \mathcal{B}.2$	different	NA	different	different	different	different

Table 1: An Illustrative Example of Agreement Patterns. The top panel of the table shows two artificial data sets, $\mathcal{A}$ and $\mathcal{B}$, each of which has two records. The bottom panel shows the agreement patterns for all possible pairs of these records. For example, the second line of the agreement patterns compares the first record of the data set $\mathcal{A}$ with the second record of the data set $\mathcal{B}$. These two records have an identical information for first name, date of birth, and house number; similar information for last name and street name; and different information for middle name. A comparison involving at least one missing value is indicated by NA.

We consider all possible pair-wise comparisons between these two data sets. For each of these $N_{\mathcal{A}} \times N_{\mathcal{B}}$ distinct pairs, we define an agreement vector of length K, denoted by $\boldsymbol{\gamma}(i, j)$, whose kth element $\gamma_k(i, j)$ represents the discrete level of within-pair similarity for the kth variable between the ith observation of data set $\mathcal{A}$ and the jth observation of data set $\mathcal{B}$. Specifically, if we have a total of L_k similarity levels for the kth variable, then the corresponding element of the agreement vector can be defined as,

$$\gamma_k(i,j) \;=\; \begin{cases} 0 & \text{different} \\ 1 \\ \vdots & \text{similar} \\ L_k - 2 \\ L_k - 1 & \text{identical} \end{cases} \tag{1}$$

The proposed methodology allows for the existence of missing data. We define a missingness vector of length K, denoted by $\delta(i, j)$, for each pair (i, j) where its kth element $\delta_k(i, j)$ equals 1 if at least one record in the pair has a missing value in the kth variable and is equal to 0 otherwise.

Table 1 presents an illustrative example of agreement patterns based on two artificial data

sets, $\mathcal{A}$ and $\mathcal{B}$, each of which has two records. In this example, we consider three possible values of $\gamma_k(i,j)$ for first name, last name, and street name, i.e., $L_k = 3$ (`different`, `similar`, nearly `identical`), whereas a binary variable is used for the other fields, i.e., $L_k = 2$ (`different`, nearly `identical`). The former set of variables require a similarity measure and threshold values. 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).[3] Since the Jaro-Winkler distance is a continuous measure whose values range from 0 (different) to 1 (identical), we discretize it so that $\gamma_k(i,j)$ takes an integer value between 0 and $L_k - 1$ as defined in equation (1). Suppose that we use three levels (i.e., `different`, `similar`, and nearly `identical`) based on the threshold values of 0.88 and 0.94 as recommended by Winkler (1990). Then, when comparing the last names in Table 1, we find that, for example, `Smith` and `Smithson` are similar (a Jaro-Winkler distance of 0.88) whereas `Smith` and `Martinez` are different (a Jaro-Winkler distance of 0.55).[4]

The above setup implies a total of $N_{\mathcal{A}} \times N_{\mathcal{B}}$ comparisons for each of K fields. Thus, the number of comparisons grows quickly as the size of data sets increases. One solution is to use blocking and avoid comparisons that should not be made. For example, we may make comparisons within gender group only. While appealing due to computational efficiency gains, Winkler (2005) notes that blocking often involves ad hoc decisions by researchers and face difficulties when variables have missing values and measurement error. Here, we focus on the data merge within a block and refer interested readers to Christen (2012) and Steorts et al. (2014) for comprehensive reviews of blocking techniques.[5] We also note a related technique, called filtering, which has the potential to overcome the weaknesses of traditional blocking methods by discarding pairs that are unlikely to be matches when fitting a probabilistic model (Murray, 2016).

2.2 The Canonical Probabilistic Model of Record Linkage

2.2.1 The Model and Assumptions

We first describe the most commonly used probabilistic model of record linkage (Fellegi and Sunter, 1969). Let a latent mixing variable M_{ij} indicate whether a pair of records (the ith record in the

[3] Online Supplementary Information (SI) A describes how the Jaro-Winkler string distance is calculated.

[4] As shown in Section 3.3 and Appendix A, the discretization of the distance measure leads to substantial computational efficiency when making pairwise comparison for each linkage field.

[5] The parameters of record linkage models must be interpreted separately for each block (Murray, 2016).

data set $\mathcal{A}$ and the jth record in the data set $\mathcal{B}$) represents a match. The model has the following simple finite mixture structure (e.g., McLaughlan and Peel, 2000; Imai and Tingley, 2012),

$$\gamma_k(i,j) \mid M_{ij} = m \overset{\text{indep.}}{\sim} \text{Discrete}(\boldsymbol{\pi}_{km}) \tag{2}$$

$$M_{ij} \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\lambda) \tag{3}$$

where $\boldsymbol{\pi}_{km}$ is a vector of length L_k, containing the probability of each agreement level for the kth variable given that the pair is a match ($m = 1$) or a non-match ($m = 0$), and λ represents the probability of a match across all pairwise comparisons. Through the parameter $\boldsymbol{\pi}_{k0}$, the model allows for the possibility that two records can have identical values for some variables even when they do not represent a match.

This model is based on two key independence assumptions. First, the latent variable M_{ij} is assumed to be independently and identically distributed. Such an assumption is necessarily violated if, for example, each record in the data set $\mathcal{A}$ should be matched with no more than one record in the data set $\mathcal{B}$. In theory, this assumption can be relaxed (e.g., Sadinle, 2017) but doing so makes the estimation significantly more complex and reduces its scalability (see Online SI H). Later in the paper, we discuss how to impose such a constraint without sacrificing computational efficiency. Second, the conditional independence among linkage variables is assumed given the match status. Some studies find that the violation of this assumption leads to unsatisfactory performance (e.g., Thibaudeau, 1993; Belin and Rubin, 1995; Larsen and Rubin, 2001; Winkler and Yancey, 2006; Herzog, Scheuren and Winkler, 2010). In Online SI D, we show how to relax the conditional independence assumption while keeping our scalable implementation.

In the literature, researchers often treat missing data as disagreements, i.e., $\gamma_k(i,j) = 0$ if $\delta_k(i,j) = 1$ (e.g., Goldstein and Harron, 2015; Sariyar, Borg and Pommerening, 2012; Ong et al., 2014)). This procedure is problematic because a true match can contain missing values. Other imputation procedures also exist but none of them has a theoretical justification or appears to perform well in practice.[6] To address this problem, following Sadinle (2014, 2017), we assume that data are missing at random (MAR) conditional on the latent variable M_{ij},

$$\delta_k(i,j) \perp\!\!\!\perp \gamma_k(i,j) \mid M_{ij}$$

[6]For example, although Goldstein and Harron (2015) suggest the possibility of treating a comparison that involves a missing value as a separate agreement value, but Sariyar, Borg and Pommerening (2012) find that this approach does not outperform the standard method of treating missing values as disagreements.

for each $i = 1, 2, \ldots, N_{\mathcal{A}}$, $j = 1, 2, \ldots, N_{\mathcal{B}}$, and $k = 1, 2, \ldots, K$. Under this MAR assumption, we can simply ignore missing data. The observed-data likelihood function of the model defined in equations (2) and (3) is given by,

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

where $\pi_{km\ell}$ represents the ℓth element of probability vector $\boldsymbol{\pi}_{km}$, i.e., $\pi_{km\ell} = \Pr(\gamma_k(i,j) = \ell \mid M_{ij} = m)$. Since the direct maximization of the observed-data log-likelihood function is difficult, we estimate the model parameters using the EM algorithm (see Online SI B).

2.2.2 The Uncertainty of Merge Process

The advantage of probabilistic models is their ability to quantify the uncertainty inherent in merging. Once the model parameters are estimated, we can compute the match probability for each pair using Bayes rule,[7]

$$\begin{aligned} \xi_{ij} &= \Pr(M_{ij} = 1 \mid \boldsymbol{\delta}(i,j), \boldsymbol{\gamma}(i,j)) \\ &= \frac{\lambda \prod_{k=1}^{K} \left(\prod_{\ell=0}^{L_k-1} \pi_{k1\ell}^{\mathbf{1}\{\gamma_k(i,j)=\ell\}} \right)^{1-\delta_k(i,j)}}{\sum_{m=0}^{1} \lambda^m (1-\lambda)^{1-m} \prod_{k=1}^{K} \left(\prod_{\ell=0}^{L_k-1} \pi_{km\ell}^{\mathbf{1}\{\gamma_k(i,j)=\ell\}} \right)^{1-\delta_k(i,j)}} \end{aligned} \tag{4}$$

In Section 2.4, we show how to incorporate this match probability into post-merge regression analysis in order to account for the uncertainty of the merge process.

While in theory a post-merge analysis can use all pairs with non-zero match probabilities, it is often more convenient to determine a threshold S when creating a merged data set. Such an approach is useful especially when the data sets are large. Specifically, we call a pair (i, j) a match if the match probability ξ_{ij} exceeds S. There is a clear trade-off in the choice of this threshold value. A large value of S will ensure that most of the selected pairs are correct matches but may fail to identify many true matches. In contrast, if we lower S too much, we will select more pairs but many of them may be false matches. Therefore, it is important to quantify the degree of these matching errors in the merging process.

One advantage of probabilistic models over deterministic methods is that we can estimate the false discovery rate (FDR) and the false negative rate (FNR). The FDR represents the proportion

[7] This is known as the maximum a posteriori (MAP) estimate.

of false matches among the selected pairs whose matching probability is greater than or equal to the threshold. We estimate the FDR using our model parameters as follows,

$$\Pr(M_{ij} = 0 \mid \ \xi_{ij} \geq S) \ = \ \frac{\sum_{i=1}^{N_{\mathcal{A}}} \sum_{j=1}^{N_{\mathcal{B}}} \mathbf{1}\{\xi_{ij} \geq S\}(1 - \xi_{ij})}{\sum_{i=1}^{N_{\mathcal{A}}} \sum_{j=1}^{N_{\mathcal{B}}} \mathbf{1}\{\xi_{ij} \geq S\}}. \tag{5}$$

whereas the FNR, which represents the proportion of true matches that are not selected, is estimated as,

$$\Pr(M_{ij} = 1 \mid \xi_{ij} < S) \ = \ \frac{\sum_{i=1}^{N_{\mathcal{A}}} \sum_{j=1}^{N_{\mathcal{B}}} \xi_{ij} \mathbf{1}\{\xi_{ij} < S\}}{\lambda N_{\mathcal{A}} N_{\mathcal{B}}}. \tag{6}$$

Researchers typically select, at their own discretion, the value of S such that the FDR is sufficiently small. But, we also emphasize the FNR since a strict threshold can lead to many false negatives.[8] In our simulations and empirical studies, we find that the results are not particularly sensitive to the choice of threshold value, although in other applications, scholars found ex-post adjustments are necessary for obtaining good estimates of error rates (e.g., Winkler, 1993; Thibaudeau, 1993; Belin and Rubin, 1995; Larsen and Rubin, 2001; Winkler, 2006a; Murray, 2016).

In the merging process, for a given record in the data set $\mathcal{A}$, it is possible to find multiple records in the data set $\mathcal{B}$ that have high match probabilities. In some cases, multiple observations have an identical value of match probability, i.e., $\xi_{ij} = \xi_{ij'}$ with $j \neq j'$. Following the literature (e.g., Tancredi and Liseo, 2011; Sadinle, 2017; McVeigh and Murray, 2017), we recommend that researchers analyze all matched observations by weighting them according to the matching probability (see Section 2.4). If researchers wish to enforce a constraint that each record in one data set is only matched at most with one record in the other data set, they may follow a procedure described in Online SI E.

