Bayesian Modeling for Simultaneous Regression and Record Linkage

Jiurui Tang(✉), Jerome P. Reiter, and Rebecca C. Steorts

Department of Statistical Science, Duke University, Durham, USA
{jiurui.tang,jreiter}@duke.edu, beka@stat.duke.edu

Abstract. Often data analysts use probabilistic record linkage techniques to match records across two data sets. Such matching can be the primary goal, or it can be a necessary step to analyze relationships among the variables in the data sets. We propose a Bayesian hierarchical model that allows data analysts to perform simultaneous linear regression and probabilistic record linkage. This allows analysts to leverage relationships among the variables to improve linkage quality. Further, it enables analysts to propagate uncertainty in a principled way, while also potentially offering more accurate estimates of regression parameters compared to approaches that use a two-step process, i.e., link the records first, then estimate the linear regression on the linked data. We propose and evaluate three Markov chain Monte Carlo algorithms for implementing the Bayesian model, which we compare against a two-step process.

1 Introduction

Increasingly, data analysts seek to link records across data sets to facilitate statistical analyses. As a prototypical example, a health researcher seeks to link data from a previously completed study to patients' electronic medical records to collect long-term outcomes, with the ultimate goal of estimating relationships between the long-term outcomes and baseline covariates. Such linkages are performed readily when unique identifiers, such as social security numbers, are available for all records in all data sets.

Often, however, one or more of the data sets do not have unique identifiers, perhaps because they were never collected or are not made available due to privacy concerns. In such cases, analysts have to link records based on indirect identifiers, such as name, date of birth, and demographic variables [1,2]. Generally, such indirect identifiers contain distortions and errors. As a result, they can differ across the data sets, which can make it difficult to determine the correct record linkages. This uncertainty should be quantified and propagated to statistical inferences, although typically this is not done.

In the statistics literature, the most popular method for linking records via indirect identifiers is based on the probabilistic record linkage (RL) approach

R. C. Steorts—This research was partially supported by the National Science Foundation through grants SES1131897, SES1733835, SES1652431 and SES1534412.

J. Domingo-Ferrer and K. Muralidhar (Eds.): PSD 2020, LNCS 12276, pp. 209–223, 2020.
https://doi.org/10.1007/978-3-030-57521-2_15

of Newcombe et al. [15], which was later extended and formalized by Fellegi and Sunter [4]. Many extensions to the Fellegi-Sunter (FS) model have been proposed [e.g., 18,23,24]. A common drawback of these and other probabilistic RL methods [e.g., 7,12] is the difficulty in quantifying linkage uncertainty, and propagating that uncertainty to statistical inferences. These limitations have led to developments of RL approaches from Bayesian perspectives [e.g., 3,5,6,9–11,14,16,17,19–21,25,26].

In this article, we propose a Bayesian model for performing probabilistic RL and linear regression simultaneously. The proposed model quantifies uncertainty about the linkages and propagates this uncertainty to inferences about the regression parameters. We focus on bipartite RL—that is, the analyst seeks to merge two data sets—assuming that individuals appear at most once in each data set. As we illustrate, the model can leverage relationships among the dependent and independent variables in the regression to potentially improve the quality of the linkages. This also can increase the accuracy of resulting inferences about the regression parameters.

We use a Bayesian hierarchical model that builds on prior work by Sadinle [17], who proposed a Bayesian version of the FS model for merging two data sets. In fact, one of our primary contributions is to turn the model in [17] into a procedure for jointly performing probabilistic RL and fully Bayesian inference for regression parameters. We also propose and evaluate the effectiveness of three algorithms for fitting the Bayesian hierarchical model, focusing on both the quality of the linkages and on the accuracy of the parameter estimates.

