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

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

July 31, 2017

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 sets may contain missing and inaccurate information. These problems are severe especially when merging large-scale administrative records. The existing algorithms to automate the merging process do not scale, fail to identify many matches, and require arbitrary decisions by researchers. We develop a faster and more scalable algorithm to implement the canonical probabilistic model of record linkage. The proposed methodology can efficiently handle 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 merge campaign contribution records, survey data, and nationwide voter files. Open-source software 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 and Seth Hill for useful comments.

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

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

[§]Professor, Department of Politics and Center for Statistics and Machine Learning, Princeton University. Professor of Visiting Status, Graduate Schools of Law and Politics, The University of Tokyo. Phone: 609–258–6601, Email: `kimai@princeton.edu`, URL: http://imai.princeton.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 Survey (ANES) validates self-reported turnout by merging their survey data with a nationwide voter file where only the matched respondents are treated as registered voters. A similar turnout validation procedure is used for the Cooperative Congressional Election Study (CCES). 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 may be due to the failure of the merge procedure rather than social desirability bias.

Merging data sets is straightforward if there exists a unique identifier that unambiguously links records from different data sets. Unfortunately, in practice, such a unique identifier is often unavailable. Under these circumstances, some researchers have used a deterministic algorithm to automate the merging process (e.g., DellaVigna and Kaplan, 2007; 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, 2016; Berent, Krosnick and Lupia, 2016; Cesarini et al., 2016; Fouirnaies and Hall, Forthcoming; Hill, Forthcoming) 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; Enos and Fowler, 2016; 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, and as a result may yield many false negatives. 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. This also means that post-merge data analyses cannot account for the uncertainty of the merging procedure, leading to the bias due to measurement error. These methodological challenges are amplified especially when merging large data sets such as administrative records.

In this paper, 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 any merge process. Therefore, they offer a principled way to calibrate and account for two types of errors, i.e., false positives and false negatives, when merging data sets without a unique identifier. Unfortunately, while there exists a well-known statistics literature on probabilistic record linkage (e.g., Winkler, 2006; Herzog, Scheuren and Winkler, 2007; Harron, Goldstein and Dibben, 2015), the current implementation does not scale to large data sets that are commonly used in today's social science research. We use a hashing technique to develop a faster and more scalable implementation of the canonical probabilistic record linkage model originally proposed by Fellegi and Sunter (1969). Together with parallelization and random sampling, 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. In addition, building on the prior methodological literature (e.g., Winkler, 2000; Lahiri and Larsen, 2005), we show (1) how to incorporate auxiliary information such as population name frequency and migration rates into the merging procedure and (2) how to conduct post-merge analyses while accounting for the uncertainty about the merge process. All of these methodological developments are described in Section 2.

In Section 3, we conduct a comprehensive set of simulation studies to evaluate the robustness of the proposed methodology fastLink to several factors including the size of data sets, the proportion of true matches, measurement error, and missing data proportion and mechanisms. Through a total of 270 different simulation settings, we consistently find 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, we show that fastLink performs at least as well as recently proposed probabilistic approaches (Steorts, 2015; Sadinle, 2017). Importantly, our merge method is fast and scales to larg data sets when compared to these state-of-art models, which rely on Markov chain Monte Carlo estimation methods.

In Section 4, we present two empirical applications. First, we revisit the study of Hill and Huber (2017) who examine the ideological differences between donors and non-donors by merging the 2012 CCES data set, which has more than 50,000 respondents, with the Database on Ideology,

Money in Politics, and Elections (DIME Bonica (2013)), which has over 5 million donor records. We compare our merge results with those of a proprietary method, which is used by the original authors. We find that the matches identified by fastLink are at least as high-quality as those identified by the proprietary method. We also improve the original analysis by incorporating the uncertainty of the merge process in the post-merge analysis. We find 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. This represents one of the largest data merges ever conducted in the social sciences.[1] By merging voter files over time, scholars can study the causes and consequences of partisan residential segregation (e.g., Bishop and Cushing, 2008; Tam Cho, Gimpel and Hui, 2013; Mummolo and Nall, 2016) and political analytics professionals can develop effective micro-targeting strategies (e.g., Nickerson and Rogers, 2014; 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 approximately 30 percentage points greater than the exact matching method. Although it is more difficult to find across-state movers, we are able to find approximately 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 describe the proposed methodology. We first introduce the canonical probabilistic model of record linkage originally proposed by Fellegi and Sunter (1969). We then describe several improvements we make to this model, including a faster and more scalable implementa-

[1] While large-scale data merges have been used in social science research (e.g., Hersh, 2015), they are based on proprietary algorithms. Other large-scale data merges such as Ansolabehere and Hersh (2016) and Tam Cho, Gimpel and Hui (2013) match datasets of several million voters each, but neither of these studies approaches the scale of our own exercise.

tion, the use of auxiliary information to inform parameter estimation, and the incorporation of uncertainty about 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. We consider all possible pair-wise comparison 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 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 also 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 $\delta_k(i,j)$ is equal to 1 if at least one of the two records has a missing value in the kth variable and is equal to 0 if both records have observed values.

Table 1 presents an illustrative example of agreement patterns based on two artificial data sets, $\mathcal{A}$ and $\mathcal{B}$, each of which has three 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 agreement 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). The Jaro-Winkler string distance between strings s_1 and s_2, which ranges from 0 to 1, is

	Name				Address	
	First	Middle	Last	Date of birth	House	Street
Data set $\mathcal{A}$						
1	James	V	Smith	04-10-1934	780	Devereux St.
2	John	NA	Martin	01-15-1942	780	Devereux St.
3	Robert	NA	Martines	02-03-1956	60	16th St.
Data set $\mathcal{B}$						
1	Michael	F	Martinez	08-07-1928	4	16th St.
2	James	NA	Smith	04-10-1934	780	Dvereuux St.
3	William	V	Smithson	09-14-1950	12	Hibben Magie Rd
Agreement patterns						
$\mathcal{A}.1 - \mathcal{B}.1$	different	different	different	different	different	different
$\mathcal{A}.1 - \mathcal{B}.2$	identical	NA	identical	identical	identical	similar
$\mathcal{A}.1 - \mathcal{B}.3$	different	identical	similar	different	different	different
$\mathcal{A}.2 - \mathcal{B}.1$	different	NA	similar	different	different	different
$\mathcal{A}.2 - \mathcal{B}.2$	different	NA	different	different	identical	similar
$\mathcal{A}.2 - \mathcal{B}.3$	different	NA	different	different	different	different
$\mathcal{A}.3 - \mathcal{B}.1$	different	NA	similar	different	different	different
$\mathcal{A}.3 - \mathcal{B}.2$	different	NA	different	different	different	different
$\mathcal{A}.3 - \mathcal{B}.3$	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 three records. The bottom panel shows the agreement patterns for all possible pairs of these records. For example, the third line of the agreement pattern compares the first record of the data set $\mathcal{A}$ with the third record of the data set $\mathcal{B}$. These two records have an identical middle name and a similar last name, but have different values for the other variables. Missing values are indicated by NA.

defined as,

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

where

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

where $|s|$ represents that length of string s, m is the number of characters in common between the two strings, t is the number of transpositions between the common characters, $\ell \in [0, 4]$ is the number of consecutive characters in common at the beginning of the two strings, and $w \in [0, 0.25]$ is the weight given to ℓ. For example, if we consider two last names, $s_1 =$ Smith and $s_2 =$ Martinez, we have that $m = 3$ (the letters: m, i, t), $|s_1| = 5$, and $|s_2| = 8$, and $t = 2$. If we set $\ell = 4$ and $w = 0.1$, as often done in practice (see e.g., Winkler, 1990; Cohen, Ravikumar and Fienberg,

2003), then the Jaro-Winkler distance for these two strings equals 0.55.

Since the Jaro-Winkler distance is a continuous measure, 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 discretize the Jaro-Winkler distance into three categories (i.e., `different`, `similar`, and (nearly) `identical`) using 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 with the Jaro-Winkler distance of 0.88 whereas `Smith` and `Martinez` are different. We now explain how a probabilistic model based on these agreement patterns can assist researchers merging data sets.

2.2 The Canonical Probabilistic Model of Record Linkage

2.2.1 The Model and Assumptions

We start with the description of the most commonly used probabilistic model of record linkage that is originally proposed by Fellegi and Sunter (1969). The model is the following simple finite mixture model (e.g., McLaughlan and Peel, 2000; Imai and Tingley, 2012), in which the latent mixing variable M_{ij} indicates whether a pair of records (the ith record in the data set $\mathcal{A}$ and the jth record in the data set $\mathcal{B}$) represents a match,

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

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

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. Since $\boldsymbol{\pi}_{k0}$ may not be zero, 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 although it is otherwise completely nonparametric. 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 unscalable. 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/non-match status. Again, this assumption can be relaxed in principle but we maintain it throughout this paper so that the model can be applied to large data sets. In our simulation studies, we find that the model performs well even when these independence assumptions are violated (see Section 3).

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)). Other adhoc imputation procedures also exist but none of them has a theoretical justification or appears to perform well in practice (Sariyar, Borg and Pommerening, 2012). Employing such techniques is problematic because a true match can contain missing values. To address this problem, following Sadinle (2014, 2017), we assume that data are missing at random (MAR) conditional on the latent variable M_{ij},

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

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 and hence 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., $\pi_{km\ell} = \Pr(\gamma_k(i,j) = \ell \mid M_{ij} = m)$.

2.2.2 The Uncertainty of Merge Process

The attractiveness of this probabilistic model is its ability to quantify the uncertainty inherent in merging. Once the model parameters are estimated, we can compute the posterior match probability for each pair using the Bayes rule,

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

In Section 2.5, we show how to incorporate this posterior match probability for post-merge analysis based on regression models so that the uncertainty of the merge process is properly reflected in the post-merge analysis.