2.3 Incorporating Auxiliary Information

Another advantage of the probabilistic model introduced above is that we can incorporate auxiliary information in parameter estimation. This point has not been emphasized enough in the literature. Here, we briefly discuss two adjustments using auxiliary data — first, how to adjust for the fact that

[8] A more principled solution to the threshold S selection problem would require data for which the true matching status $M(i, j)$ is known – so that one can select the value of S to minimize the classification error. However, in record linkage problems, only in rare occasions labeled datasets exists. See Larsen and Rubin (2001), Feigenbaum (2016), and Enamorado (2018) for approaches that directly incorporate labeled data.

some names are more common than others, and second, how to incorporate aggregate information about migration. More details can be found in Online SI Section F.

Since some first names are more common than others, they may be more likely to be false matches. To adjust for this possibility without increasing the computational burden, we formalize the conditions under which the ex-post correction originally proposed by Winkler (2000) is well-suited for this purpose. Briefly, the probability of being a match will be up-weighted or down-weighted given the true frequencies of different first names (obtained, for instance, from Census data) or observed frequencies of each unique first name in the data (see Online SI F.3.1).

Furthermore, we may know *a priori* how many matches we should find in two data sets due to knowledge and data on over-time migration. For instance, the IRS publishes detailed information on migration in the United States from tax records (see `https://www.irs.gov/uac/soi-tax-stats-migration-data`). An estimate of the share of individuals who moved out of a state or who moved in-state can be easily reformulated as a prior on relevant parameters in the Fellegi-Sunter model and incorporated into parameter estimation (see Online SI F.3.2).

2.4 Post-merge Analysis

Finally, we discuss how to conduct a statistical analysis once merging is complete. One advantage of probabilistic models is that we can directly incorporate the uncertainty inherent to the merging process in the post-merge analysis. This is important because researchers often use the merged variable either as the outcome or as the explanatory variable in the post-merge analysis. For example, when the American National Election Survey (ANES) validates self-reported turnout by merging the survey data with a nationwide voter file, respondents who are unable to be merged are coded as non-registered voters. Given the uncertainty inherent to the merging process, it is possible that a merging algorithm fails to find some respondents in the voter file even though they are actually registered voters. Similarly, we may incorrectly merge survey respondents with other registered voters. These mismatches, if ignored, can adversely affect the properties of post-match analyses (e.g., Neter, Maynes and Ramanathan, 1965; Scheuren and Winkler, 1993).

Unfortunately, most of the record linkage literature has focused on the linkage process itself without considering how to conduct subsequent statistical analyses after merging data sets.[9] Here,

[9] An important exception includes a fully Bayesian approach outside of the Fellegi-Sunter framework, which we do not pursue here due to its limited scalability (see Tancredi and Liseo, 2011; Gutman, C. and M., 2013; Gutman

we build on a small literature about post-merge regression analysis, whose goal is to eliminate possible biases due to the linkage process within the Fellegi-Sunter framework (e.g., Scheuren and Winkler, 1993, 1997; Lahiri and Larsen, 2005; Kim and Chambers, 2012; Hof and Zwinderman, 2012). We also clarify the assumptions under which a valid post-merge analysis can be conducted.

2.4.1 The Merged Variable as an Outcome Variable

We first consider the scenario, in which researchers wish to use the variable Z merged from the data set $\mathcal{B}$ as a proxy for the outcome variable in a regression analysis. We assume that this regression analysis is applied to all observations of the data set $\mathcal{A}$ and uses a set of explanatory variables $\mathbf{X}$ taken from this data set. These explanatory variables may or may not include the variables used for merging. In the ANES application mentioned above, for example, we may be interested in regressing the validated turnout measure merged from the nationwide voter file on a variety of demographic variables measured in the survey.

For each observation i in the data set $\mathcal{A}$, we obtain the mean of the merged variable, i.e., $\zeta_i = \mathbb{E}(Z_i^* \mid \boldsymbol{\gamma}, \boldsymbol{\delta})$ where Z_i^* represents the true value of the merged variable. This quantity can be computed as the weighted average of the variable Z merged from the data set $\mathcal{B}$ where the weights are proportional to the match probabilities, i.e., $\zeta_i = \sum_{j=1}^{N_\mathcal{B}} \xi_{ij} Z_j / \sum_{j=1}^{N_\mathcal{B}} \xi_{ij}$. In the ANES application, for example, ζ_i represents the probability of turnout for survey respondent i in the data set $\mathcal{A}$ and can be computed as the weighted average of turnout among the registered voters in the voter file merged with respondent i. If we use thresholding and one-to-one match assignment so that each record in the data set $\mathcal{A}$ is matched with at most one record in the data set $\mathcal{B}$ (see Section 2.2), then we compute the mean of the merged variable as $\zeta_i = \sum_{j=1}^{N_\mathcal{B}} M_{ij}^* \xi_{ij} Z_j$ where M_{ij}^* is a binary variable indicating whether record i in the data set $\mathcal{A}$ is matched with record j in the data set $\mathcal{B}$ subject to the constraint $\sum_{j=1}^{N_\mathcal{B}} M_{ij}^* \leq 1$.

Under this setting, we assume that the true value of the outcome variable is independent of the explanatory variables in the regression conditional on the information used for merging, i.e.,

$$Z_i^* \perp\!\!\!\perp \mathbf{X}_i \mid (\boldsymbol{\delta}, \boldsymbol{\gamma}) \tag{7}$$

for each $i = 1, 2, \ldots, N_\mathcal{A}$. The assumption implies that the merging process is based on all relevant information. Specifically, within an agreement pattern, the true value of the merged variable Z_i^*

et al., 2016; Dalzell and Reiter, 2018).

is not correlated with the explanatory variables $\mathbf{X}_i$. Under this assumption, the law of iterated expectation implies that regressing ζ_i on $\mathbf{X}_i$ gives the results equivalent to the ones based on the regression of Z_i^* on $\mathbf{X}_i$ in expectation.

$$\mathbb{E}(Z_i^* \mid \mathbf{X}_i) \;=\; \mathbb{E}\{\mathbb{E}(Z_i^* \mid \boldsymbol{\gamma}, \boldsymbol{\delta}, \mathbf{X}_i) \mid \mathbf{X}_i\} \;=\; \mathbb{E}(\zeta_i \mid \mathbf{X}_i) \tag{8}$$

The conditional independence assumption may be violated if, for example, within the same agreement pattern, a variable correlated with explanatory variables is associated with merging error. Without this assumption, however, only the bounds can be identified (Cross and Manski, 2002). Thus, alternative assumptions such as parametric assumptions and exclusion restrictions are needed to achieve identification (see Ridder and Moffitt, 2007, for a review).

2.4.2 The Merged Variable as an Explanatory Variable

The second scenario we consider is the case where we use the merged variable as an explanatory variable. Suppose that we are interested in fitting the following linear regression model,

$$Y_i \;=\; \alpha + \beta Z_i^* + \eta^\top \mathbf{X}_i + \epsilon_i \tag{9}$$

where Y_i is a scalar outcome variable and the strict exogeneity is assumed, i.e., $\mathbb{E}(\epsilon_i \mid \mathbf{Z}^*, \mathbf{X}) = 0$ for all i. We follow the analysis strategy first proposed by Lahiri and Larsen (2005) but clarify the assumptions required for their approach to be valid (see also Hof and Zwinderman, 2012). Specifically, we maintain the assumption of no omitted variable for merging given in equation (7). Additionally, we assume that the merging variables are independent of the outcome variable conditional on the explanatory variables $\mathbf{Z}^*$ and $\mathbf{X}$, i.e.,

$$Y_i \perp\!\!\!\perp (\boldsymbol{\gamma}, \boldsymbol{\delta}) \mid \mathbf{Z}^*, \mathbf{X}. \tag{10}$$

Under these two assumptions, we can consistently estimate the coefficients by regressing Y_i on ζ_i and $\mathbf{X}_i$,

$$\begin{aligned} \mathbb{E}(Y_i \mid \boldsymbol{\gamma}, \boldsymbol{\delta}, \mathbf{X}_i) &= \alpha + \beta \mathbb{E}(Z_i^* \mid \boldsymbol{\gamma}, \boldsymbol{\delta}, \mathbf{X}_i) + \eta^\top \mathbf{X}_i + \mathbb{E}(\epsilon_i \mid \boldsymbol{\gamma}, \boldsymbol{\delta}, \mathbf{X}_i) \\ &= \alpha + \beta \zeta_i + \eta^\top \mathbf{X}_i \end{aligned} \tag{11}$$

where the second equality follows from the assumptions and the law of iterated expectation.

We generalize this strategy to the maximum likelihood (ML) estimation, which, to the best of our knowledge, has not been considered in the literature (but see Kim and Chambers (2012) for an estimating equations approach),

$$Y_i \mid Z_i^*, \mathbf{X}_i \overset{\text{indep.}}{\sim} P_\theta(Y_i \mid Z_i^*, \mathbf{X}_i) \tag{12}$$

where θ is a vector of model parameters. To estimate the parameters of this model, we maximize the following weighted log-likelihood function,

$$\hat{\theta} \;=\; \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{N_{\mathcal{A}}} \sum_{j=1}^{N_{\mathcal{B}}} \xi_{ij}^* \log P_\theta(Y_i \mid Z_i^* = Z_j, \mathbf{X}_i) \tag{13}$$

where $\xi_{ij}^* = \xi_{ij} / \sum_{j'=1}^{N_{\mathcal{B}}} \xi_{ij'}$. Online SI G shows that under the two assumptions described earlier and mild regularity conditions, the weighted ML estimator given in equation (13) is consistent and asymptotically normal. Note that because we are considering large data sets, we ignore the uncertainty about ξ_{ij}^*.

3 Simulation Studies

We conduct a comprehensive set of simulation studies to evaluate the statistical accuracy and computational efficiency of our probabilistic modeling approach and compare them with those of deterministic methods. Specifically, we assess the ability of the proposed methodology to control estimation error, false positives and false negatives, and its robustness to missing values and noise in the linkage fields, as well as the degree of overlap between two data sets to be merged. We do so by systematically varying the amount and structure of missing data and measurement error.

3.1 The Setup

To make our simulation studies realistic, we use a data set taken from the 2006 California voter file. Since merging voter files is often done by blocking on gender, we subset the data set to extract the information about female voters only, reducing the number of observation to approximately 17 million voters to 8.3 million observations. To create a base data set for simulations, we further subset the data set by removing all observations that have at least one missing value in the following variables: first name, middle initial, last name, date of birth, registration date, address, zip code, and turnout in the 2004 Presidential election. After listwise deletion, we obtain the final

data set of 341,160 voters, from which we generate two data sets of various characteristics to be merged. From this data set, we independently and randomly select two subsamples to be merged under a variety of scenarios.

We design our simulation studies by varying the values of the five parameters as summarized below. Online SI I.1 describes in detail the precise setups of these simulations.

1. **Degree of overlap**: Proportion of records in the smaller data set that are also in the larger data set. We consider three scenarios — 20% (small), 50% (medium), and 80% (large).