2 Review of Bayesian Probabilistic Record Linkage

In this section, we review the Bayesian bipartite RL model of [17]. Consider two data sets $\mathbf{A}_1$ and $\mathbf{A}_2$, containing n_1 and n_2 records, respectively. Without loss of generality, assume $n_1 \geq n_2$. Our goal is to link records in $\mathbf{A}_1$ to records in $\mathbf{A}_2$. We further assume that $\mathbf{A}_1$ and $\mathbf{A}_2$ do not contain duplicate records; that is, each record in $\mathbf{A}_1$ corresponds to a single individual, as is the case for $\mathbf{A}_2$. We assume that some of the same individuals are in $\mathbf{A}_1$ and $\mathbf{A}_2$.

To characterize this, we define the random variable $\mathbf{Z} = (Z_1, \ldots, Z_{n_2})$ as the vector of matching labels for the records in $\mathbf{A}_2$. For $j = 1, \ldots, n_2$, let

$$Z_j = \begin{cases} i, \text{ if record } i \in \mathbf{A}_1 \text{ and } j \in \mathbf{A}_2 \text{ refer to the same entity;} \\ n_1 + j, \text{ if record } j \in \mathbf{A}_2 \text{ does not have a match in } \mathbf{A}_1. \end{cases}$$

Analysts determine whether a pair of records (i, j) is a link, i.e., whether or not $Z_j = i$, by comparing values of variables that are common to $\mathbf{A}_1$ and $\mathbf{A}_2$. Suppose we have F common variables, also known as *linking variables* or *fields*. For $f = 1, \ldots, F$, let γ_{ij}^f represent a score that reflects the similarity of field f for records i and j. For example, when field f is a binary variable, we can set $\gamma_{ij}^f = 1$ when record i agrees with record j on field f, and $\gamma_{ij}^f = 0$ otherwise. When field f is a string variable like name, we can calculate a similarity metric

like the Jaro-Winkler distance [8,22] or the Levenshtein edit distance [13]. We can convert these string metrics to γ_{ij}^f by categorizing the scores into a multinomial variable, where the categories represent the strength of agreement. We illustrate this approach in Sect. 4.

For each record (i, j) in $\mathbf{A}_1 \times \mathbf{A}_2$, let $\boldsymbol{\gamma}_{ij} = (\gamma_{ij}^1, \ldots, \gamma_{ij}^F)$. We assume $\boldsymbol{\gamma}_{ij}$ is a realization of a random vector Γ_{ij} distributed as

$$\Gamma_{ij}|Z_j = i \overset{iid}{\sim} \mathcal{M}(\mathbf{m}), \qquad \Gamma_{ij}|Z_j \neq i \overset{iid}{\sim} \mathcal{U}(\mathbf{u}), \qquad \text{where}$$

$\mathcal{M}(\mathbf{m})$ represents the model for comparison vectors among matches, and $\mathcal{U}(\mathbf{u})$ represents the model for comparison vectors among non-matches. For each field f, we let $m_{f\ell} = \mathbb{P}(\Gamma_{ij}^f = \ell | Z_j = i)$ be the probability of a match having level ℓ of agreement in field f, and let $u_{f\ell} = \mathbb{P}(\Gamma_{ij}^f = \ell | Z_j \neq i)$ be the probability of a non-match having level ℓ of agreement in field f. Let $\mathbf{m}_f = (m_{f1}, \ldots, m_{fL_f})$ and $\mathbf{u}_f = (u_{f1}, \ldots, u_{fL_f})$; let $\mathbf{m} = (\mathbf{m}_1, \ldots, \mathbf{m}_F)$ and $\mathbf{u} = (\mathbf{u}_1, \ldots, \mathbf{u}_F)$.

For computational convenience, it is typical to assume the comparison fields are conditionally independent given the matching status of the record pairs. Let $\boldsymbol{\Theta} = (\mathbf{m}, \mathbf{u})$. The likelihood of the comparison data can be written as

$$\mathcal{L}(\mathbf{Z}|\boldsymbol{\Theta}, \boldsymbol{\gamma}) = \prod_{i=1}^{n_1} \prod_{j=1}^{n_2} \prod_{f=1}^{F} \prod_{\ell=0}^{L_f} \left[m_{f\ell}^{I(Z_j=i)} u_{f\ell}^{I(Z_j \neq i)} \right]^{I(\gamma_{ij}^f = \ell)}, \tag{1}$$

where $I(\cdot) = 1$ when its argument is true and $I(\cdot) = 0$ otherwise.