While in theory a post-merge analysis can use all pairs with their corresponding posterior match probabilities, it is often more convenient to determine a threshold S in order to create a merged data set. Such an approach is useful in practice especially when the data sets to be merged are large. Specifically, we call a pair (i, j) as a match if the posterior match probability ξ_{ij} exceeds the threshold S chosen by a researcher. There is a clear trade-off in the choice of this threshold value. A large value of S ensures that most of the selected pairs are correct matches but may lead to a failure to identify many true matches. In contrast, if we lower the value of the threshold too much, we will select more pairs but many of them may not be true matches. Therefore, it is important to quantify the degree to which these matching errors occur in the merging process.

Fortunately, the probabilistic model can estimate the false discovery rate (FDR) and the false negative rate (FNR). The FDR represents the proportion of false matches among the selected pairs whose posterior matching probability is greater than or equal to the chosen threshold. We can estimate the FDR using our model parameters as follows,

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

In addition, we can also estimate the FNR, which represents the proportion of true matches that are not selected,

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

In practice, researchers may select the threshold value S such that the FDR is sufficiently small. Nevertheless, it is important to recognize another type of error where a strict threshold can lead to many false negatives. In our simulations and empirical applications, we find that the results are not particularly sensitive to the choice of threshold value. However, in other applications, the further calibration of error rates may be necessary (see e.g., Belin and Rubin, 1995; Larsen and Rubin, 2001).

2.2.3 Enforcing One-to-One Merge

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 posterior match probabilities. In some cases, multiple observations may have an identical value of posterior match probability, i.e., $\xi_{ij} = \xi_{ij'}$ with $j \neq j'$. If researchers wish to enforce a constraint that each record $\mathcal{A}$ is only matched at most with one record in the data set $\mathcal{B}$, there are several possible procedures that can be used. First, researchers may choose the record with the greatest posterior match probability where a tie is broken with random sampling. Second, researchers may randomly sample a record with the selection probability proportional to the posterior match probability. However, these two procedures may end up matching multiple records in $\mathcal{A}$ with an identical record in the data set $\mathcal{B}$, which may not be desirable in some cases. To avoid this problem, matching can be done without replacement but this means that the results may depend on the order, in which matching is done for each record.

For moderate or small data sets, many researchers use the method proposed by Jaro (1989), which resolves the problem of duplicate matches by maximizing the sum of posterior probabilities for matched pairs. Jaro (1989) shows that this can be formulated as the linear sum assignment problem, for which an efficient algorithm exists. However, we find that the algorithm is prohibitively slow for large data sets. Thus, we reduce the number of assignments by focusing on those pairs whose posterior probabilities are greater than a threshold c, which could be set, for example, to 0.85. Specifically, suppose that there are D_c pairs with $\xi_{ij} \geq c$ and let $\mathcal{A}_c$ and $\mathcal{B}_c$ represent the set of unique observations in the data set $\mathcal{A}$ and $\mathcal{B}$, respectively, that belong to at least one of these pairs. Then, the assignment problem can be written as,

$$\begin{aligned} \text{maximize} \sum_{i \in \mathcal{A}_c} \sum_{j \in \mathcal{B}_c} \xi_{ij} M_{ij} \quad \text{subject to} \quad & \sum_{i \in \mathcal{A}_c} M_{ij} \;\leq\; 1 \text{ for each } j \in \mathcal{B}_c, \\ \text{and} \quad & \sum_{j \in \mathcal{B}_c} M_{ij} \;\leq\; 1 \text{ for each } i \in \mathcal{A}_c \end{aligned} \tag{7}$$

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

2.3 Computationally Efficient Implementation

2.3.1 The EM Algorithm

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

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

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

$$\lambda = \frac{1}{N_{\mathcal{A}} N_{\mathcal{B}}} \sum_{i=1}^{N_{\mathcal{A}}} \sum_{j=1}^{N_{\mathcal{B}}} \xi_{ij} \tag{8}$$

$$\pi_{km\ell} = \frac{\sum_{i=1}^{N_{\mathcal{A}}} \sum_{j=1}^{N_{\mathcal{B}}} \mathbf{1}\{\gamma_k(i,j) = l)\}(1-\delta_k(i,j))\xi_{ij}^m (1-\xi_{ij})^{1-m}}{\sum_{i=1}^{N_{\mathcal{A}}} \sum_{j=1}^{N_{\mathcal{B}}} (1-\delta_k(i,j))\xi_{ij}^m (1-\xi_{ij})^{1-m}} \tag{9}$$

Then with a suitable set of starting values, we repeat the E-step and M-step until convergence. When setting the starting values for the model parameters, we impose inequality constraints based on the following two ideas: (1) the set of matches is strictly smaller than the set on non-matches $\lambda \ll 1 - \lambda$, and (2) for binary comparisons, we have $\pi_{k10} \ll \pi_{k11}$ and $\pi_{k01} \ll \pi_{k00}$ for each k (Jaro, 1989; Winkler, 1993; Sadinle and Fienberg, 2013). The latter implies that agreement (disagreement) is more likely among matches (non-matches).

While the EM algorithm described above is relatively simple, as shown below, we find that the existing implementation is computationally inefficient. In particular, as shown later in this section, the open-source software packages available in R (Borg and Sariyar, 2016) and python (de Bruin, 2017) do not scale if the data sets to be merged contain more than tens of thousands of records each.

2.3.2 Hashing for Efficient Memory Management

To overcome this challenge, we develop a computationally efficient implementation of the EM algorithm. First, for implementing the E-step, notice that the posterior 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 posterior 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 posterior match probability to λ. Thus, the E-step can be implemented by computing the posterior 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 (8) and (9) requires the summation of posterior 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. We use the following hash function for efficient counting,

$$\mathbf{H} \;=\; \sum_{k=1}^{K} \mathbf{H}_k \quad \text{where} \quad \mathbf{H}_k \;=\; \begin{bmatrix} h_k^{(1,1)} & h_k^{(1,2)} & \dots & h_k^{(1,N_{\mathcal{B}})} \\ \vdots & \vdots & \ddots & \vdots \\ h_k^{(N_{\mathcal{A}},1)} & h_k^{(N_{\mathcal{A}},2)} & \dots & h_k^{(N_{\mathcal{A}},N_{\mathcal{B}})} \end{bmatrix} \tag{10}$$

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 are not true matches, 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 $O(T)$ where T is the number of unique agreement patterns observed. In most applications, T 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.

2.3.3 Parallelization and Random Sampling

Under the proposed probabilistic modeling approach, a vast majority of the computational burden is due to the enumeration of agreement patterns. In fact, the actual computation time of implementing the E and M steps, once hashing is done, is fast even for large data sets. Therefore, for further computational efficiency, we parallelize the enumeration of agreement patterns. Specifically, we take a divide-and-conquer approach by partitioning the two data sets, $\mathcal{A}$ and $\mathcal{B}$, into equally-sized subsets such that $\mathcal{A} = \{\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_{M_\mathcal{A}}\}$ and $\mathcal{B} = \{\mathcal{B}_1, \mathcal{B}_2, \ldots, \mathcal{B}_{M_\mathcal{B}}\}$. Then, using OpenMP, we count the agreement patterns for each partition pair $\{\mathcal{A}_i, \mathcal{B}_j\}$ in parallel using the hash function given in equation (10). As explained above, we utilize sparse matrix objects to efficiently store agreement patterns for all pairwise comparisons. Finally, the entire pattern-counting procedure is implemented in C++ for additional performance gains. Taken together, our approach provides simple parallelization of the pattern-counting procedure and efficient memory usage so that our linkage procedure can be applied to arbitrarily large problems.

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

2.4 Incorporating Auxiliary Information

Another advantage of the probabilistic model introduced above is that we can incorporate auxiliary information. In our view, this point has not been emphasized enough in the literature. We consider two settings. We first describe how to adjust for the fact that some names are more common than others. We then consider how to incorporate aggregate information about migration.

2.4.1 Incorporating Name Frequencies

Some names are more common than others and as such they are more likely to contribute to false matches. Here, we consider how to incorporate name frequencies into the calculation of

posterior match probabilities. It is difficult to incorporate this information directly into the model and estimation without significantly increasing computational burden. Therefore, inspired by the work of Fellegi and Sunter (1969) and Winkler (2000), we make an ex-post adjustment to the posterior match probabilities rather than the likelihood ratio as done in the previous approaches. Unlike Winkler (2000), however, we do not assume that the frequency of a name in the set of true matches is equal to the frequency of the same name in a smaller data set.

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

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

for $m = 0, 1$ and $p = 1, 2, \ldots, P$. The assumption may be violated, for example, if certain names are more likely to be spelled incorrectly (even after conditioning on the true match status). Under this assumption, we have,

$$\begin{aligned} & \Pr(\gamma_{\text{name}}(i,j) = 1, \text{name}_i = \text{name}_j = p \mid M_{ij} = m) \\ = \;& \Pr(\gamma_{\text{name}}(i,j) = 1 \mid \text{name}_i = \text{name}_j, M_{ij} = m) \times \Pr(\text{name}_i = \text{name}_j = p \mid M_{ij} = m) \end{aligned} \tag{11}$$

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

Under this setting, we can make an ex-post adjustment to the posterior match probability given in equation (4) for any given pair whose records have an identical name. First, we have,

$$\begin{aligned} & \Pr(M_{ij} = 1 \mid \boldsymbol{\delta}(i,j), \boldsymbol{\gamma}(i,j), \text{name}_i = \text{name}_j = p) \\ = \;& \frac{\lambda \cdot \Pr(\boldsymbol{\gamma}(i,j), \boldsymbol{\delta}(i,j), \text{name}_i = \text{name}_j = p \mid M_{ij} = 1)}{\sum_{m=0}^{1} \lambda^m (1-\lambda)^{1-m} \cdot \Pr(\boldsymbol{\gamma}(i,j), \boldsymbol{\delta}(i,j), \text{name}_i = \text{name}_j = p \mid M_{ij} = m)} \end{aligned} \tag{12}$$

where the conditional independence assumption given in equation (11) implies,

$$\Pr(\boldsymbol{\gamma}(i,j), \boldsymbol{\delta}(i,j), \text{name}_i = \text{name}_j = p \mid M_{ij} = m)$$

$$= \Pr(\text{name}_i = \text{name}_j = p \mid M_{ij} = m) \times \prod_{k=1}^{K} \left(\prod_{\ell=0}^{L_k - 1} \pi_{km\ell}^{\mathbf{1}\{\gamma_k(i,j)=\ell\}} \right)^{1-\delta_k(i,j)} \tag{13}$$