2. **Size balance**: Balance of sample sizes between the two data sets to be merged. We consider three ratios — 1:1 (equally sized), 1:10 (imbalanced), and 1:100 (lopsided).

3. **Missing data**: We consider five different mechanisms, missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR). For MAR and NMAR, we consider independent and dependent missingness patterns across linkage fields

4. **Amount of missing data**: Proportion of missing values in each linkage variable other than year of birth. We consider three scenarios — 5% (small), 10% (medium), and 15% (large).

5. **Measurement error**: Proportion of records (6%) whose first name, last name, and street name contains classical measurement error.

Together, we conduct a total of 135 ($= 3^3 \times 5$) simulation studies where missing data are of main concern. We also conduct another set of 135 simulations with various types of non-classical measurement errors, while keeping the amount of missing values fixed (see Online SI I.2).

3.2 Results

Figure 1 compares the performance of fastLink (blue solid bars) to the two deterministic methods often used by social scientists. The first is the merging method based on exact matches (red shaded bars), while the second is the recently proposed partial match algorithm (ADGN; light green solid bars) that considers two records as a match if at least three fields of their address, date of birth, gender, and name are identical (Ansolabehere and Hersh, 2017). The top panel of Figure 1 presents the FNR while the bottom panel presents the absolute error for estimating the 2004 turnout rate. We merge two data sets of equal size (100,000 records each) after introducing

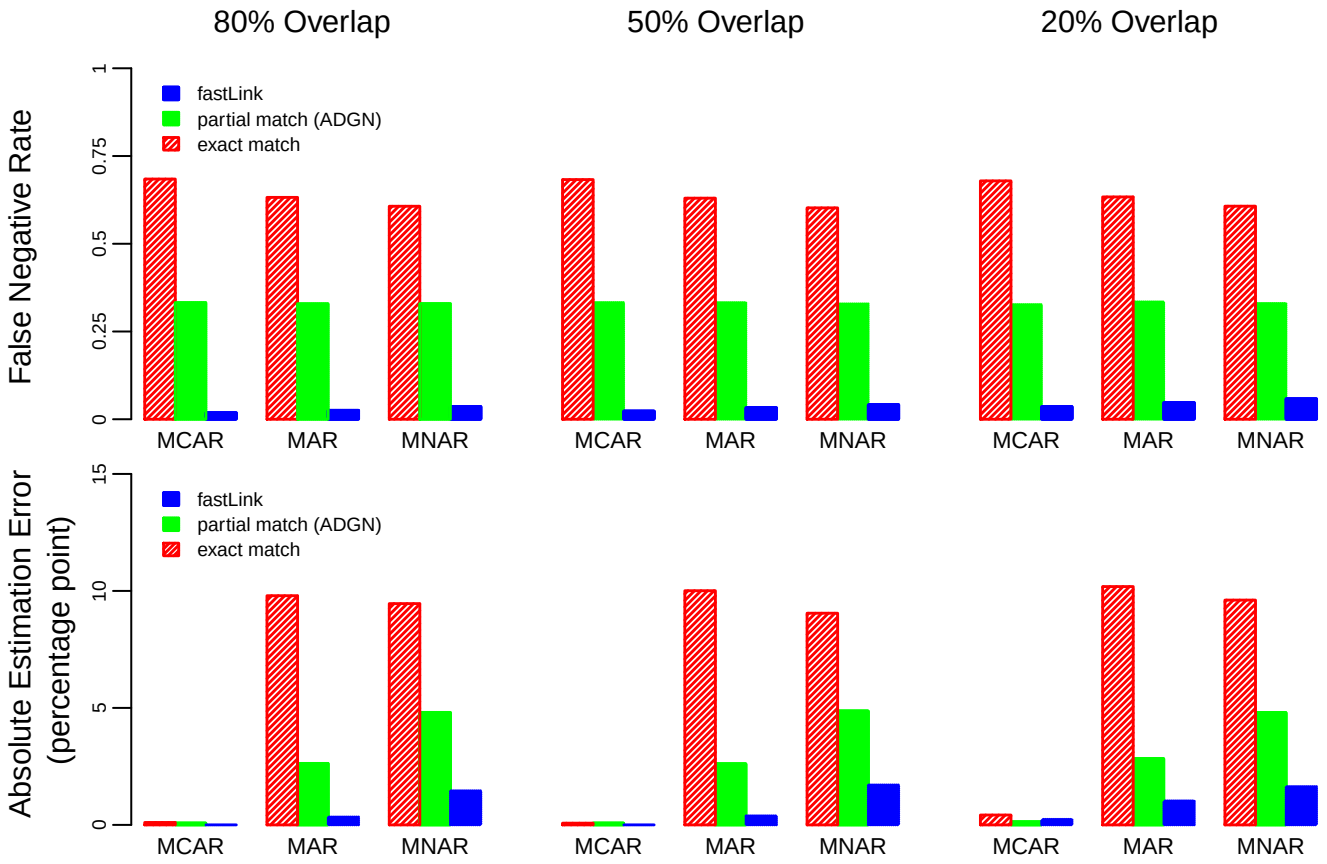

Figure 1: Accuracy of Data Merge. The top and bottom panels present the false discovery rate (FNR) and the absolute estimation error (for estimating the turnout rate), respectively, when merging datasets of 100,000 records each across with different levels of overlap (measured as a percentage of a data set). Three missing data mechanisms are studied with the missing data proportion of 10% for each linkage field other than year of birth: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). Classical measurement error is introduced to several linkage fields. The proposed probabilistic methodology (fastLink; blue solid bars) significantly outperforms the two deterministic algorithms, i.e., exact match (red shaded bars) and partial match (ADGN; light green solid bars), across simulation settings.

the classical measurement error and the medium amount of missing data as explained above. For fastLink, only pairs with a match probability ≥ 0.85 are considered to be matches, but the results remain qualitatively similar if we change the threshold to 0.75 or 0.95.

We find that fastLink significantly outperforms the two deterministic methods.[10] While all three methods are designed to control the FDR, only fastLink is able to keep the FNR low (less than 5 percent in all cases considered here). The deterministic algorithms are not robust to missing data and measurement error, yielding a FNR of much greater magnitude. Additionally, we observe

[10]In Online SI H, we compare fastLink to the state-of-the-art probabilistic methods, and finds that fastLink performs as well as these methods.

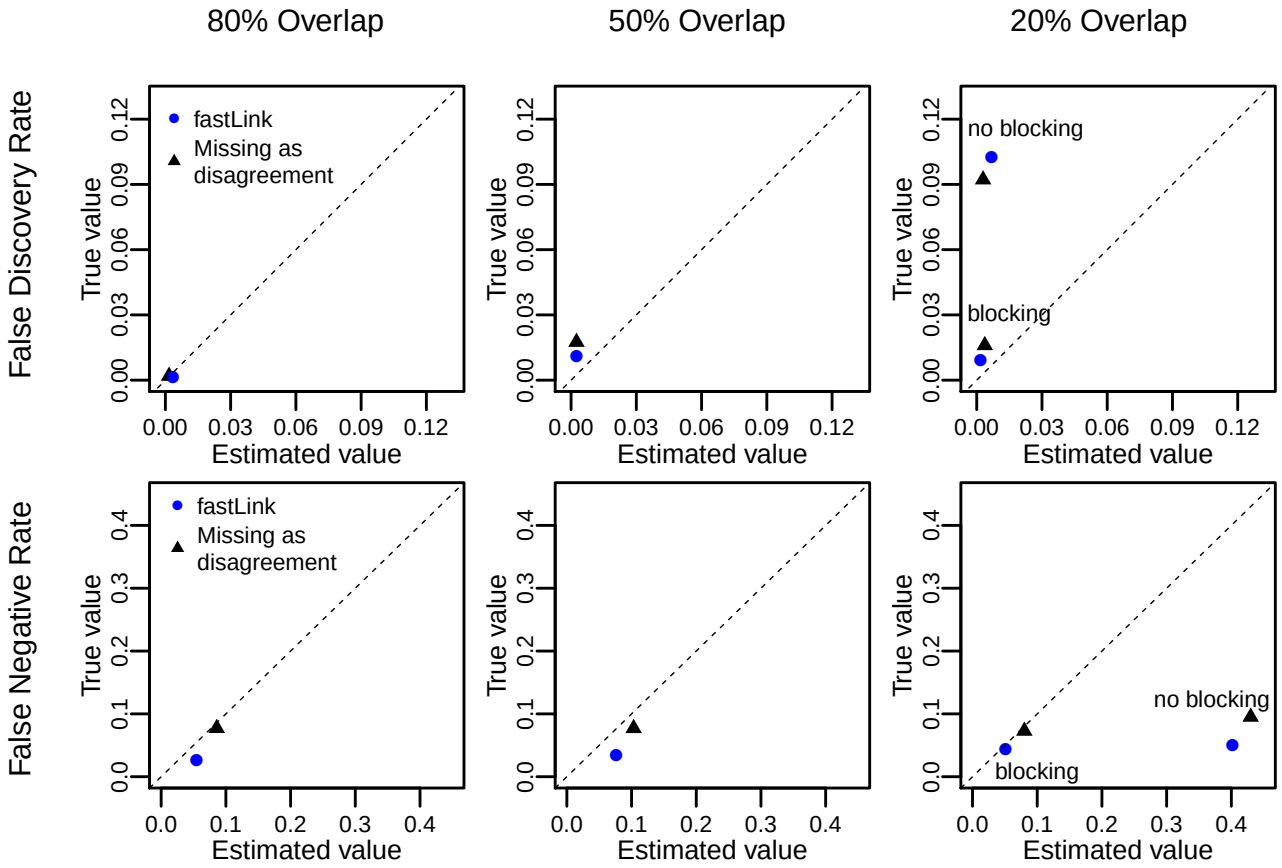

Figure 2: Accuracy of FDR and FNR Estimates. The top panel compares the estimated FDR (x-axis) with its true value (y-axis) whereas the bottom panel compares the estimated FNR against its true value. We consider the medium amount of missing data generated under MAR as a missingness mechanism and add 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. Additionally, 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 improves the precision of our estimates.

that the deterministic methods yield a substantially greater estimation bias than fastLink unless the data are missing completely at random. Under the other two missing data mechanisms, the magnitude of the bias is substantially greater than that of fastLink. While fastLink has an absolute estimation error of less than 1.5 percentage points even under MNAR, the other two methods have an absolute estimation error of more than 7.5 percentage points under both MAR and MNAR. Finally, the performance of fastLink worsens as the size of overlap reduces and the missing data mechanism becomes less random.

We next evaluate the accuracy of FDR and FNR estimates in the top and bottom panels, respectively. Since the deterministic methods do not give such error estimates, we compare the

performance of the proposed methodology (indicated by blue solid circles) with that of the same probabilistic modeling approach, which treats missing values as disagreements following a common practice in the literature (indicated by solid triangles). Figure 2 presents the results of merging two data sets of equal size where the medium amount of data are assumed to be missing at random and some noise are added as described earlier. In the top panel of the figure, we find that the true FDR is low and its estimate is accurate unless the degree of overlap is small. With a small degree of overlap, both methods significantly underestimate the FDR. A similar finding is obtained for the FNR in the bottom panel of the figure where estimated FNR is biased upward.