Sometimes, there are missing values in the linking variables. Although we do not consider this possibility in our simulations, we summarize how to handle missing values in the model, assuming ignorable missing data. With conditional independence and ignorability, we can marginalize over the missing comparison variables. The likelihood of the observed comparison data can be written as

$$\mathcal{L}(\mathbf{Z}|\boldsymbol{\Theta}, \boldsymbol{\gamma}^{\text{obs}}) = \prod_{f=1}^{F} \prod_{\ell=0}^{L_f} m_{f\ell}^{a_{f\ell}(\mathbf{Z})} u_{f\ell}^{b_{f\ell}(\mathbf{Z})}, \qquad \text{where} \tag{2}$$

$a_{f\ell}(\mathbf{Z}) = \sum_{i,j} I_{\text{obs}}(\gamma_{ij}^f) I(\gamma_{ij}^f = \ell) I(Z_j = i)$ and $b_{f\ell}(\mathbf{Z}) = \sum_{i,j} I_{\text{obs}}(\gamma_{ij}^f) I(\gamma_{ij}^f = \ell) I(Z_j \neq i)$. For a given $\mathbf{Z}$, these represent the number of matches and non-matches with observed disagreement level ℓ in field f. Here, $I_{\text{obs}}(\cdot) = 1$ when its argument is observed, and $I_{\text{obs}}(\cdot) = 0$ when its argument is missing.

To define the prior distributions for $\mathbf{m}$ and $\mathbf{u}$, for all fields f, let $\boldsymbol{\alpha}_f = (\alpha_{f0}, \ldots, \alpha_{fL_f})$ and $\boldsymbol{\beta}_f = (\beta_{f0}, \ldots, \beta_{fL_f})$. We assume that $\mathbf{m}_f \sim \text{Dirichlet}(\boldsymbol{\alpha}_f)$ and $\mathbf{u}_f \sim \text{Dirichlet}(\boldsymbol{\beta}_f)$, where $\boldsymbol{\alpha}_f$ and $\boldsymbol{\beta}_f$ are known parameters. In our simulation studies, we set all entries of $\boldsymbol{\alpha}_f$ and $\boldsymbol{\beta}_f$ equal to 1 for every field.

We use the prior distribution for $\mathbf{Z}$ from [17]. For $j = 1, \ldots, n_2$, let the indicator variable $I(Z_j \leq n_1) \mid \pi \overset{iid}{\sim} \text{Bernoulli}(\pi)$, where π is the proportion of matches expected a priori. Let $\pi \sim \text{Beta}(\alpha_\pi, \beta_\pi)$, where the prior mean $\alpha_\pi/(\alpha_\pi + \beta_\pi)$ represents the expected percentage of overlap. Let $n_{12}(\mathbf{Z}) = \sum_j I(Z_j \leq n_1)$

be the number of matches according to $\mathbf{Z}$. The prior specification implies that $n_{12}(\mathbf{Z}) \sim$ Beta-Binomial$(n_2, \alpha_\pi, \beta_\pi)$ after marginalizing over π. That is,

$$\mathbb{P}(n_{12}(\mathbf{Z})) = \binom{n_2}{n_{12}(\mathbf{Z})} \frac{\mathrm{B}(n_{12}(\mathbf{Z}) + \alpha_\pi, n_2 - n_{12}(\mathbf{Z}) + \beta_\pi)}{\mathrm{B}(\alpha_\pi, \beta_\pi)}. \tag{3}$$