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

$$\begin{aligned} & \Pr(\text{name}_i = \text{name}_j = p \mid M_{ij} = m) \\ = \; & \frac{\sum_{i'=1}^{N_{\mathcal{A}}} \sum_{j'=1}^{N_{\mathcal{B}}} \xi_{i'j'}^{m} (1 - \xi_{i'j'})^{1-m} \mathbf{1}\{\text{name}_{i'} = \text{name}_{j'} = p\}}{\sum_{i'=1}^{N_{\mathcal{A}}} \sum_{j'=1}^{N_{\mathcal{B}}} \mathbf{1}\{\text{name}_{i'} = \text{name}_{j'} = p\}} \times \frac{f_p^{\mathcal{A}} f_p^{\mathcal{B}}}{N_{\mathcal{A}} \times N_{\mathcal{B}}} \times \frac{1}{\lambda^m (1-\lambda)^{1-m}} \end{aligned} \tag{14}$$

If we substitute equations (13) and (14) into equation (12), we obtain,

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

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

2.4.2 Incorporating Aggregate Migration Rates

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

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

$$\lambda^{\text{prior}} \;=\; \frac{\#\text{ of non-movers} + \#\text{ of within-state movers}}{N_{\mathcal{A}} \times N_{\mathcal{B}}} \tag{15}$$

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

$$\pi_{\text{adr},1,0}^{\text{prior}} \;=\; \frac{\#\text{ of in-state movers}}{\#\text{ of in-state movers} + \#\text{ of non-movers}} \tag{16}$$

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

We use the above prior mean for the conjugate prior distributions on λ, i.e., $\text{Beta}(a_\lambda, b_\lambda)$, and on $\pi_{\text{adr},1,0}$, i.e., $\text{Beta}(a_{\text{adr}}, b_{\text{adr}})$, while leaving the priors for the other parameters improper. To specify these two hyperparameters of each prior distribution, we require users to specify the weight they would attach to the prior information relative to the data as well as the prior means given in equations (15) and (16). Under the current setting, we use w_λ and w_{adr} to denote these weights for λ and $\pi_{\text{adr},1,0}$, which range from zero to one. For example, if we specify $w_\lambda = w_{\text{adr}} = 0.9$, then the resulting estimates of λ and $\pi_{\text{adr},1,0}$ will be approximately equal to the weighted averages of their corresponding ML estimates and the prior mean where the former accounts for 10% and the latter makes up 90%. Appendix A.1 describes the details of the EM algorithm that integrates this prior information.

2.5 Post-merge Analysis

Finally, we discuss how to conduct a statistical analysis after merging is complete. One advantage of the probabilistic model 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).

Even today, most of the literature on record linkage has focused on the linkage process itself, while just a few works have either connected the results obtained from Fellegi-Sunter model to subsequent statistical analysis using the linked data and/or have described under what conditions

such a connection would lead to valid estimates. In this paper, we build on the existing works on post-merge regression analysis whose goal is to eliminate possible biases due to the linkage process (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.

2.5.1 The Merged Variable as an Outcome Variable

We first consider the case where 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 taken 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 posterior 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 posterior 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 posterior 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 deduplication 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 can compute the posterior 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.,

$$Z_i^* \perp\!\!\!\perp \mathbf{X}_i \mid (\boldsymbol{\delta}, \boldsymbol{\gamma}) \tag{17}$$

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^* is

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

$$\mathbb{E}(Z_i^* \mid \mathbf{X}_i) \;=\; \mathbb{E}\{\mathbb{E}(Z_i^* \mid \boldsymbol{\gamma}, \boldsymbol{\delta}, \mathbf{X}_i) \mid \mathbf{X}_i\} \;=\; \mathbb{E}(\zeta_i \mid \mathbf{X}_i) \tag{18}$$

2.5.2 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,

$$Y_i \;=\; \alpha + \beta Z_i^* + \eta^\top \mathbf{X}_i + \epsilon_i \tag{19}$$

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 (17). In addition, we assume that the merging variables are independent of the outcome variable conditional on the explanatory variables $\mathbf{Z}^*$ and $\mathbf{X}$, i.e.,

$$Y_i \perp\!\!\!\perp (\boldsymbol{\gamma}, \boldsymbol{\delta}) \mid \mathbf{Z}^*, \mathbf{X}. \tag{20}$$

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

$$\begin{aligned} \mathbb{E}(Y_i \mid \boldsymbol{\gamma}, \boldsymbol{\delta}, \mathbf{X}_i) &= \alpha + \beta \mathbb{E}(Z_i^* \mid \boldsymbol{\gamma}, \boldsymbol{\delta}, \mathbf{X}_i) + \eta^\top \mathbf{X}_i + \mathbb{E}(\epsilon_i \mid \boldsymbol{\gamma}, \boldsymbol{\delta}, \mathbf{X}_i) \\ &= \alpha + \beta \zeta_i + \eta^\top \mathbf{X}_i \end{aligned} \tag{21}$$

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

We can 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),

$$Y_i \mid Z_i^*, \mathbf{X}_i \;\overset{\text{indep.}}{\sim}\; P_\theta(Y_i \mid Z_i^*, \mathbf{X}_i) \tag{22}$$

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

$$\hat{\theta} \;=\; \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{N_{\mathcal{A}}} \sum_{j=1}^{N_{\mathcal{B}}} \xi_{ij}^{*} \log P_{\theta}(Y_i \mid Z_i^{*} = Z_j, \mathbf{X}_i) \tag{23}$$

where $\xi_{ij}^{*} = \xi_{ij} / \sum_{j'=1}^{N_{\mathcal{B}}} \xi_{ij'}$ (if deduplication is done, then we have $\xi_{ij}^{*} = M_{ij}^{*}\xi_{ij} / \sum_{j'=1}^{N_{\mathcal{B}}} M_{ij'}^{*}\xi_{ij'}$ 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 $\mathcal{B}$). Under the two assumptions described earlier and mild regularity conditions, it can be shown that the sample average of this weighted log-likelihood function converges to the expected value of the original log-likelihood function (see Appendix A.2). This result implies that the parameter estimator given in equation (23) is consistent. Finally, because we are considering large data sets, we can typically ignore the uncertainty about ξ_{ij}^{*}. Then, as shown in Appendix A.2, under suitable regularity conditions, this weighted ML estimator is asymptotically normal.

3 Simulation Studies

The methods described in Section 2 are implemented and freely available as an R package fastLink. 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.

3.1 The Setup

To make our simulation studies realistic, we use a real 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 name, 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 following five parameters:

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

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

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

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

5. **Degree of measurement error**: the proportion of records whose first name, last name, and street name contains classical measurement error. We consider 0% (no measurement error), and 6%.

Together, we have conducted a total of 270 ($= 3^3 \times 5 \times 2$) simulation studies. The details of how the data sets are generated under each simulation study appear in Appendix A.3.

3.2 Results

Figure 1 compares the performance of fastLink (blue solid bars) to the two existing 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 individual records as a match if at least three fields of their address, date of birth, gender, and name are identical (Ansolabehere and Hersh, 2016). 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)