One way to address the problem of having small overlap would be to use blocking based on a set of fully observed covariates. For example, in our simulations, since the year of birth is observed for each record in both datasets, we block the data by making comparisons only across individuals within a window of $\pm$ 1 year around each birth year.[11] Then, we apply fastLink to each block separately. As shown in the right most column of Figure 2, blocking significantly improves the estimation accuracy for the FDR and FNR estimates as well as their true values although the bias is not eliminated. The reason for this improvement is that traditional blocking increases the degree of overlap. For example, in this simulation setting for each of the 94 blocks under consideration, the ratio of true matches to all possible pairs is at least 8×10^{-5}, which is more than 15 times as large as the corresponding ratio for no blocking and is comparable to the case of overlap of 50%.

We present the results of the remaining simulation studies in the Online Simulation Appendix. Two major patterns discussed above are also found under these other simulation scenarios. First, regardless of the missing data mechanisms and the amount of missing observations, fastLink controls FDR, FNR, and estimation error well. Second, a greater degree of overlap between datasets leads to better merging results in terms of FDR and FNR as well as the accuracy of their estimates. Blocking can ameliorate these problems caused by small overlap to some extent. These empirical patterns are consistently found across simulations even when two datasets have unequal sizes.

3.3 Computational Efficiency

We compare the computational performance of fastLink with that of the RecordLinkage package in R (Sariyar and Borg, 2016) and the Record Linkage package in Python (de Bruin, 2017) in terms

[11] In Online SI L, we also present results using a clustering method, i.e., k-means, to group similar observations.

of running time. The latter two are the only other open source packages in R and Python that implement a probabilistic model of record linkage under the Fellegi-Sunter framework. To mimic a standard computing environment of applied researchers, all the calculations are performed in a Macintosh laptop computer with a 2.8 GHz Intel Core i7 processor and 8 GB of RAM.

While `fastLink` takes advantage of a multicore machine via the OpenMP-based parallelization (the other two packages do not have a parallelization feature), we perform the comparison on a single-core computing environment so that we can assess the computational efficiency of our algorithm itself. Additionally, we include runtime results where we parallelize computation across eight cores. For all implementations, we set the convergence threshold to 1×10^{-5}.[12]

We consider the setup in which we merge two datasets of equal size with 50% overlap, 10% missing proportion under MCAR, and no measurement error. Our linkage variables are first name, middle initial, last name, house number, street name, and year of birth. We vary the size of each data set from 1,000 records to 300,000 observations. As in the earlier simulations, each dataset is based on the sample of 341,160 female registered voters in California, for whom we have complete information in each linkage field. To build the agreement patterns, we use the Jaro-Winkler string distance with a cutoff of 0.94 for first name, last name, and street name. For the remaining fields, we only consider exact matches as agreements.

Figure 3 presents the results of this running time comparison. We find that although all three packages take a similar amount of time for data sets of 1,000 records, the running time increases exponentially for the other packages in contrast to fastLink (black solid triangles connected by a dashed line, single core; blue solid circles connected by a solid line, 8 cores), which exhibits a near linear increase. When matching data sets of 150,000 records each, fastLink takes less than 6 hours to merge using a single core (under 3 hours when parallelized across 8 cores). In contrast, it takes more than 24 hours for Record Linkage (Python; solid purple squares connected by a dotted line), to merge two data sets of only 20,000 observations each. The performance is not as bad for Record Linkage (R; red crosses connected by a dashed line), but it still takes over 6 hours to merge data sets of 40,000 records each. Moreover, an approximation based on an exponential regression

[12]Starting values differ across methods because other methods do not allow us to change their default starting values. However, the EM algorithm converges quickly regardless of the choice of starting values. In fact, it is well known that the bottleneck is a large number of required comparisons (e.g., Jaro, 1972; Christen, 2012), for which we use a hashing technique as described below in Appendix A.2.

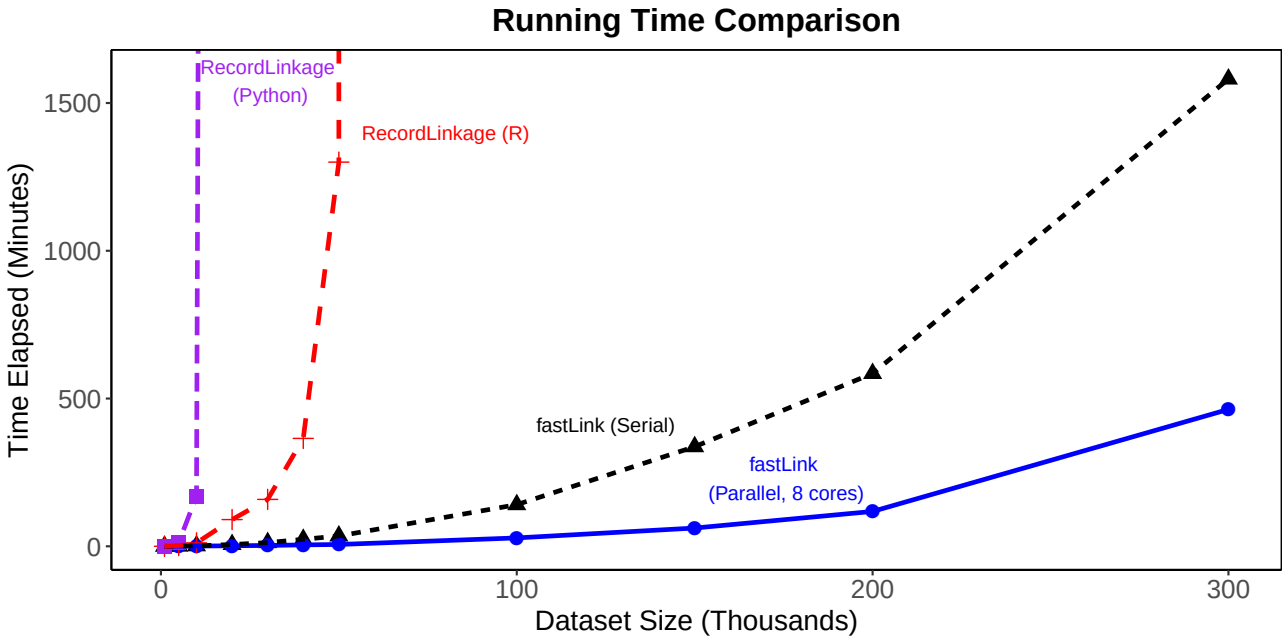

Figure 3: Running Time Comparison. The plot presents the results of merging datasets of equal size using different implementations of the Fellegi-Sunter model. The datasets were constructed from a sample of female registered voters in California. The amount of overlap between datasets is 50%, and, for each dataset, there are 10% missing observations in each linkage variable: first name, middle initial, last name, house number, street name, and year of birth. The missing data mechanism is Missing Completely at Random (MCAR). The computation is performed on a Macintosh laptop computer with a 2.8 GHz Intel Core i7 processor and 8 GB of RAM. The proposed implementation fastLink (single-core runtime as black solid triangles connected by a dashed line, and parallelized over eight cores as blue solid dots connected by a solid line) is significantly faster than the other open-source packages.

model suggests that Record Linkage (R) would take around 22 hours to merge two data sets of 50,000 records each, while Record Linkage (Python) would take about 900 days to accomplish the same merge. In Online SI C.1, we further decompose the runtime comparison to provide more detail on the sources of our computational improvements. We detail the choices we make in the computational implementation that yields these substantial efficiency gains in Appendix A.

4 Empirical Applications

In this section, we present two empirical applications of the proposed methodology. First, we merge election survey data (about 55,000 observations) with political contribution data (about 5 million observations). The major challenge of this merge is the fact that the expected number of matches between the two data sets is small. Therefore, we utilize blocking and conduct the

data merge within each block. The second application is to merge two nationwide voter files, each of which has more than 160 million records. This may, therefore, represent the largest data merge ever conducted in the social sciences. We show how to use auxiliary information about within-state and across-state migration rates to inform the match.

4.1 Merging Election Survey Data with Political Contribution Data

Hill and Huber (2017) study differences between donors and non-donors by merging the 2012 Cooperative Congressional Election Study (CCES) survey with the Database on Ideology, Money in Politics, and Elections (DIME, Bonica (2013)). The 2012 CCES is based on a nationally representative sample of 54,535 individuals recruited from the voting-age population in the United States. The DIME data, on the other hand, provide the information about individual donations to political campaigns. For the 2010 and 2012 elections, the DIME contains over 5 million donors.

The original authors asked YouGov, the company which conducted the survey, to merge the two data sets using a proprietary algorithm. This yielded a total of 4,432 CCES respondents matched to a donor in the DIME data. After the merge, Hill and Huber (2017) treat each matched CCES respondent as a donor and conduct various analyses by comparing these matched respondents with those who are not matched with a donor in the DIME data and hence are treated as non-donors. Below, we apply the proposed methodology to merge these two data sets and conduct a post-merge analysis by incorporating the uncertainty about the merge process.

4.1.1 Merge Procedure

We use the name, address, and gender information to merge the two data sets. In order to protect the anonymity of CCES respondents, YouGov used fastLink to merge the data sets on our behalf. Moreover, due to contractual obligations, the merge was conducted only for 51,184 YouGov panelists, which is a subset of the 2012 CCES respondents. We block based on gender and state of residence, resulting in 102 blocks (50 states plus Washington DC $\times$ two gender categories). The size of each block ranges from 175,861 (CCES = 49, DIME = 3589) to 790,372,071 pairs (CCES = 2,367, DIME = 333,913) with the median value of 14,048,151 pairs (CCES = 377, DIME = 37,263). Within each block, we merge the data sets using the first name, middle initial, last name, house number, street name, and postal code. As done in the simulations, we use three levels of agreement for the string valued variables based on the Jaro-Winkler distance with 0.85 and 0.92 as

the thresholds. For the remaining variables (i.e., middle initial, house number, and postal code), we utilize a binary comparison indicating whether they have an identical value.

To construct our set of matched pairs between CCES and DIME, first, we use the one-to-one matching assignment algorithm described in Online SI E and find the best match in the DIME data for each CCES respondent. Then, we declare as a match any pair whose matching probability is above a certain threshold. We use three thresholds, i.e., 0.75, 0.85, and 0.95, and examine the sensitivity of the empirical results to the choice of threshold value.[13] Finally, in the original study of Hill and Huber (2017), noise is added to the amount of contribution in order to protect the anonymity of matched CCES respondents. However, we signed a non-disclosure agreement with YouGov for our analysis so that we can make a precise comparison between the proposed methodology and the proprietary merge method used by YouGov.

4.1.2 Merge Results

Table 2 presents the merge results. We begin by assessing the match rates, which represent the proportion of CCES respondents who are matched with donors in the DIME data. While the match rates are similar between the two methods, fastLink appears to find slightly more (less) matches for male (female) respondents than the proprietary method regardless of the threshold used. However, this does not mean that both methods find identical matches. In fact, out of 4,797 matches identified by fastLink (using the threshold of 0.85), the proprietary method does not identify 861 or 18% of them as matches.

As discussed in Section 2.2, one important advantage of the probabilistic modeling approach is that we can estimate the FDR and FNR, which are shown in the table. Such error rates are not available for the proprietary method. As expected, the overall estimated FDR is controlled to less than 1.5% for both male and female respondents. The estimated FNR, on the other hand, is large, illustrating the difficulty of finding some donors. In particular, we find that female donors are much more difficult to find than male donors.