We assume that, conditional on the value of $n_{12}(\mathbf{Z})$, all the possible bipartite matchings are equally likely a priori. There are $n_1!/(n_1 - n_{12}(\mathbf{Z}))!$ such bipartite matchings. Thus, the prior distribution for $\mathbf{Z}$ is

$$\mathbb{P}(\mathbf{Z}|\alpha_\pi, \beta_\pi) = \mathbb{P}(n_{12}(\mathbf{Z})|\alpha_\pi, \beta_\pi)\mathbb{P}(\mathbf{Z}|n_{12}(\mathbf{Z}), \alpha_\pi, \beta_\pi) \tag{4}$$

$$= \frac{(n_1 - n_{12}(\mathbf{Z}))!}{n_1!} \frac{\mathrm{B}(n_{12}(\mathbf{Z}) + \alpha_\pi, n_2 - n_{12}(\mathbf{Z}) + \beta_\pi)}{\mathrm{B}(\alpha_\pi, \beta_\pi)}. \tag{5}$$

3 The Bayesian Hierarchical Model for Simultaneous Regression and Record Linkage

In this section, we present the Bayesian hierarchical model for regression and RL, and propose three Markov chain Monte Carlo (MCMC) algorithms for fitting the model in practice. Throughout, we assume the explanatory variables $\mathbf{X}$ are in $\mathbf{A}_1$, and the response variable $\mathbf{Y}$ is in $\mathbf{A}_2$.

3.1 Model Specification

We assume the standard linear regression, $\mathbf{Y}|\mathbf{X}, \mathbf{V}, \mathbf{Z} \sim N(\mathbf{X}\beta, \sigma^2\mathbf{I})$. Here, $\mathbf{V}$ are linking variables used in the RL model but not in the regression model. Analysts can specify prior distributions on (β, σ^2) that represent their beliefs. A full specification of the joint distribution of $(\mathbf{Y}, \mathbf{X}|\mathbf{V})$ requires analysts to specify some marginal model for $\mathbf{X}$, written generically as $f(\mathbf{X}|\mathbf{V})$. In some contexts, however, it is not necessary to specify $f(\mathbf{X}|\mathbf{V})$, as we explain in Sect. 3.2. Critically, this model assumes that the distribution of $(\mathbf{Y}, \mathbf{X}|\mathbf{V})$ is the same for matches and non-matches. Finally, for the RL component of the model, we model $\mathbf{Z}$ using the Bayesian FS approach in Sect. 2.

For the simulation studies, we illustrate computations with the Bayesian hierarchical model using univariate Y and univariate X. Assume $X \sim N(\mu, \tau^2)$. As a result, in the simulations, the random variable (X, Y) follows a bivariate normal distribution with

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim N\Big(\begin{bmatrix} \mu \\ \beta_0 + \beta_1\mu \end{bmatrix}, \begin{bmatrix} \tau^2 & \beta_1\tau^2 \\ \beta_1\tau^2 & \sigma^2 + \beta_1^2\tau^2 \end{bmatrix} \Big).$$

We assume a normal-Gamma prior on the regression parameters. Letting $\phi = 1/\sigma^2$, we have $\phi \sim G(.5, .5)$ and $\beta|\phi \sim N(b_0, \Phi_0\phi^{-1})$ where $b_0 = [3, 1]^T$ and Φ_0 is a 2×2 identity matrix. When needed, we assume Jeffrey's prior $p(\mu, \tau^2) \propto 1/\tau^2$.

3.2 Estimation Strategies

Even in the relatively uncomplicated set-up of the simulation study, it is not possible to compute the posterior distribution of the model parameters in closed form. Therefore, we consider three general strategies for MCMC sampling in order to approximate the posterior distribution.

First, we propose an MCMC sampler that uses only the linked records when estimating the full conditional distribution of the regression parameters in each iteration of the sampler. This method generates potentially different samples of linked records at each iteration. A key advantage of this method is that it does not require imputation of $\mathbf{X}$; hence, analysts need not specify $f(\mathbf{X}|\mathbf{V})$. We call this the joint model without imputation, abbreviated as JM. Second, we propose an MCMC sampler that imputes the missing values of $\mathbf{X}$ for those records in $\mathbf{A}_2$ without a link, and updates the regression parameters in each iteration using the linked pairs as well the imputed data. We call this the joint model with imputation, abbreviated as JMI. Third, we propose an MCMC sampler that is similar to JMI but uses an extra step when imputing the missing values of $\mathbf{X}$. Specifically, at each iteration, we (i) sample values of the regression parameters from a conditional distribution based on only the linked records, (ii) use the sampled parameters to impute missing values in $\mathbf{X}$, and (iii) update regression coefficients based on linked as well as imputed pairs. By adding step (i), we aim to reduce potential effects of a feedback loop in which less accurate estimates of regression parameters result in less accurate estimates of the conditional distribution of $\mathbf{X}$, and so on through the MCMC iterations. We call this the joint model with imputation and reduced feedback (JMIF).