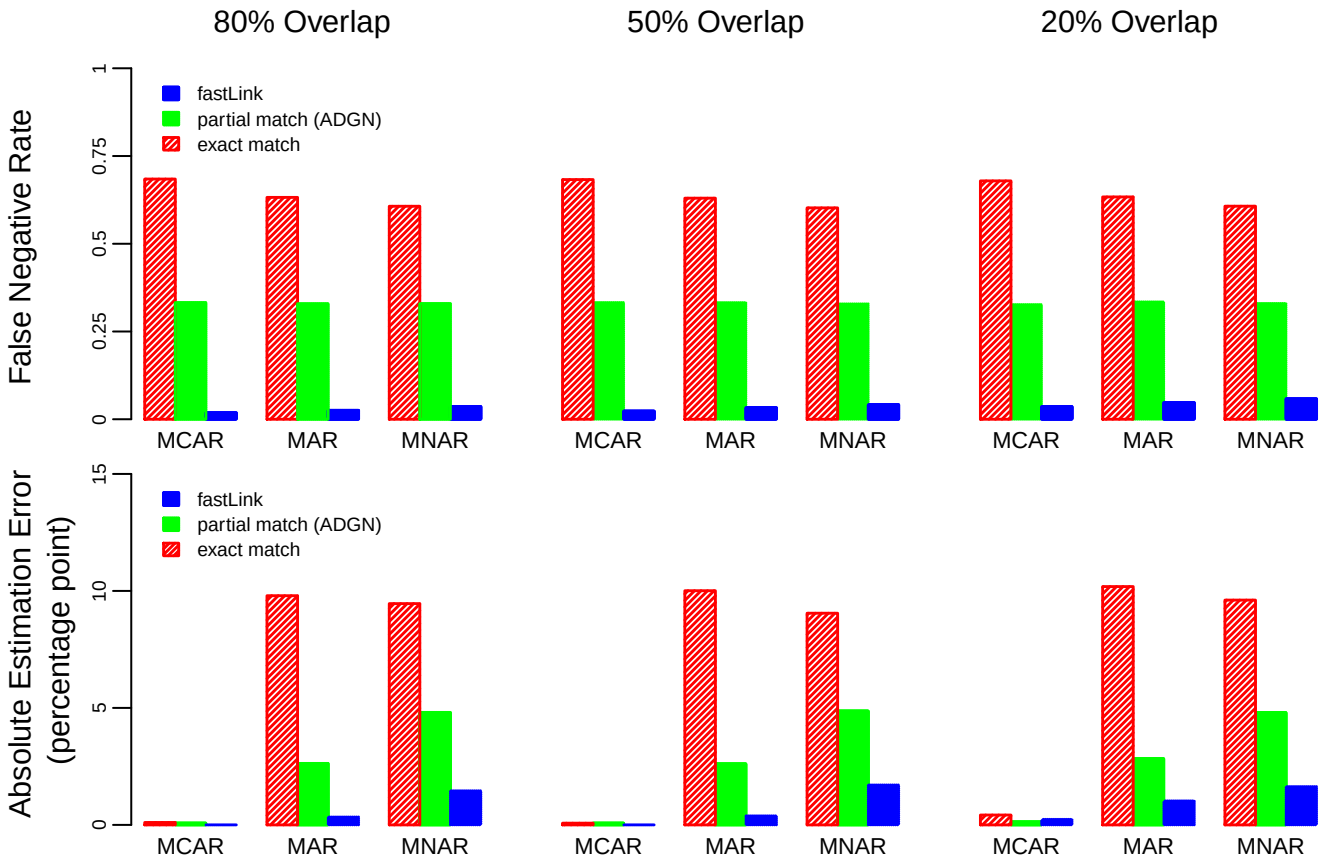

Figure 1: Accuracy of Data Merge. The top and bottom panels present the false discovery rate (FNR) and the absolute estimation error (for estimating the turnout rate), respectively, when merging datasets of 100,000 records each across with different levels of overlap (measured as a percentage of a data 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). 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.

after introducing the measurement error and the medium amount of missing data as explained above. For fastLink, only pairs with a posterior probability ≥ 0.85 are considered to be matches, but the results remain qualitatively similar if we change the threshold to 0.9 or 0.95.

We find that the proposed probabilistic methodology fastLink significantly outperforms the two deterministic methods. While all three methods are designed to control the FDR, only the proposed methodology 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 and as a result yield a FNR of much greater magnitude. In addition, we observe that the deterministic methods yield substantially greater estimation bias than fastLink unless the data are missing

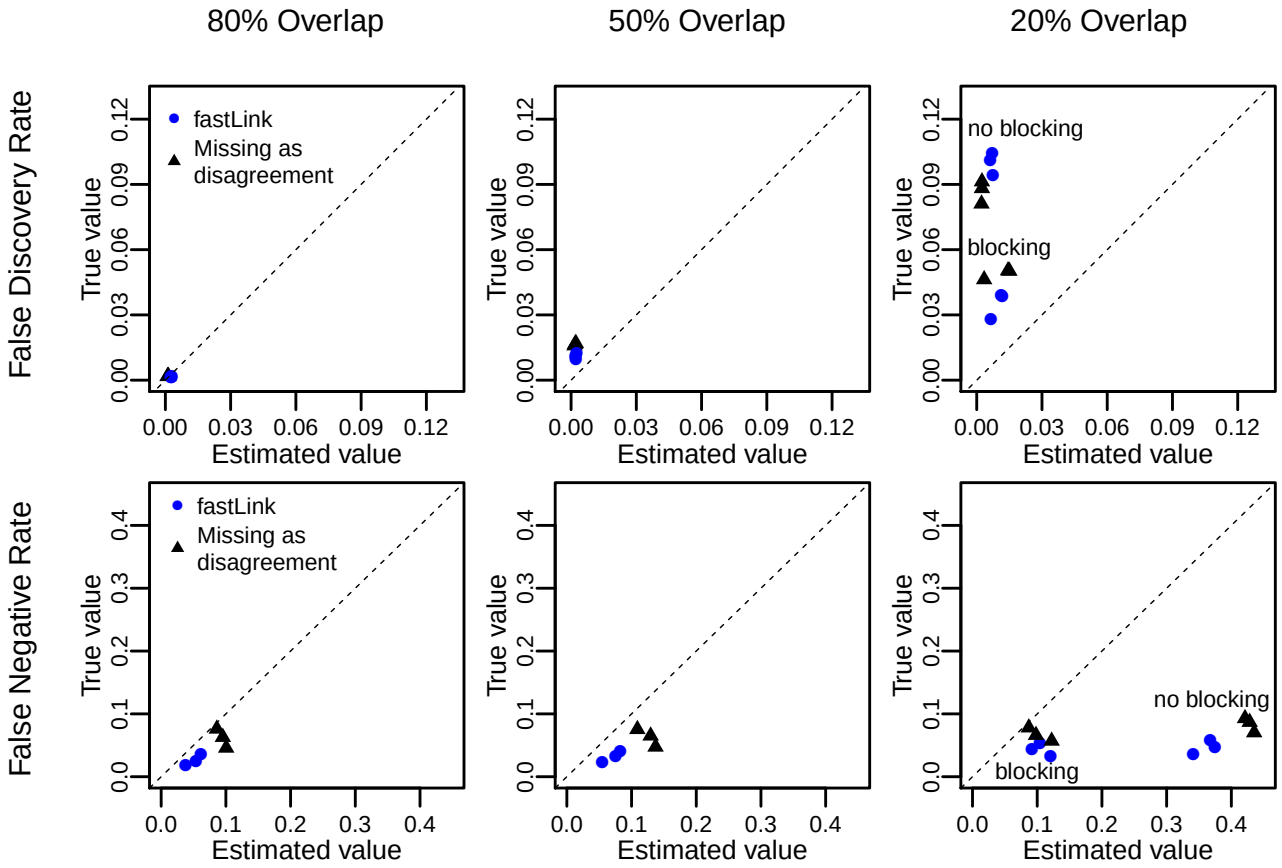

Figure 2: Accuracy of FDR and FNR Estimates. The top panel compares the estimated FDR (x-axis) with its true value (y-axis) whereas the bottom panel compares the estimated FNR against its true value. We consider the medium amount of missing data generated under MAR as a missingness mechanism and add 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. In addition, fastLink gives lower FDR and FNR than the same algorithm that treats missing values as a disagreement. Note that in cases where the overlap is small (20%), blocking improves the precision of our estimates.

completely at random. Under the other two missing data mechanisms, the magnitude of the bias is substantially greater than that of fastLink. While the proposed algorithm 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 ameliorate the problem of having small overlap would be to use blocking based on a set of covariates that are fully observed. For example, in our simulations, the year of birth is observed for each record in both datasets, and thus we can use clustering methods such as k-means to group similar observations. Then, we apply fastLink to each group 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 completely eliminated. The reason for this improvement is that blocking increases the degree of overlap within each group. For example, in this simulation setting with four clusters, the ratio of true matches to all possible pairs is approximately 6×10^{-6}, which is more than three times as large as the corresponding ratio for no blocking and is comparable to the case of overlap of 50%.

Due to space constraints, we present the results of the remaining simulation studies in a separate 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 Comparison with the Recently Proposed Probabilistic Methods

Although many social scientists use deterministic methods, the probabilistic modeling approach has been a predominant approach in the statistics literature ever since the publication of Fellegi and Sunter (1969). Recently, some researchers have proposed alternative probabilistic models (e.g.,

Steorts, 2015; Steorts, Hall and Fienberg, 2016; Sadinle, 2017). These methodologies consider data merging as the problem of forming a bipartite graph. One advantage of this approach is that the one-to-one match restriction can be imposed as part of the probabilistic models. However, their major drawback is the lack of scalability. In particular, they are implemented using Markov chain Monte Carlo methods and do not easily scale to the large-scale administrative records our methodology is designed to handle. For the sake of completeness, we compare the performance of our methodology with that of these recently proposed probabilistic methods using the small-scale validation data sets analyzed in the original papers. We find that the accuracy of our algorithm is comparable to that of these state-of-art methodologies.

For validation, Steorts (2015) uses the RLdata500 data set of only 500 records, which is part of the RecordLinkage package in R. The validation data set contains five linkage variables (first and last name, day, month, and year of birth) where noise is added to one randomly selected variable for any given record. The goal of this validation exercise is to identify 50 records that are known to be duplicates by matching one observation against another within this data set. Steorts (2015) reports that the FNR and FDR of her methodology are 0.02 and 0.04, respectively. When applying our algorithm, we use three categories for the string valued variables (first and last names), i.e., exact (or nearly identical) match, partial match, and disagreement, based on the Jaro-Winkler string distance with 0.94 and 0.88 as the cutpoints as recommended by Winkler (1990). For the numeric valued fields (day, month, and year of birth), we use a binary comparison, based on exact matches. Using fastLink, we found both FNR and FDR of our methodology to be exactly zero.

Sadinle (2017) used a synthetic data set for a validation study by varying the degree of overlap (10%, 50%, and 100%) as well as the number of variables to which noise is added (1, 2, and 3 variables). The original author used the first and last name, age, and occupation as linkage variables. Consistent with out simulation studies, Sadinle (2017) finds that as the degree of overlap between datasets decreases, the performance of the Fellegi-Sunter model worsens while this is not the case for his proposed probabilistic model. In the original simulations, Sadinle (2017) uses four agreement levels for the string valued fields (first and last name) based on the Levenshtein edit distance, while binary comparisons are applied to age and occupation. When we apply fastLink using the same setup as described above (e.g., Jaro-Winkler distance with the aforementioned recommended cutoff values for string valued variables), we find that our methodology performs as

well as the method of Sadinle (2017) except in the case when the overlap is 10% and measurement error is present in three out of four linkage variables.

In sum, using the existing small-scale validation data sets, we find that our methodology, as implemented via the fastLink package, is more or as accurate as the state-of-art probabilistic models. The main advantage of our methodology, however, is computational efficiency, to which we now turn.

3.4 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 J., 2017) in terms 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. For all implementations, we set the convergence threshold to 1×10^{-5}.[2]

We first consider the setup where 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 name, last name, house number, street name, and year of birth. We vary the size of each data set from 1,000 records to 30,000 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.