Specifically, there are 12,803 CCES respondents who said they made a campaign contribution during the last 12 months before the 2012 election. Among them, 5,206 respondents claimed

[13] In Online SI J.3, instead of a one-to-one matching restriction used here, we present the results of the weighted approach described in Section 2.2.2. As shown in Figure S7 of Online SI J.3, there is no distinguishable difference in the results obtained from either approach.

		fastLink			Proprietary method
		0.75	0.85	0.95	
Number of matches	All	4948	4797	4576	4534
	Female	2198	2156	2067	2210
	Male	2750	2641	2524	2324
Overlap between fastLink and proprietary method	All	3959	3936	3881	
	Female	1877	1866	1844	
	Male	2082	2070	2037	
Match rate (%)	All	9.67	9.37	8.94	8.85
	Female	8.12	7.96	7.63	8.16
	Male	11.40	10.95	10.40	9.64
False discovery rate (FDR; %)	All	1.24	0.65	0.21	
	Female	0.92	0.53	0.14	
	Male	1.50	0.75	0.28	
False negative rate (FNR; %)	All	15.25	17.35	20.81	
	Female	5.35	6.80	10.30	
	Male	21.83	24.36	27.79	

Table 2: The Results of Merging the 2012 Cooperative Congressional Election Study (CCES) with the 2010 and 2012 Database on Ideology, Money in Politics, and Elections (DIME) Data. The table presents the merging results for both fastLink and the proprietary method used by YouGov. The results of fastLink are presented for one-to-one match with three different thresholds (i.e., 0.75, 0.85, 0.95) for the matching probability to declare a pair of observations as a successful match. The number of matches, the amount of overlap, and the overall match rates are similar between the two methods. The table also presents information on the estimated false discovery and false negative rates (FDR and FNR, respectively) obtained using fastLink. These statistics are not available for the proprietary method.

to have donated at least 200 dollars. Interestingly, both fastLink and the proprietary method matched an essentially identical number of self-reported donors with a contribution of over 200 dollars (2,431 and 2,434 or approximately 47%, respectively), whereas among the self-reported small donors both methods can only match approximately 16% of them.

Next, we examine the quality of matches for the two methods (see also Online SI M). We begin by comparing the self-reported donation amount of matched CCES respondents with their actual donation amount recorded in the DIME data. While only donations greater than 200 dollars are recorded at the federal level, the DIME data include some donations of smaller amounts, if not all, at the state level. Thus, while we do not expect a perfect correlation between self-reported and actual donation amount, under the assumption that donors do not systematically under- or over-report the amount of campaign contributions, a high correlation between the two measures

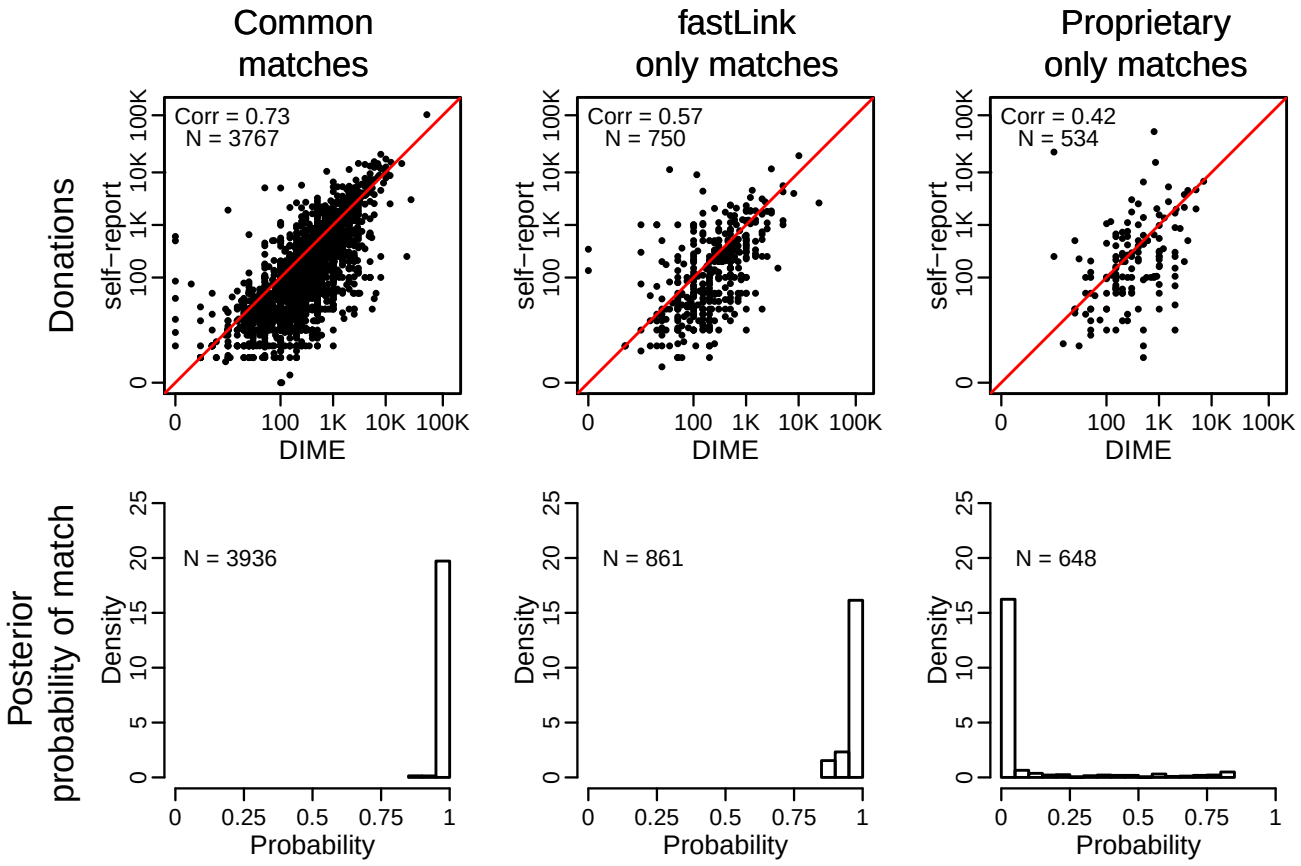

Figure 4: Comparison of fastLink and the Proprietary Method. The top panel compares the self-reported donations (y-axis) by matched CCES respondents with their donation amount recorded in the DIME data (x-axis) for the three different groups of observations: those declared as matches by both fastLink 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 0.85.

implies a more accurate merging process.

The upper panel of Figure 4 presents the results where for fastLink, we use one-to-one match with the threshold of 0.85.[14] We find that for the respondents who are matched by both methods, the correlation between the self-reported and matched donation amounts is reasonably high (0.73). In the case of respondents who are matched by fastLink only, we observe that the correlation is low (0.57) but is greater than the correlation for those matches identified by the proprietary method alone (0.42). We also examine the distribution of match probabilities for these three groups of matches. The bottom panel of the figure presents the results, which are consistent with the patterns of correlation identified in the top panel. That is, those matches identified by the two

[14]Figures S5 and S6 in Online SI J present the results under two different thresholds: 0.75 and 0.95, respectively. The results under those thresholds are similar to the those with the threshold of 0.85 presented here.

methods have the highest match probability whereas most of the matches identified only by the proprietary method have extremely low match probabilities. In Online SI M we also examine the quality of the agreement patterns separately for the matches identified by both methods, fastLink only, and the proprietary method only. Overall, our results indicate that fastLink produces matches whose quality is often better than those based on the proprietary method.

4.1.3 Post-merge Analysis

An important advantage of the probabilistic modeling approach is its ability to account for the uncertainty of the merge process in post-merge analyses. We illustrate this feature by revisiting the post-merge analysis of Hill and Huber (2017). The original authors are interested in the comparison of donors (defined as those who are matched with records in the DIME data) and non-donors (defined as those who are not matched) among CCES respondents. Using the matches identified by a proprietary method, Hill and Huber (2017) regress policy ideology on the matching indicator variable, which is interpreted as a donation indicator variable, the turnout indicator variables for the 2012 general election and 2012 congressional primary elections, as well as several demographic variables. Policy ideology, which ranges from -1 (most liberal) to 1 (most conservative), is constructed by applying a factor analysis to a series of questions on various issues.[15] The demographic control variables include income, education, gender, household union membership, race, age in decades, and importance of religion. The same model is fitted separately for Democrats and Republicans.

To account for the uncertainty of the merge process, as explained in Section 2.4, we fit the same linear regression except that we use the mean of the match indicator variable as the main explanatory variable rather than the match indicator variable. Table 3 presents the estimated coefficients of the aforementioned linear regression models with the corresponding heteroskedasticity-robust standard errors in parentheses. Generally, the results of our improved analysis agree with those of the original analysis, showing that donors tend to be more ideologically extreme than non-donors.

While the overall conclusion is similar, the estimated coefficients are smaller in magnitude when accounting for the uncertainty of merge process. In particular, according to fastLink, for Republican respondents, the estimated coefficient of being a donor represents only 12% of the

[15] They include gun control, climate change, immigration, abortion, jobs versus the environment, gay marriage, affirmative action, and fiscal policy.

	Republicans		Democrats	
	Original	fastLink	Original	fastLink
Contributor	0.080***	0.046***	−0.180***	−0.165***
	(0.016)	(0.015)	(0.008)	(0.009)
Turnout for 2012	0.095***	0.095***	−0.060***	−0.060***
general election	(0.013)	(0.013)	(0.010)	(0.010)
Turnout for 2012	0.094***	0.095***	−0.019**	−0.022***
primary election	(0.009)	(0.009)	(0.009)	0.009)
Demographic Controls	Yes	Yes	Yes	Yes
Number of observations	17386	17386	20925	20925

Table 3: Predicting Policy Ideology Using Contributor Status. The estimated coefficients from the linear regression of policy ideology score on the contributor indicator variable and a set of demographic controls. Along with the original analysis, the table presents the results of the improved analysis based on fastLink, which accounts for the uncertainty of the merge process. *** p<0.001, ** p<0.01, * p<0.05. Robust standard errors in parentheses.

standard deviation of their ideological positions (instead of 21% given by the proprietary method). Indeed, the difference in the estimated coefficients between fastLink and the proprietary method is statistically significant for both Republicans (0.035, $s.e.$ = 0.014), and Democrats (−0.015, $s.e.$ = 0.007). Moreover, although the original analysis find that the partisan mean ideological difference for donors (1.108, $s.e.$ = 0.018) is 31 percent larger than that for non-donors (0.848, $s.e.$ = 0.001), the results based on fastLink shows that this difference is only 25 percent larger for donors (1.058, $s.e.$ = 0.018). Thus, while the proprietary method suggests that the partisan gap for donors is similar to the partisan gap for those with a college degree or higher (1.100, $s.e.$ = 0.036), fastLink shows that it is closer to the partisan gap for those with just some college education but without a degree (1.036, $s.e.$ = 0.035).

4.2 Merging Two Nationwide Voter Files over Time

Our second application is what might be the largest data merging exercise ever conducted in social sciences. Specifically, we merge the 2014 nationwide voter file to the 2015 nationwide voter file, each of which has over 160 million records. The data sets are provided by L2, Inc., a leading national non-partisan firm and the oldest organization in the United States that supplies voter data and related technology to candidates, political parties, pollsters and consultants for use in campaigns. In addition to the sheer size of the data sets, merging these nationwide voter files

is methodologically challenging because some voters change their residence over time, making addresses uninformative for matching these voters.