For JMI and JMIF, inferences are based on every entity in $\mathbf{A}_2$, whereas for JM inferences are based on the subsets of linked pairs, which can differ across MCMC iterations. Analysts should keep these differences in mind when selecting an algorithm that suits their goals.

3.3 Details of the MCMC Samplers

In this section, we present the mathematical details for implementing the three proposed MCMC samplers. Before doing so, we present an algorithm for a two-step approach, where we perform RL and then use the linked data for regression. The three proposed MCMC samplers for the Bayesian hierarchical model utilize parts of the algorithm for the two-step approach.

3.3.1 Two Step Approach (TS)

Given the parameter values at iteration t of the sampler, we need to sample new values $\mathbf{m}_f^{[t+1]} = (m_{f0}^{[t+1]}, \ldots, m_{fL_f}^{[t+1]})$ and $\mathbf{u}_f^{[t+1]} = (u_{f0}^{[t+1]}, \ldots, u_{fL_f}^{[t+1]})$, where $f = 1, \ldots, F$. We then sample a new value $\mathbf{Z}^{[t+1]} = (Z_1^{[t+1]}, \ldots, Z_{n_2}^{[t+1]})$. The steps are as follows.

T.1 For $f = 1, \ldots, F$, sample

$$\mathbf{m}_f^{[t+1]} | \boldsymbol{\gamma}^{obs}, \mathbf{Z}^{[t]} \sim \text{Dirichlet}(a_{f0}(\mathbf{Z}^{[t]}) + \alpha_{f0}, \ldots, a_{fL_f}(\mathbf{Z}^{[t]}) + \alpha_{fL_f}), \quad (6)$$

$$\mathbf{u}_f^{[t+1]}|\boldsymbol{\gamma}^{obs}, \mathbf{Z}^{[t]} \sim \text{Dirichlet}(b_{f0}(\mathbf{Z}^{[t]}) + \beta_{f0}, \ldots, b_{fL_f}(\mathbf{Z}^{[t]}) + \beta_{fL_f}). \quad (7)$$

Collect these new draws into $\boldsymbol{\Theta}^{[t+1]}$. Here, each

$$a_{fl}(\mathbf{Z}) = \sum_{i,j} I_{obs}(\boldsymbol{\gamma}_{ij}^f) I(\gamma_{ij}^f = l) I(Z_j = i), \quad (8)$$

$$b_{fl}(\mathbf{Z}) = \sum_{i,j} I_{obs}(\boldsymbol{\gamma}_{ij}^f) I(\gamma_{ij}^f = l) I(Z_j \neq i). \quad (9)$$

T.2 Sample the entries of $\mathbf{Z}^{[t+1]}$ sequentially. Having sampled the first $j-1$ entries of $\mathbf{Z}^{[t+1]}$, we define $\mathbf{Z}_{-j}^{[t+(j-1)/n_2]} = (Z_1^{[t+1]}, \ldots, Z_{j-1}^{[t+1]}, Z_{j+1}^{[t]}, \ldots, Z_{n_2}^{[t]})$. We sample a new label $Z_j^{[t+1]}$, with the probability of selecting label $q \in \{1, \ldots, n_1, n_1 + j\}$ given by $p_{qj}(\mathbf{Z}_{-j}^{[t+(j-1)/n_2]} \mid \boldsymbol{\Theta}^{[t+1]})$. This can be expressed for generic $\mathbf{Z}_{-j}$ and $\boldsymbol{\Theta}$ as