The left plot of 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

[2]We use different starting values for different methods because other methods do not allow users to specify starting values. However, we believe this is unlikely to be an issue because the EM algorithm converges quickly. In fact, the real bottleneck is the data preparation, for which we use a hashing technique as described in Section 2.3.2.

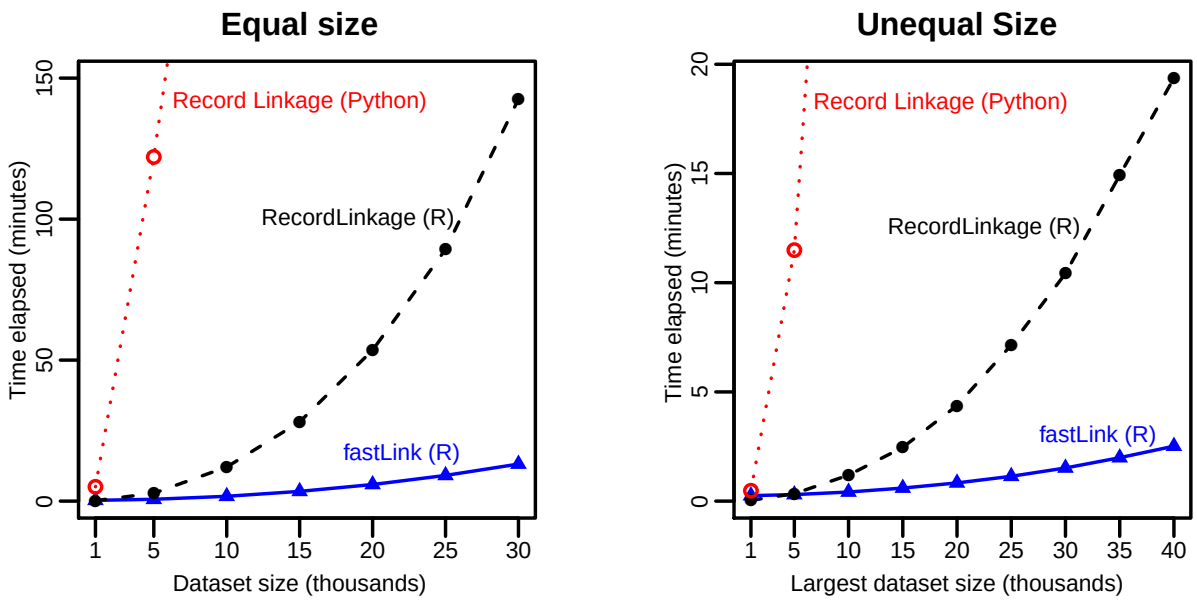

Figure 3: Running Time Comparison. The left plot presents the results of merging datasets of equal size while the right plot shows the results of merging datasets of unequal size (1:10 ratio). The datasets were constructed from a sample of female registered voters in California. The amount of overlap between datasets is 50% of the smaller data set, and, for each dataset, there are 10% missing observations in each linkage variable: first name, middle name, 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. In both cases, the proposed implementation fastLink (blue solid triangles connected by a solid line) is significantly faster than the other open source packages.

running time increases exponentially for the other packages in contrast to fastLink (blue solid triangles connected by a solid line), which exhibits a near linear increase. For the data sets of all sizes considered here, fastLink takes less than 10 minutes to merge. In contrast, it takes more than 2 hours for Record Linkage (Python; indicated by red open circles connected by a dotted line), to merge two data sets of only 5,000 observations each. The performance is not as bad for Record Linkage (R; represented by black solid circles connected by a dashed line), but it still takes over 2.5 hours to merge data sets of 30,000 records each.

We emphasize that fastLink is also far more memory-efficient than the other packages. For example, RecordLinkage (R) writes all objects to disk, requiring more than 200GB of free disk space when merging two equally sized data sets of 30,000 records. Similarly, Record Linkage (Python) tries to use as much free RAM memory as possible, requiring the system to swap at between 8GB and 12GB of disk space to act as RAM when merging data sets of 10,000 records

each. While Record Linkage (Python) is more memory efficient than RecordLinkage (R), merging data sets of 10,000 observations each took over 5 hours. In contrast, fastLink uses less than 100MB of disk space with 1GB of RAM memory needed when merging data sets of 30,000 records each.

Similar results are obtained when we merge two data sets of different sizes. In the right plot of Figure 3, we merge two data sets where the size of one data set is 10% of that of the other data set. We keep the overlap of the two data sets to be 50% of the smaller data set, and vary the size of the larger data set from 1,000 to 40,000 records. We use the same settings regarding missing data and measurement error as the one used in the comparison based on two data sets of equal size. Consistent with the above results, we find that while fastLink scales almost linearly in the dataset size, the computational time for the other two packages grows exponentially.

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 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 more than 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 name, 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 name, 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 Section 2.2 and find the best match in the DIME data for each CCES respondent. Then, we declare as a match any pair whose posterior 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. 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.

		fastLink			Proprietary method
		0.75	0.85	0.95	
Number of matches	All	4945	4794	4573	4534
	Female	2198	2156	2067	2210
	Male	2747	2638	2506	2324
Overlap between fastLink and proprietary method	All	3958	3935	3880	
	Female	1878	1867	1845	
	Male	2080	2068	2035	
Match rate (%)	All	9.66	9.37	8.93	8.85
	Female	8.11	7.96	7.63	8.16
	Male	11.40	10.94	10.40	9.64
False discovery rate (FDR; %)	All	1.24	0.65	0.21	
	Female	0.91	0.52	0.14	
	Male	1.49	0.75	0.27	
False negative rate (FNR; %)	All	15.25	17.35	20.81	
	Female	5.34	6.79	10.29	
	Male	21.84	24.37	27.81	

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 posterior 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.

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 the same matches all the time. In fact, out of 4794 matches identified by fastLink (using the threshold of 0.85), the proprietary method does not identify 859 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 FDR is controlled to less than

1.5% for both male and female respondents. The 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 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. 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 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.[3] 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 (0.58) but is greater than the correlation for those matches identified by the proprietary method alone (0.46). We also examine the distribution of posterior 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 methods have the highest posterior match probability whereas most of the matches identified only by the proprietary method have extremely low posterior match probabilities.

We further examine the quality of the matches through Table 3, which presents the five most

[3]Figures 7 and 8 in Appendix A.4 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.

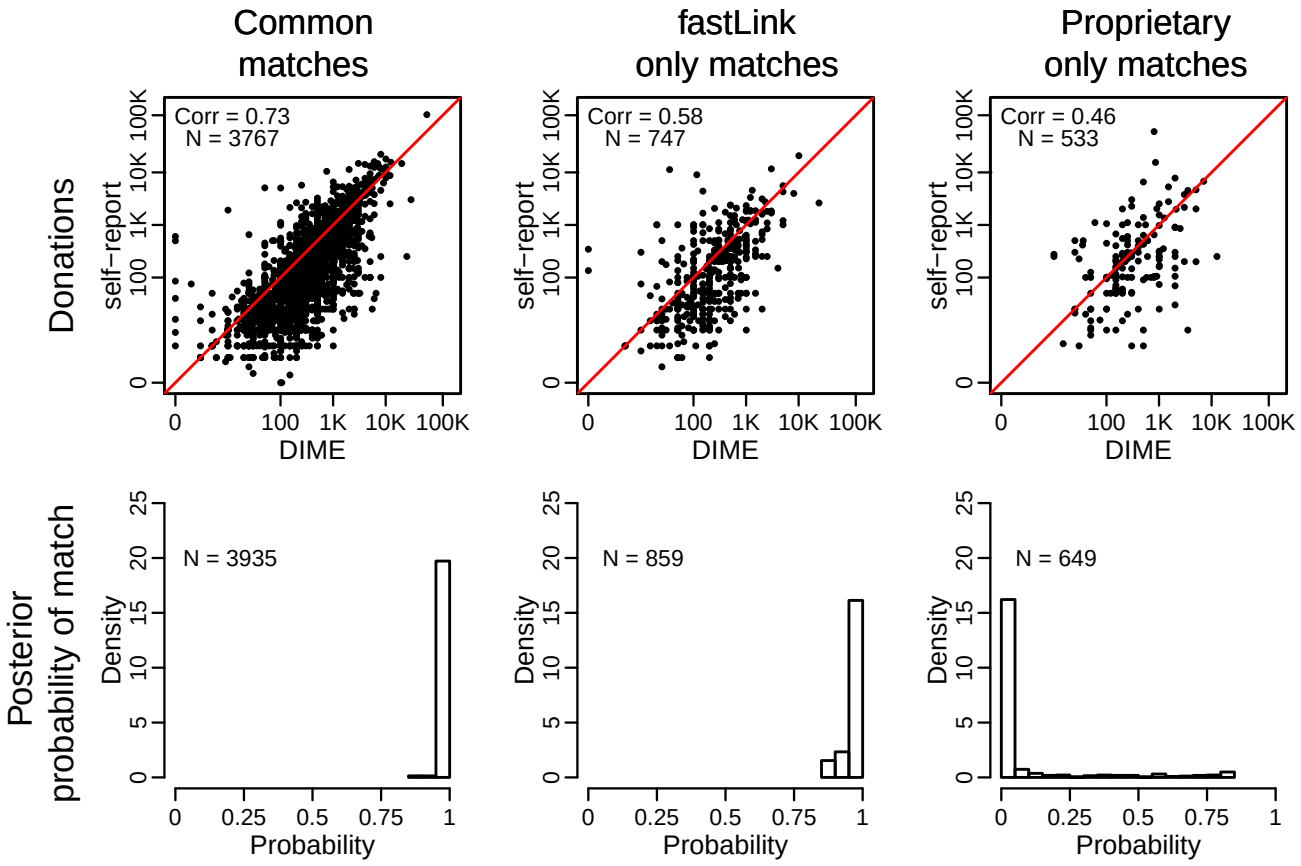

Figure 4: Comparison of fastLink and the Proprietary Method. The top panel compares the self-reported donations (y-axis) by matched CCES respondents with their donation amount recorded in the DIME data (x-axis) for the three different groups of observations: those declared as matches by both fastLink 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 posterior match probability for each group. For fastLink, we use one-to-one match with the threshold of 0.85.