4.2.1 Merge Procedure

When merging data sets of this scale, we must drastically reduce the number of comparisons. In fact, if we examine all possible pairwise comparisons between the two voter files, the total number of such pairs exceeds 2.5×10^{16}. It is also important to incorporate auxiliary information about movers since the address variable is non-informative when matching these voters. We use the Internal Revenue Service Statistics of Income (IRS SOI) to calibrate match rates for within-state and across-state movers. Details on incorporating migration rates into parameter estimation can be found in Section 2.3 and Online SI F.2. The IRS SOI data is a definitive source of migration data in the United States that tracks individual residences year-to-year across all states through their tax returns.

We develop the following two-step procedure that utilizes random sampling and blocking of voter records to reduce the computational burden of the merge (see Online SI C.2 and Section F.2). Our merge is based on first name, middle initial, last name, house number, street name, date/year/month of birth, date/year/month of registration, and gender. The first step uses each of these fields to inform the merge, while the second step uses only first name, middle initial, last name, date/year/month of birth, and gender. For both first name and last name, we include a partial match category based on the Jaro-Winkler string distance calculation, setting the cutoff for a full match at 0.92 and for a partial match at 0.88.

As described in Online SI F.2, we incorporate auxiliary information into the model by moving from the likelihood framework to a fully Bayesian approach. Due to conjugacy of our priors, we can obtain the estimated parameters by maximizing the log posterior distribution via the EM algorithm. This approach allows us to maintain the computational efficiency.[16]

Step 1: Matching within-state movers and non-movers for each state.

[16]Specifically, we set prior parameters on the expected match rate and expected within-state movers rate using the IRS data, giving 75% weight to the prior estimate and 25% weight to the maximum likelihood estimate. For the first step, we set priors on both $\pi_{\text{address},1,0}$ (the probability of a voter's address not matching conditional on being in the matched set, which is equivalent to the share of in-state movers in the matched set) and λ. For the second step, we set a prior on λ.

(a) Obtain a random sample of voter records from each state file

(b) Fit the model to this sample using the within-state migration rates from the IRS data to specify prior parameters

(c) Create blocks by first stratifying on gender and then applying the k-means algorithm to the first name

(d) Using the estimated model parameters, conduct the data merge within each block

Step 2: Matching across-state movers for each pair of states.

(a) Set aside voters who are identified as successful matches in Step 1

(b) Obtain a random sample of voter records from each state file as done in Step 1(a)

(c) Fit the model using the across-state migration rates from the IRS data to specify prior parameters

(d) Create blocks by first stratifying on gender and then applying the k-means algorithm to the first name as done in Step 1(c)

(e) Using the estimated model parameters, conduct the data merge within each block as done in Step 1(e)

In Step 1, we apply random sampling, rather than blocking, strategy in order to use the within-state migration rates from the IRS data and fit the model to a representative sample for each state. For the same reason, we use a random sampling strategy in Step 2 to exploit the availability of IRS across-state migration rates. We obtain a random sample of 800,000 voter records for files with over 800,000 voters and use the entire state file for states with fewer than 800,000 voter records on file. Online SI K shows through simulation studies that for datasets as small as 100,000 records, a 5% random sample leads to parameter estimates nearly indistinguishable from those obtained using the full data set. Based on this finding, we choose 800,000 records as the size of the random samples, corresponding to a 5% of records from California, the largest state in the United States.

Second, within each step, we conduct the merge by creating blocks in order to reduce the number of pairs for consideration. We block based on gender, first name, and state, and we select the number of blocks so that the average size of each blocked dataset is approximately 250,000

		fastLink			
		0.75	0.85	0.95	Exact
Match count (millions)	All	135.60	129.69	128.73	91.62
	Within-state	127.38	127.12	126.80	91.36
	Across-state	8.22	2.57	1.93	0.27
Match rate (%)	All	97.25	93.67	93.04	66.24
	Within-state	92.06	91.87	91.66	66.05
	Across-state	5.19	1.80	1.38	0.19
False discovery rate (FDR; %)	All	1.02	0.10	0.03	
	Within-state	0.08	0.04	0.01	
	Across-state	0.95	0.06	0.02	
False negative rate (FNR; %)	All	3.35	3.63	3.86	
	Within-state	2.63	2.83	3.05	
	Across-state	0.72	0.80	0.81	

Table 4: The Results of Merging the 2014 Nationwide Voter File with the 2015 Nationwide Voter File. This table presents the merging results for fastLink for three different thresholds (i.e., 0.75, 0.85, 0.95) for the matching probability to declare a pair of observations a successful match. Across the different thresholds, the match rates do not change substantially and are significantly greater than the corresponding match rates of the exact matching technique.

records. To block by first name, we rank-ordered the first names alphabetically and ran the k-means algorithm on this ranking in order to create clusters of maximally similar names.[17] Finally, the entire merge procedure is computationally intensive. The reason is that we need to repeat Step 1 for each of 50 states plus Washington DC and apply Step 2 to each of 1275 pairs. Thus, as explained in Online SI C.2, we use parallelization whenever possible. All merges were run on a Linux cluster with 16 2.4-GHz Broadwell 28-core nodes with 128 GB of RAM per node.

4.2.2 Merge Results

Table 4 presents the overall match rate, FDR, and FNR obtained from fastLink. We assess the performance of the match at three separate matching probability thresholds to declare a pair of observations a successful match: 0.75, 0.85, and 0.95. We also break out the matches by within-state matches only and across-state matches only. Across the three thresholds, the overall match rate remains very high, at 93.04% under a 95% acceptance threshold, while the estimated FDR and FNR remain controlled at 0.03% and 3.86%. All three thresholds yield match rates that are

[17] See Online SI N for evidence that this blocking strategy performs similarly to a blocking strategy based on age windowing.

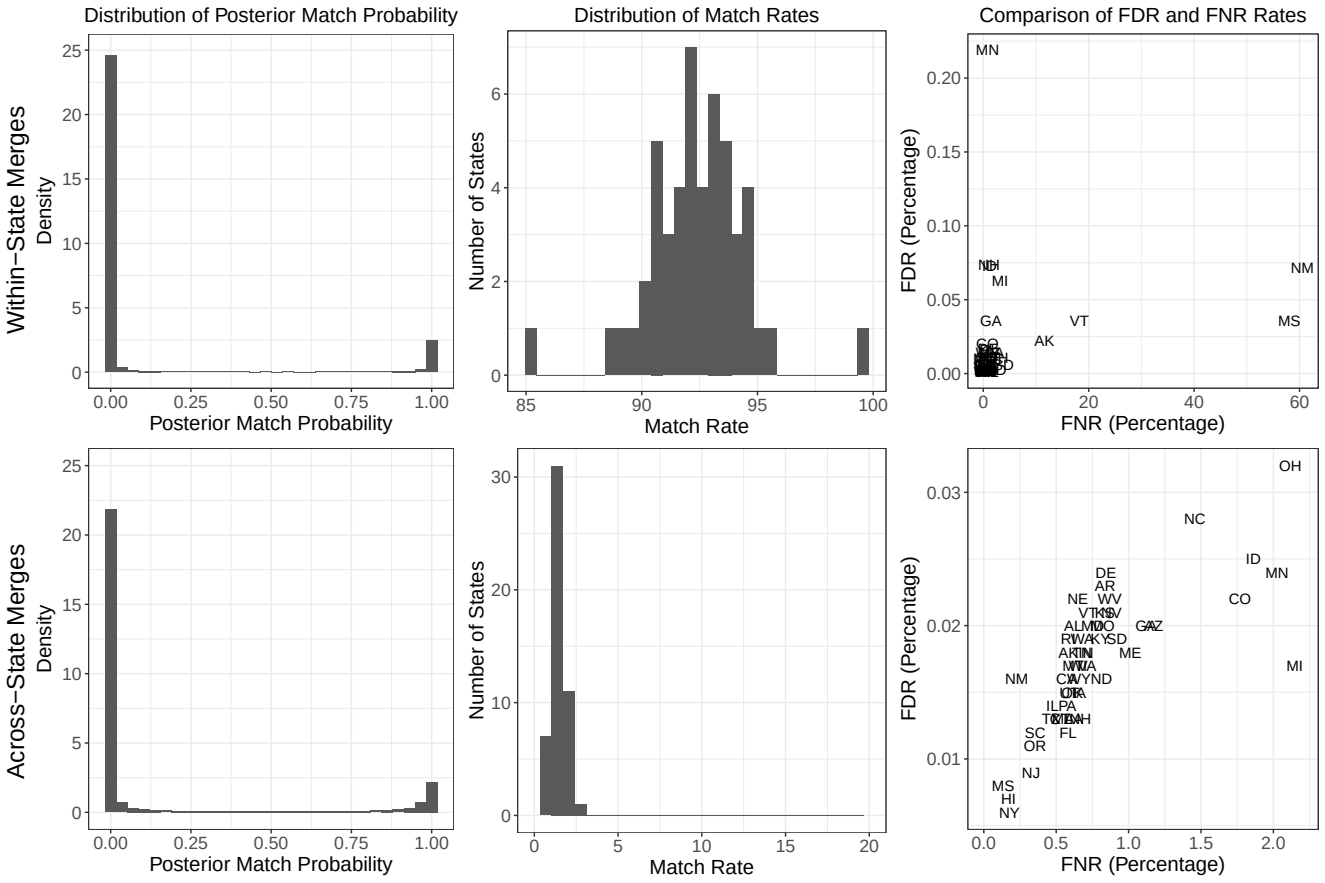

Figure 5: Graphical Diagnostics from Merging the 2014 Nationwide Voter File with the 2015 Nationwide Voter File. This figure presents graphical diagnostics for fastLink for within-state matches (top panel) and across-state matches (bottom panel). The first column plots the distribution of the matching probability across all patterns. The second column plots the distribution of the match rate for each state. Lastly, the third column compares the FNR against the FDR for each state separately.

significant higher than the corresponding match rates of the exact matching technique.