$$p_{qj}(\mathbf{Z}_{-j}|\boldsymbol{\Theta}) \propto \begin{cases} \exp[w_{qj}] I(Z_{j'} \neq q, \forall j' \neq j), \text{ if } q \leq n_1 \\ [n_1 - n_{12}(\mathbf{Z}_{-j})] \dfrac{n_2 - n_{12}(\mathbf{Z}_{-j}) - 1 + \beta_\pi}{n_{12}(\mathbf{Z}_{-j}) + \alpha_\pi}, \text{ if } q = n_1 + j; \end{cases} \quad (10)$$

where $w_{qj} = \log[\mathbb{P}(\boldsymbol{\gamma}_{qj}^{obs}|Z_j = q, \mathbf{m})/\mathbb{P}(\boldsymbol{\gamma}_{qj}^{obs}|Z_j \neq q, \mathbf{u})]$ is equivalently

$$w_{qj} = \sum_{f=1}^{F} I_{obs}(\boldsymbol{\gamma}_{qj}^f) \sum_{l=0}^{L_f} \log\Big(\frac{m_{fl}}{u_{fl}}\Big) I(\gamma_{qj}^f = l). \quad (11)$$

The normalizing constant for $p_{qj}(\mathbf{Z}_{-j}|\boldsymbol{\Theta})$ is

$$\prod_{i=1}^{n_1} \prod_{k=1}^{n_2} \prod_{f=1}^{F} \prod_{\ell=0}^{L_f} \Big[m_{f\ell}^{I(Z_k=i)} u_{f\ell}^{I(Z_k\neq i)}\Big]^{I(\gamma_{ij}^f=\ell)} u_{fl}^{I(Z_j\neq i)} \times \frac{\big(n_1 - (n_{12}(\mathbf{Z}_{-j}) + 1)\big)!}{n_1!} \frac{(n_{12}(\mathbf{Z}_{-j}) + \alpha_\pi)!(n_2 - (n_{12}(\mathbf{Z}_{-j}) + 1) + \beta_\pi - 1)!}{(n_2 + \alpha_\pi + \beta_\pi - 1)} \quad (12)$$

where $k \neq j$.

T.3 We now add the regression parameters to the sampler. For any draw of $\mathbf{Z}^{[t+1]}$, we sample $\beta^{[t+1]}$ and $(\sigma^2)^{[t+1]} = \phi^{-1}$ from

$$\beta^{[t+1]}|\phi, \mathbf{Y}, \mathbf{Z}^{[t+1]} \sim N(b_n, (\phi\Phi_n)^{-1}) \quad (13)$$

$$\phi|\mathbf{Y}, \mathbf{Z}^{[t]} \sim G\Big(\tfrac{n+1}{2}, \tfrac{1}{2}(SSE + 1 + \hat{\beta}^T\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}}\hat{\beta} + b_0^T\phi_0 b_0 - b_n^T\Phi_n b_n)\Big) \quad (14)$$

where $SSE = \widetilde{\mathbf{Y}}^T[\mathbf{I} - \widetilde{\mathbf{X}}(\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}^T]\widetilde{\mathbf{Y}}$,$\Phi_n = \widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}} + \Phi_0$, $b_n = \Phi_n^{-1}(\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}}\hat{\beta} + \phi_0 b_0)$, and $\hat{\beta} = (\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}^T\widetilde{\mathbf{Y}}$. Here, $\widetilde{\mathbf{X}}$ and $\widetilde{\mathbf{Y}}$ are the subsets of $\mathbf{X}$ and $\mathbf{Y}$ belonging to only the linked cases at iteration t.

Steps **T.1** and **T.2** are the same as those used in [17]; we add **T.3** to sample the regression parameters. Alternatively, and equivalently, analysts can run **T.1** and **T.2** until MCMC convergence, then apply **T.3** to each of the resulting draws of $\mathbf{Z}^{[t]}$ to obtain the draws of the regression parameters.