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

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

Table 3: Five Most Frequent Agreement Patterns for Matches Identified by Both Methods, fastLink Only, and the Proprietary Method Only. For a given agreement pattern, the number of matches and the posterior match probability (according to fastLink) are presented in the "Counts" and "Prob." columns, respectively. For fastLink we use one-to-one match with the threshold of 0.85.

or at least as good as 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 conser-

	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 4: 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.

vative), is constructed by applying a factor analysis to a series of questions on various issues.[4] 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 2.5, we fit the same linear regression except that we use the posterior mean of the match indicator variable as the main explanatory variable rather than the match indicator variable. Table 4 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 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

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

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, which tracks the number of taxpayers who move within-state and across-state through their tax returns.

The SOI data has several advantages over other official statistics on migration rates, such as the Census Bureau's Current Population Survey (CPS) Annual Social and Economic Supplement and the U.S. Postal Service's National Change of Address (NCOA) registry. First, rather than relying on self-reported migration status as in the CPS, the IRS can use an individual's Social

Security Number to track their residence year-to-year. Furthermore, because of its comprehensive coverage, we can use SOI to precisely estimate migration rates for even small states. In contrast, survey-based measures suffer due to small sample size, leading to noisy estimates of state-to-state migration rates. For instance, in the 2015 CPS March Supplement, which is based on approximately 75,000 households, only 4,089 respondents report moving to another state within the past year. This makes the accurate estimation of moving rates across 2,450 state pairs essentially impossible. Lastly, all taxpaying citizens are accounted for in the SOI data, whereas citizens opt in to registering with the U.S. Postal Service's NCOA service, creating bias in the merge results.

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 Sections 2.3.3 and 2.4.2). Our merge is based on first name, middle name, last name, house number, street name, date/year/month of birth, date/year/month of registration, and gender. Step 1 uses each of these fields to inform the merge, while Step 2 uses only first name, middle name, 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 Section 2.4.2, 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 Step 1, 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 Step 2, we set a prior on λ.

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

(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 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)

Several remarks about this merge procedure are in order. First, in Step 1, we use a 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 more than 800,000 voters and use the entire state file for states with fewer than 800,000 voter records on file. In Figure 9 of Appendix A.5, we show through simulation studies that for datasets as small as 100,000 records, a 5% random sample leads to parameter estimates that are 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, which corresponds to a 5% of records from California, the largest state in the United States. Therefore, estimation for every other state is based on random samples of no less than 5%, suggesting that the sample size should be sufficiently large to yield precise estimates.

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 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. We chose to cluster on first name rather than last name because clustering on last name would fail to properly cluster women who changed their last name in their last year after marriage. This procedure ensures a large number of possible similar comparisons in each block while allowing us to parallelize computation across blocks. Finally, the entire merge procedure is computationally

		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 5: 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 posterior 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.

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 Section 2.3.3, 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 5 presents the overall match rate, FDR, and FNR statistics for fastLink. We assess the performance of the match at three separate posterior 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, we find that the overall match rate remains very high, at 93.04% under a 95% acceptance threshold, while the FDR and FNR remain controlled at 0.03% and 3.86%. All three thresholds yield match rates that are 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 posterior matching probability across all potential match pairs. For both within-state and across-state

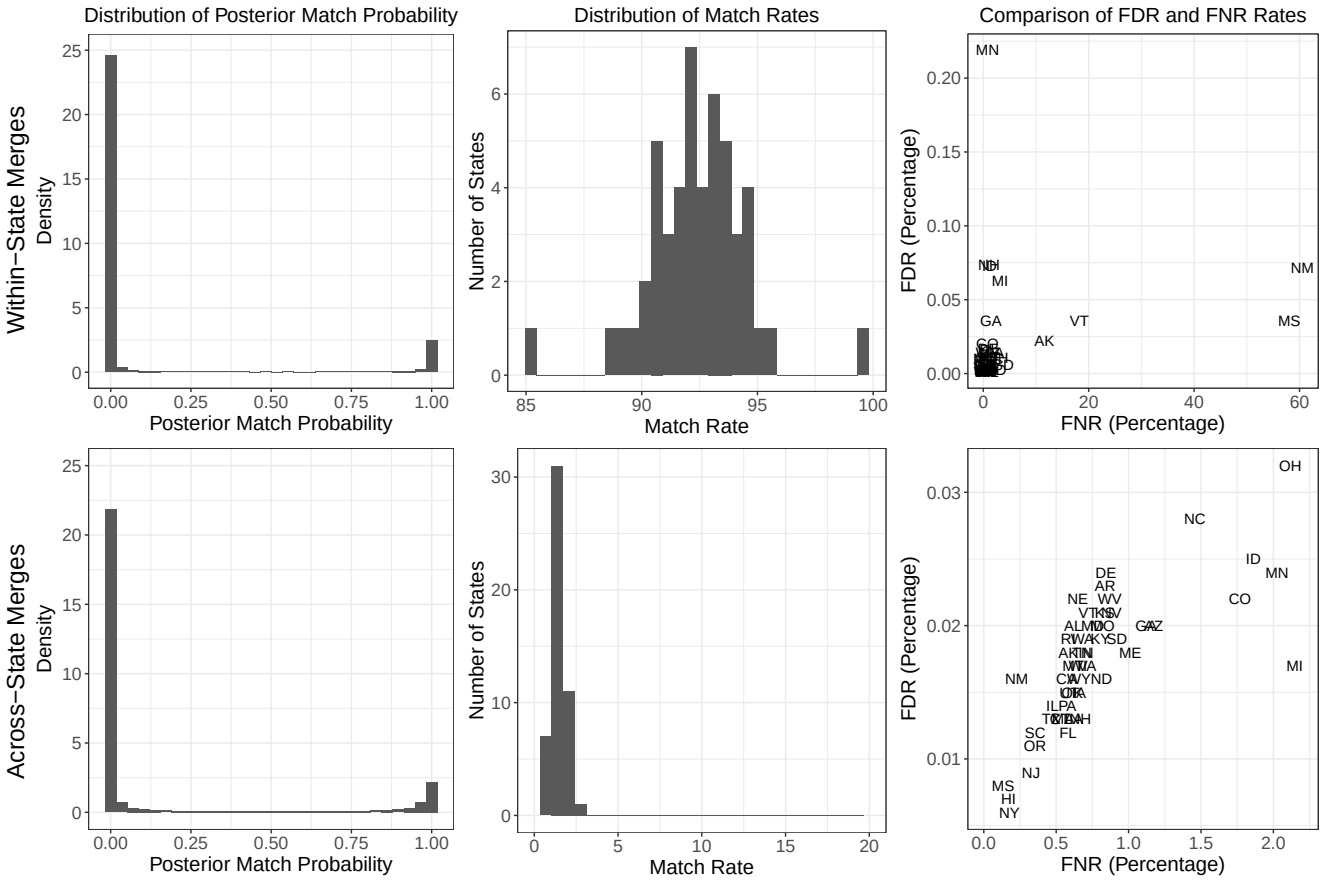

Figure 5: Graphical Diagnostics from Merging the 2014 Nationwide Voter File with the 2015 Nationwide Voter File. This figure presents graphical diagnostics for fastLink for within-state matches (top panel) and across-state matches (bottom panel). The first column plots the distribution of the posterior 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.

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 clustered tightly between 0% and 5%.

In the third column, we plot the False Discovery Rate (FDR) against the False Negative Rate (FNR) for each state. For the within-state merge, the FDR is controlled extremely well — every state other than Minnesota has an FDR below 0.1%. In addition, there are only two states,

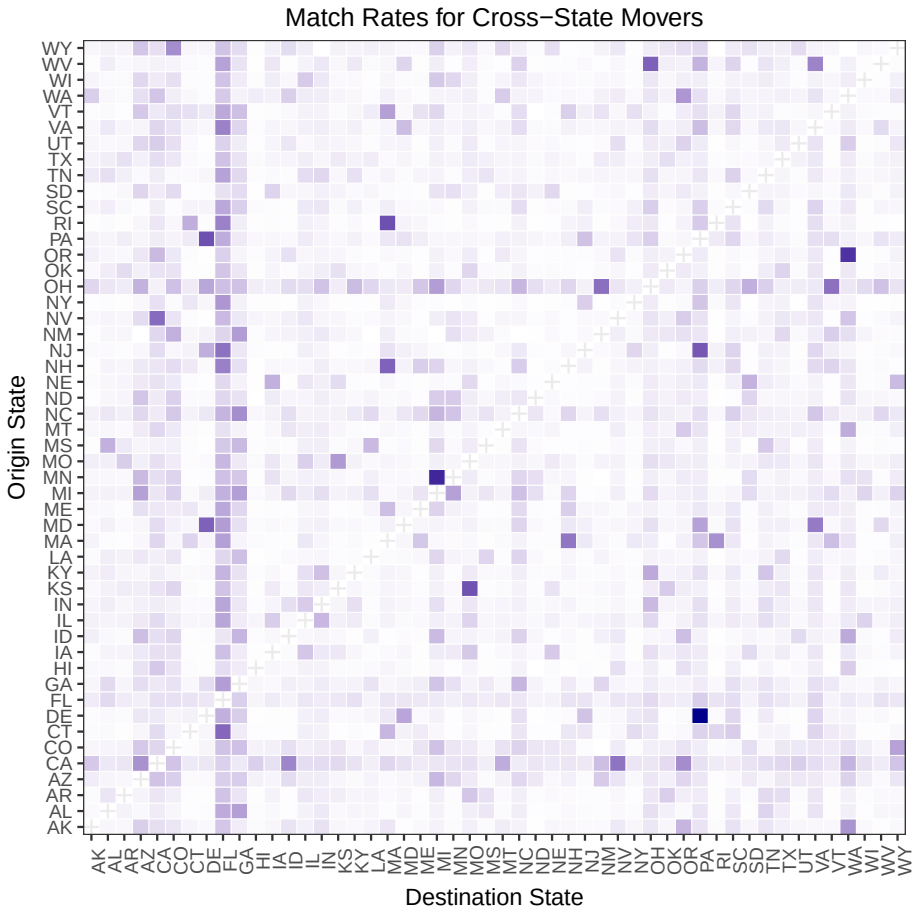

Figure 6: Across-State Match Rates for the 2014 Nationwide Voter File to 2015 Nationwide Voter File Merge. We plot the match rates from each across-state match pair as a heatmap, where darker colors indicate a higher match rate.