In Figure 5, we examine the quality of the merge separately for the within-state merge (top panel) and across-state merge (bottom panel). The first column plots the distribution of the matching probability across all potential match pairs. For both within-state and across-state merge, we observe a clear separation between the successful matches and unsuccessful matches, with very few matches falling in the middle. This suggests that the true and false matches are identified reasonably well. In the second column, we examine the distribution of the match rate by state. Here, we see that most states are tightly clustered between 88% and 96%. Only Ohio, with a match rate of 85%, has a lower match rate. For the across state merge, the match rate is

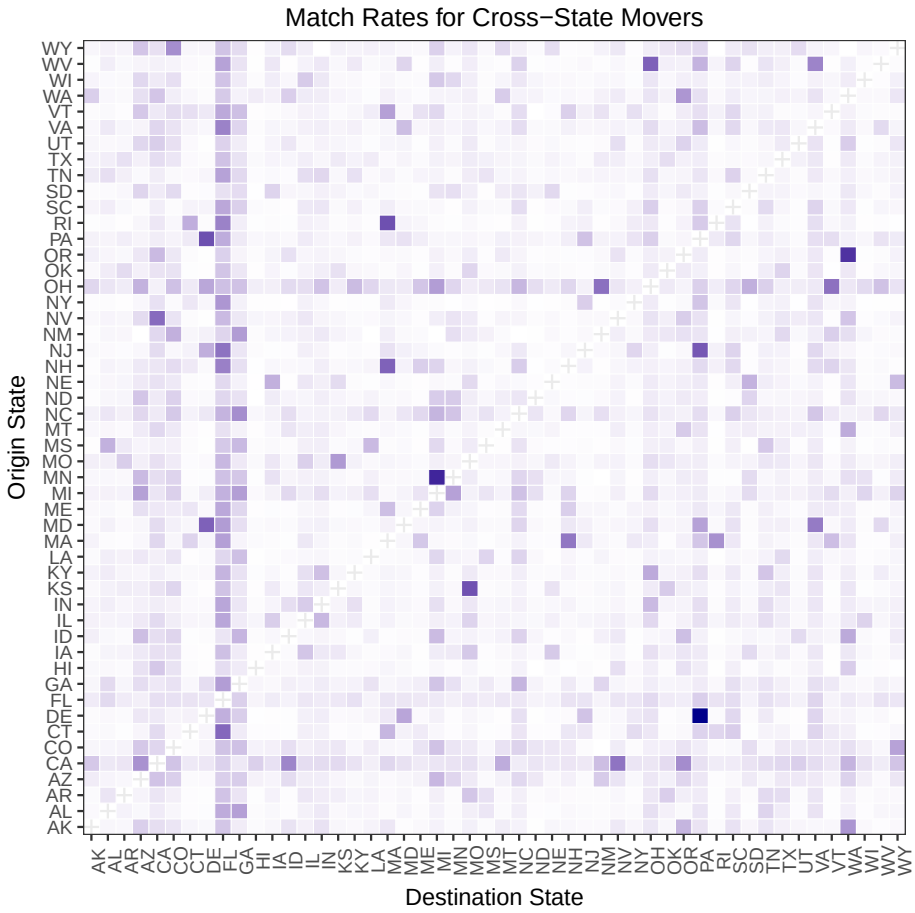

Figure 6: Across-State Match Rates for the 2014 Nationwide Voter File to 2015 Nationwide Voter File Merge. We plot the match rates from each across-state match pair as a heatmap, where darker colors indicate a higher match rate.

clustered tightly between 0% and 5%.

In the third column, we plot the estimated FDR against the estimated FNR for each state. For the within-state merge, the FDR is controlled well — every state other than Minnesota has an FDR below 0.1%. Additionally, there are only two states, Mississippi and New Mexico, where fastLink seems to have trouble identifying true matches, as measured by the FNR. In the across-state merge, the FDR for every state is below 0.1%, suggesting that the resulting matches are of high quality. Furthermore, fastLink appears to be finding a high share of true movers across voter files, as the FNR for all but three states falls under 2%.

Finally, we examine the across-state migration patterns recovered from our matching procedure. Figure 6 displays a heatmap of the migration patterns obtained from fastLink with darker purple colors indicating a higher match rate when merging the 2014 nationwide voter file for a

given state (Origin State) to the 2015 nationwide voter file for a given state (Destination State). We uncover several regional migration patterns. First, we find a migration cluster in New England, where voters from New Hampshire and Rhode Island migrated to Massachusetts between 2014 and 2015. Another strong migration cluster exists between New Jersey, Delaware, and Pennsylvania in the mid-Atlantic region. Both patterns suggest that most migration occurs between clusters of adjacent states and urban centers. Lastly, we find a large volume of out-migration to Florida from across the United States, and the out- migration is particularly concentrated in states on the Eastern seaboard such as Virginia, New Hampshire, New Jersey, and Connecticut. This possibly reflects the flow of older voters and retirees to the more temperate climate.

5 Concluding Remarks

With the advance of the Internet, the last two decades have witnessed a "data revolution" in the social sciences where diverse and large data sets have become electronically available to researchers. Much of today's cutting-edge quantitative social science research results from researchers' creativity to link multiple data sets that are collected separately. In many cases, however, a unique identifier that can be used to merge multiple data sources does not exist. Currently, most social scientists rely on either deterministic or proprietary methods. Yet, deterministic methods are not robust to measurement errors and missing data, cannot quantify the uncertainty inherent in merge process, and often require arbitrary decisions from researchers. Proprietary methods, many of which are also deterministic, lack transparency and hence are not suitable for academic and policy research where reproducibility and transparency play an essential role.

Here, we advocate the use of probabilistic modeling to assist merging large-scale data sets. The main advantage of probabilistic models is their ability to estimate false positive and false negative rates that arise when linking multiple data sets. We contribute to the statistical literature of record linkage by developing a fast and scalable implementation of the canonical model. Through simulation and empirical studies, we demonstrate that the proposed methodology can quickly and reliably merge data sets even when they have millions of records.

Like any methods, however, the proposed record linkage technology has important limitations of which researchers must be aware. Most importantly, the proposed methodology is likely to have a difficult time producing high-quality matches when the overlap between two data sets is

expected to be small. As shown in our simulation studies, for these difficult merge problems, effective blocking is essential. Blocking is even more important when linking many data sets at once. Other important research question are how to merge more than two files at the same time and how to efficiently use a small amount of hand-coded data to improve the quality of record linkage. We leave these methodological challenges to future research.

References

Adena, Maja, Ruben Enikolopov, Maria Petrova, Veronica Santarosa and Ekaterina Zhuravskaya. 2015. "Radio and the Rise of The Nazis in Prewar Germany." *Quarterly Journal of Economics* 130:1885–1939.

Ansolabehere, Stephen and Eitan Hersh. 2012. "Validation: What Big Data Reveal About Survey Misreporting and the Real Electorate." *Political Analysis* 20:437–459.

Ansolabehere, Stephen and Eitan Hersh. 2017. "ADGN: An Algorithm for Record Linkage Using Address, Date of Birth, Gender and Name.".

Belin, Thomas R. and Donald B. Rubin. 1995. "A Method for Calibrating False-Match Rates in Record Linkage." *Journal of the American Statistical Association* 90:694–707.

Berent, M. K., J. A. Krosnick and A. Lupia. 2016. "Measuring Voter Registration and Turnout in Surveys. Do Official Government Records Yield More Accurate assessments?" *Public Opinion Quarterly.* 80:597–621.

Bolsen, Toby, Paul J. Ferraro and Juan Jose Miranda. 2014. "Are Voters More Likely to Contribute to Other Public Goods? Evidence from a Large-Scale Randomized Policy Experiment." *American Journal of Political Science* 58:17–30.

Bonica, Adam. 2013. "Database on Ideology, Money in Politics, and Elections: Public version 1.0 [Computer file]." Stanford, CA: Stanford University Libraries.

Cesarini, David, Erik Lindqvist, Robert Ostling and Bjorn Wallace. 2016. "Wealth, Health, and Child Development: Evidence from Administrative Data on Swedish Lottery Players." *Quarterly Journal of Economics* 131:687–738.

Christen, Peter. 2012. *Data Matching. Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection.* Springer.

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.

Cross, Philip J. and Charles F. Manski. 2002. "Regressions, Short and Long." *Econometrica* 70:357–368.

Dalzell, N. M. and J. P. Reiter. 2018. "Regression Modeling and File Matching Using Possibly Erroneous Matching Variables." *Journal of Computational and Graphical Statistics*.

de Bruin, Jonathan. 2017. "Record Linkage. Python library. Version 0.8.1." `https://recordlinkage.readthedocs.io/`.

Einav, Liran and Jonathan Levin. 2014. "Economics in the age of big data." *Science* 346.

Enamorado, Ted. 2018. "Active Learning for Probabilisitic Record Linkage." Social Science Research Network (SSRN).
URL: *https://ssrn.com/abstract=3257638*

Engbom, Niklas and Christian Moser. 2017. "Returns to Education through Access to Higher-Paying Firms: Evidence from US Matched Employer-Employee Data." *American Economic Review: Papers and Proceedings* 107:374–78.

Feigenbaum, James. 2016. "Automated Census Record Linking: A Machine Learning Approach." Boston University, technical report.
URL: *https://jamesfeigenbaum.github.io/research/pdf/census-link-ml.pdf*

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

Figlio, David and Jonathan Guryan. 2014. "The Effects of Poor Neonatal Health on Children's Cognitive Development." *American Economic Review* 104:3921–55.

Giraud-Carrier, C., J. Goodlife, B. M. Jones and S. Cueva. 2015. "Effective record linkage for mining campaign contribution data." *Knowledge and Information Systems* 45:389–416.

Goldstein, H. and K. Harron. 2015. *Methodological Developments in Data Linkage.* John Wiley & Sons, Ltd. Chapter 6: Record Linkage: A Missing Data Problem., pp. 109–124.

Gutman, R., Afendulis C. C. and Zaslavsky A. M. 2013. "A Bayesian Procedure for File Linking to End-of-Life Medical Costs." *Journal of the American Medical Informatics Association.* 103:34–47.

Gutman, R., C.J. Sammartino, T.C. Green and B.T. Montague. 2016. "Error Adjustments for File Linking Methods Using Encrypted Unique Client Identifier (eUCI) with Application to Recently Released Prisoners who are HIV+." *Statistics in Medicine* 35:115–129.

Harron, Katie, Harvey Goldstein and Chris Dibben, eds. 2015. *Methodological Developments in Data Linkage.* West Sussex: John Wiley & Sons.

Hersh, E. D. 2015. *Hacking the Electorate: How Campaigns Perceive Voters.* Cambridge, U.K.: Cambridge University Press.

Herzog, Thomas H., Fritz Scheuren and William E. Winkler. 2010. "Record Linkage." *Wiley Interdisciplinary Reviews: Computational Statistics* 2:535–43.

Herzog, Thomas N., Fritz J. Scheuren and William E. Winkler. 2007. *Data Quality and Record Linkage Techniques.* New York: Springer.

Hill, Seth. 2017. "Changing Votes or Changing Voters: How Candidates and Election Context Swing Voters and Mobilize the Base." *Electoral Studies* 48:131–148.

Hill, Seth J. and Gregory A. Huber. 2017. "Representativeness and Motivations of the Contemporary Donorate: Results from Merged Survey and Administrative Records." *Political Behavior* 39:3–29.

Hof, M. H. P. and A.H. Zwinderman. 2012. "Methods for analyzing data from probabilistic linkage strategies based on partially identifying variables." *Statistics in Medicine* 31:4231–4242.

Imai, Kosuke and Dustin Tingley. 2012. "A Statistical Method for Empirical Testing of Competing Theories." *American Journal of Political Science* 56:218–236.

Jaro, Matthew. 1972. UNIMATCH-A Computer System for Generalized Record Linkage Under Conditions of Uncertainty. Technical Report. Spring Joint Computer Conference.

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.

Jutte, Douglas P., Leslie L. Roos and Marni D. Browne. 2011. "Administrative Record Linkage as a Tool for Public Health Research." *Annual Review of Public Health* 32:91–108.

Kim, Gunky and Raymond Chambers. 2012. "Regression analysis under incomplete linkage." *Computational Statistics and Data Analysis* 56:2756–2770.

Lahiri, P. and Michael D. Larsen. 2005. "Regression Analysis with Linked Data." *Journal of the American Statistical Association* 100:222–230.