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 matches we are identifying are still 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 recovered by 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 through our procedure. First, we find a migration cluster in New England, where voters from New Hampshire and Rhode Island migrated to Mas-

sachusetts 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 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 plays an essential role.

In this paper, we advocate the use of probabilistic modeling to assist merging large-scale data sets. The main advantage of probabilistic models is their ability to quantify 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 faster and more scalable implementation of the canonical model, proposing ways to incorporate auxiliary information such as name frequency and migration rates, and showing how to incorporate the uncertainty about the merge process in post-merge analyses. 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. Our method is as accurate as state-of-art probabilistic models and yet is much faster and scalable.

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 proportion of true matches is expected to be small. For example, turnout validation of survey respondents remains a challenging task since one must find a couple of thousand voters in a nationwide voter file of over 160 million records. Similarly, studying voter fraud based on linking non-citizen survey respondents with voter file records is difficult, requiring caution in interpreting its results (Richman, Chattha and Earnest, 2014). 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. We leave this and other important 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. 2016. "ADGN: An Algorithm for Record Linkage Using Address, Date of Birth, Gender and Name." Working paper.

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.

Bishop, B. and R. G. Cushing. 2008. *The Big Sort: Why the Clustering of Like-minded America is Tearing Us Apart.* Boston, MA: Houghton Mifflin Harcourt.

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.

Borg, Andreas and Murat Sariyar. 2016. *RecordLinkage: Record Linkage in R.* R package version 0.4-10.
URL: *https://CRAN.R-project.org/package=RecordLinkage*

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.

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.

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

de Bruin, Jonathan. 2017. *The Python Record Linkage Toolkit*. python package version 0.8.1. **URL:** *https://pypi.python.org/pypi/recordlinkage/*

DellaVigna, Stefano and Ethan Kaplan. 2007. "The Fox News Effect: Media Bias and Voting." *Quarterly Journal of Economics* 122:1187–1234.

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

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

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.

Enos, R. D. and A. Fowler. 2016. "Aggregate Effects of Large-Scale Campaigns on Voter Turnout." *Political Science Research and Methods*.

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.

Fouirnaies, A. and A. B. Hall. Forthcoming. "How Do Interest Groups Seek Access to Committees?" *American Journal of Political Science*.

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.

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

Hill, Seth. Forthcoming. "Changing Votes or Changing Voters: How Candidates and Election Context Swing Voters and Mobilize the Base." *Electoral Studies*.

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. 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.

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.

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.

Nickerson, D. W. and T. Rogers. 2014. "Political Campaigns and Big Data." *Journal of Economic Perspectives* 28:51–74.

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.

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.*

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

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

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.

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

Winkler, William E. 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.

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.

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. 2006. Overview of record linkage and current research directions. Technical Report. United States Bureau of the Census.

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

A Supplementary Appendix

A.1 Incorporating the Prior Information about Migration

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

$$\lambda \;=\; \frac{1}{b_\lambda - a_\lambda + N_{\mathcal{A}} N_{\mathcal{B}}} \left(a_\lambda - 1 + \sum_{i=1}^{N_{\mathcal{A}}} \sum_{j=1}^{N_{\mathcal{B}}} \xi_{ij} \right) \tag{24}$$

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

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

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

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

The approximation follows because for large data sets, we typically have $a_\lambda \gg 1$ so that $(a_\lambda - 1)/a_\lambda \approx 1$ (see equation (24)). In addition, since λ is a small number and at most $\frac{\min(N_{\mathcal{A}}, N_{\mathcal{B}})}{N_{\mathcal{A}} N_{\mathcal{B}}}$, a_λ is negligible compared to b_λ. Hence, the sum of the weights effectively reduces to 1. This leads to the following two equalities,

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

Solving these equations yields,

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

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

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

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

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

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

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

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

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

The result is given by,

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

A.2 The Properties of the Weighted Maximum Likelihood Estimator

We show that the expected value of the weighted log-likelihood function given in equation (23) is equal to the expected value of the log-likelihood function of the original model defined in equation (22),

$$\mathbb{E}\left[\int \xi_{ij}^* \log P_\theta(Y_i \mid Z_i^*, \mathbf{X}_i)\ dZ_i\right]$$

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

where the second equality follows from equations (17) and (20). Under mild regularity conditions, we can show that the weighted ML estimator is asymptotically normal,

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

where

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

and θ_0 is the true value of θ.

A.3 The Details of Simulation Setups

In this section, we describe the details of simulation setups for a total of 540 simulation studies we conducted. Let $\mathcal{A}$ and $\mathcal{B}$ denote the smaller and larger dataset of two data sets to be merged, respectively. We first create the larger data set $\mathcal{B}$ by randomly selecting 100,000 records out of our pool of 341,160 voters, i.e., $N_{\mathcal{B}} = 100,000$.

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

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

- **Missing data:** we consider five missing data mechanisms and three missing data proportions, 5%, 10%, and 15%, i.e., $\omega \in \{0.05, 0.1, 0.15\}$. We introduce missing data into the

following five variables, first name, middle initial, last name, house number, and street name.

1. Missing completely at random (MCAR): For each of the aforementioned five linkage variables, we independently and randomly select $\omega N_{\mathcal{A}}$ and $\omega N_{\mathcal{B}}$ observations for the data sets $\mathcal{A}$ and $\mathcal{B}$, respectively, and recode their values as missing.
2. Missing at Random (MAR) with independence across the linkage variables: We make the missing probability of each variable dependent on year of birth. For each of the linkage variables in the first data set $\mathcal{A}$, we first compute the quantiles of the year of birth variable. Among the observations whose quantile is greater than or equal to 0.4, we shuffle these quantile values. In contrast, those records whose quantile values are less than 0.4, we leave them unchanged. This induces a moderate amount of correlation between age and the probability of missing. Finally, we set the probability of missing to be proportional to the resulting quantile. Using these probabilities, we independently and randomly select $\omega N_{\mathcal{A}}$ observations to have missing values for each variable. We repeat the same procedure for the second data set, $\mathcal{B}$.
3. Missing not at Random (MNAR) with independence across the linkage variables: The only difference between MNAR and MAR is that once we shuffle the quantiles of year of birth we multiply each quantile by 0.3 for all individuals that had voted in the 2004 Presidential election. This makes those who voted less likely to have missing values.
4. MAR and MNAR with dependence across the linkage variables: The only difference between these settings and their independence counterpart is that for each linkage field (other than year of birth), we use the missing probabilities that are proportional to the same set of quantiles (without shuffling them). This makes the same group of observations likely to have missing values in multiple covariates.

- **Measurement error**: For first name, last name, and street name, we added three types of typographical errors. They are transpositions (John $\rightarrow$ Jonh), deletions (John $\rightarrow$ Jon), and replacements (John $\rightarrow$ Jobn). For each of the aforementioned linkage variables, we randomly selected 6% of their observations, and then we added each type of noise to one third of those observations.

A.4 Additional Empirical Results with Different Thresholds

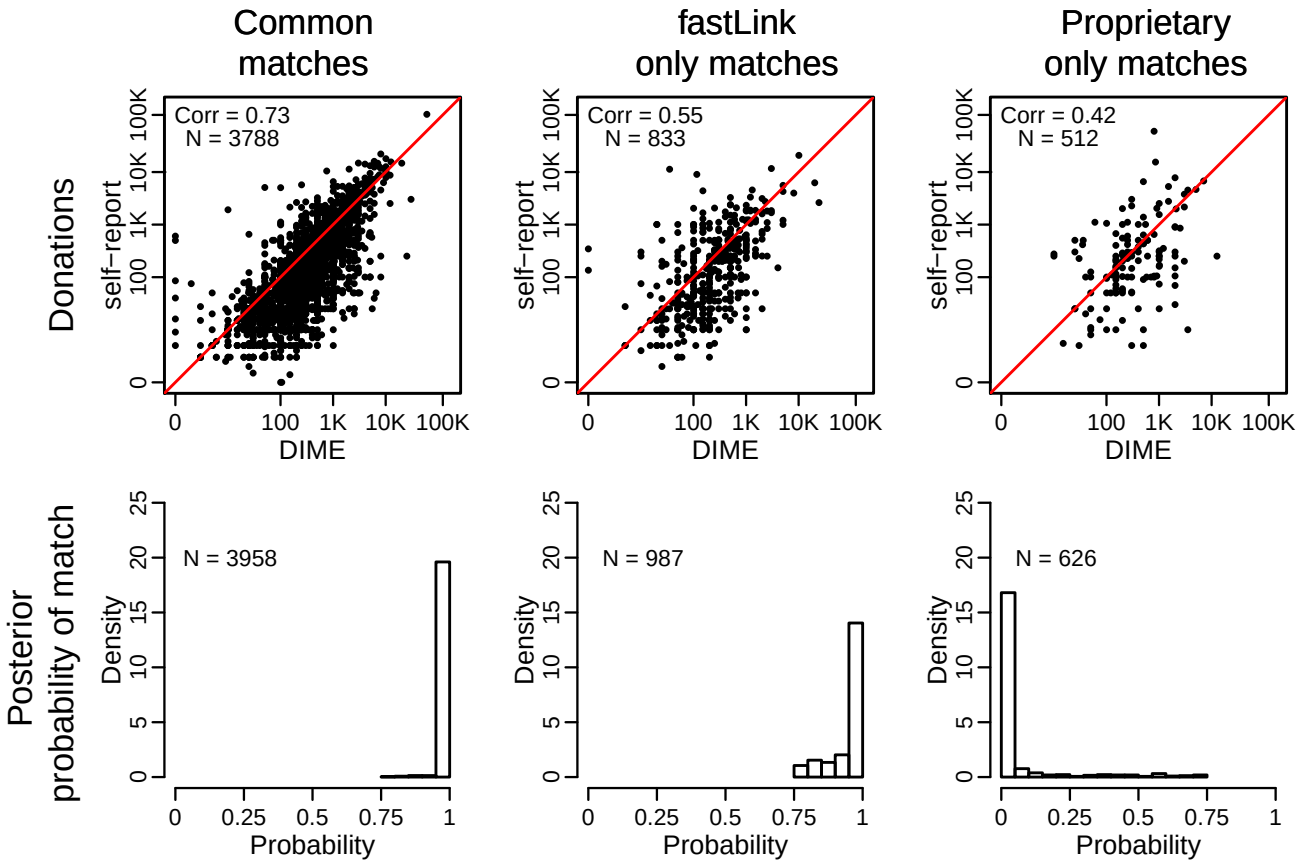

Figure 7: Comparison of fastLink and the Proprietary Method with the Threshold of 0.75. The top panel compares the self-reported donations (y-axis) by matched CCES respondents with their donation amount recorded in the DIME data (x-axis) for the three different groups of observations: those declared as matches by both fastLink and the proprietary method (left), those identified by fastLink only (middle), and those matched by the proprietary method only (right). The bottom panel presents histograms for the posterior match probability for each group. For fastLink, we use one-to-one match with the threshold of 0.75.

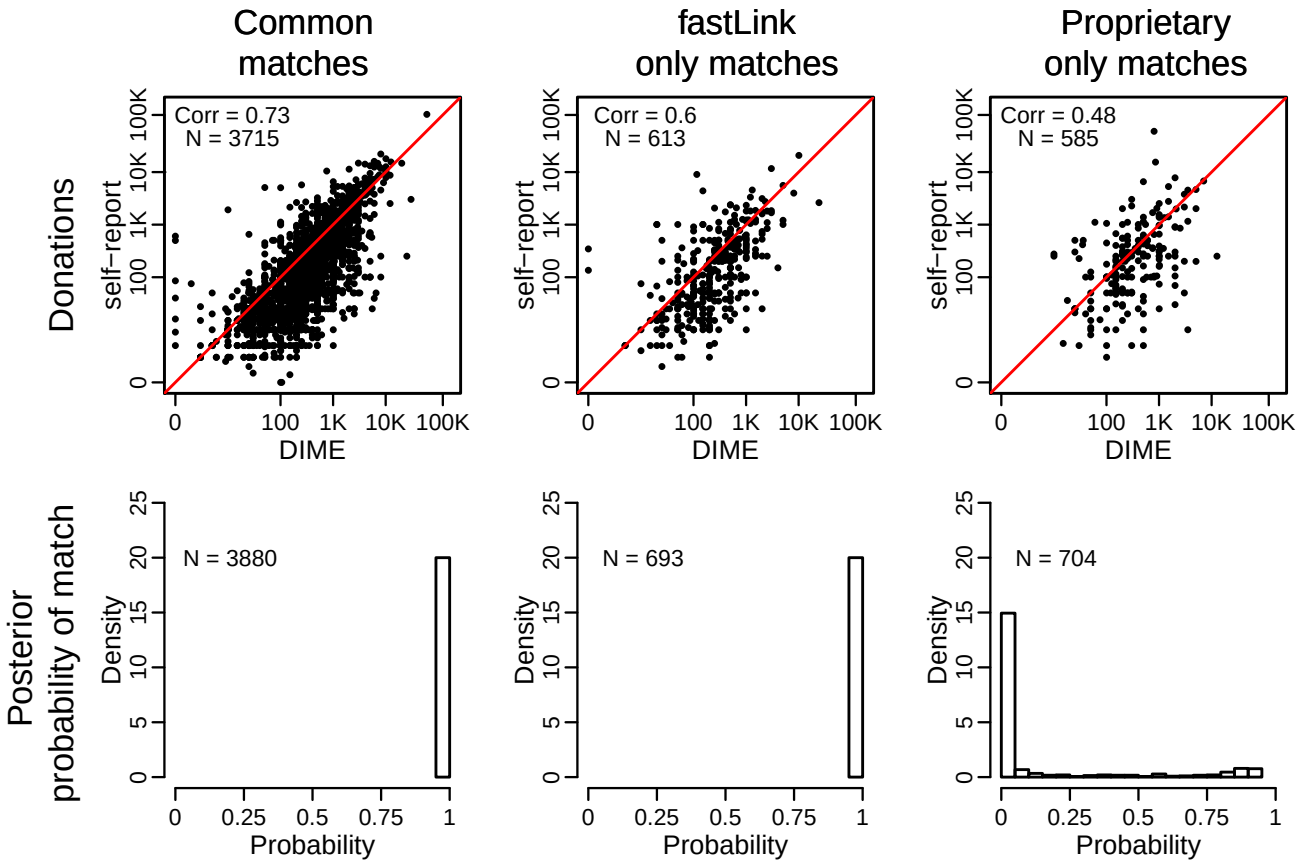

Figure 8: Comparison of fastLink and the Proprietary Method with the Threshold of 0.95. The top panel compares the self-reported donations (y-axis) by matched CCES respondents with their donation amount recorded in the DIME data (x-axis) for the three different groups of observations: those declared as matches by both fastLink and the proprietary method (left), those identified by fastLink only (middle), and those matched by the proprietary method only (right). The bottom panel presents histograms for the posterior match probability for each group. For fastLink, we use one-to-one match with the threshold of 0.95.

A.5 Simulation Results for Random Sampling

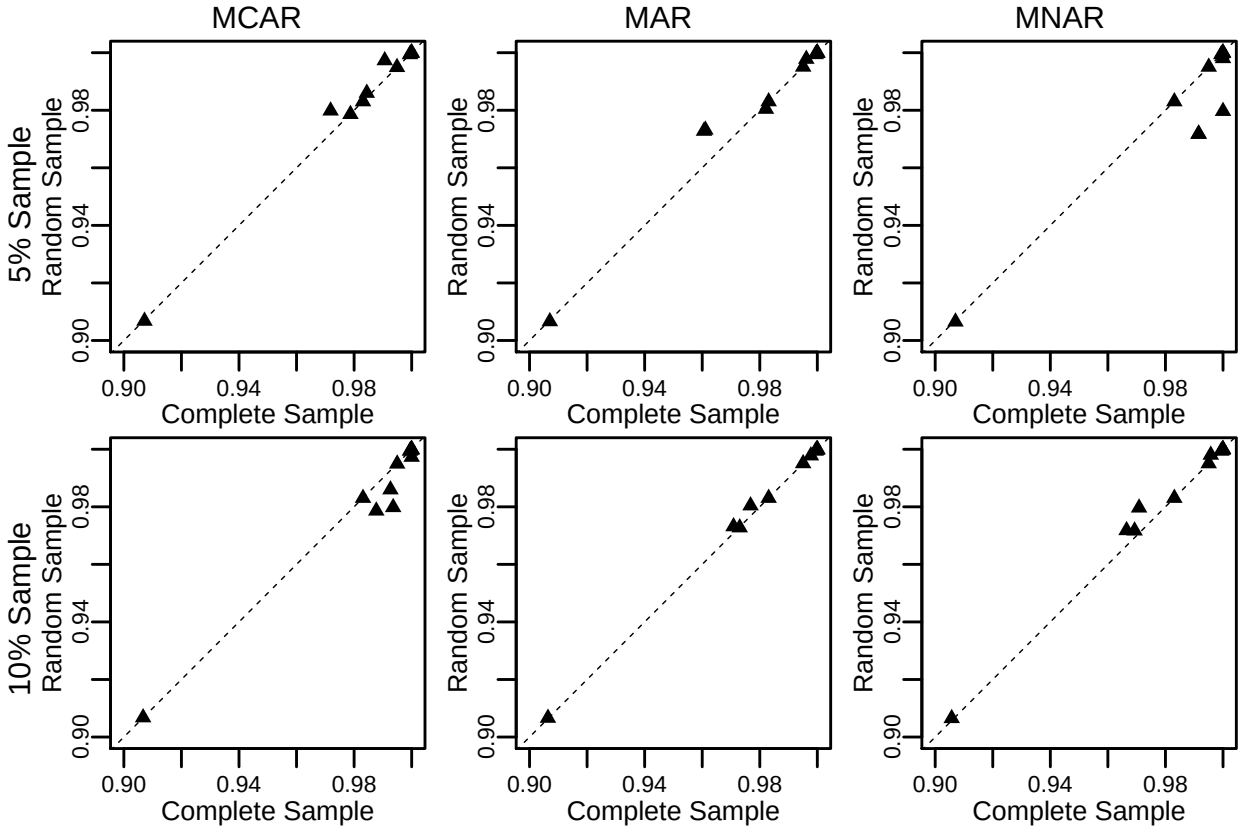

Figure 9: Parameter Estimates from Random Samples Compared Against the Parameter Estimates from the Full Dataset. The top panel compares parameter estimates from running fastLink on a full simulated dataset of size 100,000 (x-axis) against a 5% random sample from that same dataset (y-axis) under three different missing data mechanisms. The bottom panel compares parameter estimates from running fastLink on the same full simulated dataset against a 10% random sample from that same dataset. For all exercises except for the 5% random sample under MNAR, the parameter estimates from the random sample approximate the full-sample parameter estimates very closely.