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

McLaughlan, Geoffrey and David Peel. 2000. *Finite Mixture Models.* New York: John Wiley & Sons.

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

Meredith, M. and M. Morse. 2014. "Do Voting Rights Notification Laws Increase Ex-Felon Turnout?" *The ANNALS of the American Academy of Political and Social Science* 651:220–249.

Mummolo, J. and C. Nall. 2016. "Why Partisans Don't Sort: The Constraints on Political Segregation." *Journal of Politics* 79:45–59.

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

Neter, John, E. Scott Maynes and R. Ramanathan. 1965. "The Effect of Mismatching on the Measurement of Resopnse Errors." *Journal of the American Statistical Association* 60:1005–1027.

Ong, Toan C., Michael V. Mannino., Lisa M. Schilling and Michael G. Kahn. 2014. "Improving Record Linkage performance in the Presence of Missing Linkage Data." *Journal of Biomedical Informatics.* 52:43–54.

Richman, Jesse T., Gulshan A. Chattha and David C. Earnest. 2014. "Do non-citizens vote in U.S. elections?" *Electoral Studies* 36:149–157.

Ridder, Geert and Robert Moffitt. 2007. *Handbook of Econometrics.* Vol. 6 Elsevier Chapter The Econometrics of Data Combination, pp. 5469–5547.

Sadinle, Mauricio. 2014. "Detecting Duplicates in a Homicide Registry Using a Bayesian Partitioning Approach." *Annals of Applied Statistics.* 8:2404–2434.

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

Sariyar, Murat and Andreas Borg. 2016. "Record Linkage in R. R package. Version 0.4-10." `http://cran.r-project.org/package=RecordLinkage`.

Sariyar, Murat, Andreas Borg and K. Pommerening. 2012. "Missing Values in Deduplication of Electronic Patient Data." *Journal of the American Medical Informatics Association.* 19:e76–e82.

Scheuren, Fritz and William E. Winkler. 1993. "Regression Analysis of Data Files that are Computer Matched." *Survey Methodology* 19:39–58.

Scheuren, Fritz and William E. Winkler. 1997. "Regression Analysis of Data Files That Are Computer Matched II." *Survey Methodology.* 23:157–165.

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

Steorts, Rebecca C., Samuel L. Ventura, Mauricio Sadinle and Stephen E. Fienberg. 2014. A Comparison of Blocking Methods for Record Linkage. In *Lecture Notes in Computer Science.* Vol. 8744 Privacy in Statistical Databases pp. 253–268.

Tam Cho, W., J. Gimpel and I. Hui. 2013. "Voter Migration and the Geographic Sorting of the American Electorate." *Annals of the American Association of Geographers* 103:856–870.

Tancredi, A. and B. Liseo. 2011. "A Hierachical Bayesian Approach to Record Linkage and Population Size Problems." *Annals of Applied Statistics.* 5:1553–1585.

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

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. 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.

Winkler, William E. 2005. Approximate String Comparator Search Strategies for Very Large Administrative Lists. Research Report Series (Statistics) No. 2005-02. Statistical Research Division U.S. Census Bureau.

Winkler, William E. 2006a. Automatic Estimation of Record Linkage False Match Rates. In *Proceedings of the Section on Survey Research Methods.* American Statistical Association.

Winkler, William E. 2006b. Overview of record linkage and current research directions. Technical Report. United States Bureau of the Census.

Winkler, William E. and Willian Yancey. 2006. Record Linkage Error-Rate Estimation without Training Data. In *Proceedings of the Section on Survey Research Methods.* American Statistical Association.

Winkler, William E., Willian Yancey and E. H. Porter. 2010. Fast Record Linkage of Very Large Files in Support of the Decennial and Administrative Record Projects. In *Proceedings of the Secion on Survet Research Methods.*

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

A Appendix: Computationally Efficient Implementation

In this appendix, we describe the details of our computationally efficient implementation of the canonical probabilistic record linkage model.

A.1 Reverse Data Structures for Field Comparisons

The critical step in record linkage is to compare pairs of records across the K fields used to link two datasets, which is often regarded as the most expensive step in terms of computational time (Christen, 2012). To do so, for each linkage field k, we first compare observation i of dataset $\mathcal{A}$ and j from dataset $\mathcal{B}$ via a pre-defined distance metric (e.g., Jaro-Winkler for string-valued fields) and obtain a value which we call $S_k(i,j)$. However, comparisons in the Fellegi-Sunter model are represented in terms of a discrete agreement levels per linkage field, not a continuous measure of agreement as the one implied by the distance metric. In other words, we need a discrete representation of $S_k(i,j)$. Specifically, if we have a total of L_k agreement levels for the kth variable, then,

$$\gamma_k(i,j) \quad = \quad \begin{cases} 0 & \text{if } S_k(i,j) \leq \tau_0 \\ 1 & \text{if } \tau_0 < S_k(i,j) \leq \tau_1 \\ \vdots & \\ L_k - 1 & \text{if } \tau_{L_k-2} < S_k(i,j) \leq \tau_{L_k-1} \end{cases} \tag{14}$$

where $\gamma_k(i,j)$ represents the agreement level between the values for variable k for the pair (i,j) and $\boldsymbol{\tau} = \{\tau_0, \tau_1, \ldots, \tau_{L_{k-1}}\}$ the set of predetermined thresholds use to define the agreement levels. For example, to compare names and last names, some authors such as Winkler (1990) argue in favor of using the Jaro-Winkler string distance to produce S_k, where one could use $\boldsymbol{\tau} = \{0.88, 0.94\}$ to construct γ_k for three agreement levels.

Still the problem with constructing γ_k is that the number of comparisons we have to make is often large. In our proposed implementation we exploit the following characteristics of typical record linkage problems in social sciences:

- The number of unique values observed in each linkage field is often less than the number of observations in each dataset. For example, consider a variable such as first name. Naively, one may compare the first name of each observation in dataset $\mathcal{A}$ with that of every obser-

vation in $\mathcal{B}$. In practice, however, we can reduce the number of comparisons by considering only unique first name that appears in each data set. The same trick can be used for all linkage fields by focusing on the comparison of the unique values of each variable.

- For each comparison between two unique first names ($name_{1,\mathcal{A}}$ and $name_{1,\mathcal{B}}$), for example, we only keep the indices of the original datasets and store them using what is often referred as a reverse data structure in the literature (Christen, 2012). In such an arrangement, a pair of names ($name_{1,\mathcal{A}}$, $name_{1,\mathcal{B}}$) becomes a key with two lists, one containing the indices from dataset $\mathcal{A}$ that have a first name equal to $name_{1,\mathcal{A}}$, and another list that does the same for $name_{1,\mathcal{B}}$ in dataset $\mathcal{B}$.

- Comparisons involving a missing value need not be made. Instead, we only need to store the indices of the observations in $\mathcal{A}$ and $\mathcal{B}$ that contain missing information for field k.

- Since the agreement levels are mutually exclusive, we use the lowest agreement level as the base category. Once a set of threshold values has been defined, then a pair of names can only be categorized in one of the L_k agreement levels. The indices for the the pairs of values that can be categorized as disagreements (or nearly disagreements) do not need to be stored. For most variables, disagreement is the category that encompasses the largest number of pairs. Thus, our reverse data structure lists become quite sparse. This sparsity can be exploited by the use of sparse matrix, yielding a substantially memory efficient implementation.

A.2 Sparse Matrix Representation of Hash Tables to Count Agreement Patterns

Next, we describe our computationally efficient implementation of the Fellegi-Sunter model via the EM algorithm (see Online SI B for the exact algorithm we use). First, for implementing the E-step, notice that the match probability given in equation (5) takes the same value for two pairs if their agreement patterns are identical. For the sake of illustration, consider a simple example where two variables are used for merging, i.e., $K = 2$, and binary comparison is made for each variable, i.e., $L_k = 2$. Under this setting, there are a total of nine agreement patterns: $(\texttt{0}, \texttt{0})$, $(\texttt{0}, \texttt{1})$, $(\texttt{1}, \texttt{0})$, $(\texttt{1}, \texttt{1})$, $(\texttt{NA}, \texttt{0})$, $(\texttt{NA}, \texttt{1})$, $(\texttt{0}, \texttt{NA})$, $(\texttt{1}, \texttt{NA})$, and $(\texttt{NA}, \texttt{NA})$ where `1` and `0` represent agreement and disagreement, respectively while `NA` represents a missing value. Then, for instance,

the match probability for $(\texttt{0}, \texttt{1})$ is given by $\lambda\pi_{110}\pi_{211}/\{\lambda\pi_{110}\pi_{211} + (1-\lambda)\pi_{100}\pi_{201}\}$ whereas that for $(\texttt{1}, \texttt{NA})$ is equal to $\lambda\pi_{111}/\{\lambda\pi_{111} + (1-\lambda)\pi_{101}\}$. If all comparison values are missing, e.g., $(\texttt{NA}, \texttt{NA})$, then we set the match probability to λ. Thus, the E-step can be implemented by computing the match probability for each of the *realized* agreement patterns. Often, the total number of realized agreement patterns is much smaller than the number of all *possible* agreement patterns.

Second, the M-step defined in equations (S1) and (S2) requires the summation of match probabilities across all pairs or their subset. Since this probability is identical within each agreement pattern, all we have to do is to count the total number of pairs that have each agreement pattern. In other words, the number of pairs per agreement pattern becomes our sufficient statistic. We use the following hash function for efficient counting,[18]

$$\mathbf{H} \;=\; \sum_{k=1}^{K} \mathbf{H}_k \quad \text{where} \quad \mathbf{H}_k \;=\; \begin{bmatrix} h_k^{(1,1)} & h_k^{(1,2)} & \dots & h_k^{(1,N_{\mathcal{B}})} \\ \vdots & \vdots & \ddots & \vdots \\ h_k^{(N_{\mathcal{A}},1)} & h_k^{(N_{\mathcal{A}},2)} & \dots & h_k^{(N_{\mathcal{A}},N_{\mathcal{B}})} \end{bmatrix} \tag{15}$$

where $h_k^{(i,j)} \;=\; \mathbf{1}\left\{\gamma_k(i,j) > 0\right\} 2^{\gamma_k(i,j) + (k-1) \times L_k}$. The matrix $\mathbf{H}_k$ maps each pair of records to a corresponding agreement pattern in the kth variable that is represented by a unique hash value based on the powers of 2. These hash values are chosen such that the matrix $\mathbf{H}$ links each pair to the corresponding agreement pattern across K variables.

Since an overwhelming majority of pairs do not agree in any of the linkage fields, most elements of the $\mathbf{H}_k$ matrix are zero. As a result, the $\mathbf{H}$ matrix also has many zeros. In our implementation, we utilize sparse matrices whose lookup time is $O(P)$ where P is the number of unique agreement patterns observed. In most applications, P is much less than the total number of possible agreement patterns, i.e., $\prod_{k=1}^{K} L_k$. This hashing technique is applicable if the number of variables used for merge is moderate. If many variables are used for the merge, approximate hashing techniques such as min hashing and locally sensitive hashing are necessary.

[18] Since the work of Jaro (1972), the use of table-like objects to store agreement patterns has been recognized as an important step to improve computational efficiency. Our contribution goes beyond by tying together, under a unified framework, reverse data structures and novel use of a sparse matrix representation of a hash table to store agreement patterns.