3.3.2 Joint Method Without Imputation (JM)

The sampler for the JM method uses **T.1**, but it departs from **T.2** and **T.3**. As we shall see, in JM we need the marginal density $f(\mathbf{Y})$. This can be approximated with a standard univariate density estimator. Alternatively, one can derive it from $f(\mathbf{Y}|\mathbf{X})$ and extra assumptions about $f(\mathbf{X})$, although these extra assumptions obviate one of the advantages of JM compared to JMI and JMIF. In the simulations, for convenience we use the fact that (Y, X) are bivariate normal when computing the marginal density of $\mathbf{Y}$, as evident in step **J.2**. Step **J.2** can be omitted when using means to compute $f(\mathbf{Y})$ that do not leverage a joint model for $(\mathbf{Y}, \mathbf{X})$.

J.1 Sample $\mathbf{m}_f^{[t+1]}$ and $\mathbf{u}_f^{[t+1]}$ using **T.1**.

J.2 Sample $\mu^{[t+1]}$ and $(\tau^2)^{[t+1]}$ using $1/\tau^2 \sim G((n-1)/2, \sum(X_i - \bar{X})^2)$ and $\mu|\tau^2 \sim N(\bar{X}, \tau^2)$. We use all of $\mathbf{X}$ in this step.

J.3 Given $\mathbf{Z}^{[t]}$, sample $\beta^{[t+1]}$ and $(\sigma^2)^{[t+1]}$ from (13) and (14).

J.4 Sample $\mathbf{Z}^{[t+1]}$ sequentially. Having sampled the first $j-1$ entries of $\mathbf{Z}^{[t+1]}$, we define $\mathbf{Z}_{-j}^{[t+(j-1)/n_2]} = (Z_1^{[t+1]}, \ldots, Z_{j-1}^{[t+1]}, Z_{j+1}^{[t]}, \ldots, Z_{n_2}^{[t]})$. Then we sample a new label $Z_j^{[t+1]}$ with probability $p_{qj}(\mathbf{Z}_{-j}^{[t+(j-1)/n_2]} \mid \boldsymbol{\Theta}^{[t+1]}, \mathbf{X}, \mathbf{Y})$ of selecting label $q \in \{1, \ldots, n_1, n_1 + j\}$. For generic $(\mathbf{Z}_{-j}, \boldsymbol{\Theta}, \mathbf{X}, \mathbf{Y})$, we have $f(\mathbf{Z}_{-j}|\boldsymbol{\Theta}, \mathbf{X}, \mathbf{Y}) \propto f(\mathbf{Y}, \mathbf{X}|\boldsymbol{\Theta}, \mathbf{Z}_{-j}) f(\mathbf{Z}_{-j}|\boldsymbol{\Theta})$. For $q \leq n_1$, and using the definition for w_{qj} in (11), we thus have

$$p_{qj}(\mathbf{Z}_{-j}|\boldsymbol{\Theta}, \mathbf{X}, \mathbf{Y}) \propto \exp[w_{qj}] I(Z_{j'} \neq q, \forall j' \neq j) \prod_{i \neq q, i \in \mathbf{A}_{12}} f(X_i, Y_i|\mathbf{Z}_{-j})$$
$$\times \prod_{i \neq q, i \in \mathbf{A}_{1-}} f(X_i) \prod_{i \neq q, i \in \mathbf{A}_{2-}} f(Y_i) f(Y_j, X_q) \tag{15}$$
$$\propto \exp[w_{qj}] I(Z_{j'} \neq q, \forall j' \neq j) \frac{f(Y_j, X_q)}{f(Y_j) f(X_q)} \tag{16}$$
$$= \exp[w_{qj}] I(Z_{j'} \neq q, \forall j' \neq j) \frac{f(Y_j|X_q)}{f(Y_j)}. \tag{17}$$

Here, $\mathbf{A}_{12}$ is the set of matched records, $\mathbf{A}_{1-}$ is the set of records in $\mathbf{A}_1$ without a match in $\mathbf{A}_2$, and $\mathbf{A}_{2-}$ is the set of records in $\mathbf{A}_2$ without a match in $\mathbf{A}_1$.

For $q = n_1 + j$, after some algebra to collect constants, we have

$$p_{qj}(\mathbf{Z}_{-j}|\boldsymbol{\Theta}, \mathbf{X}, \mathbf{Y}) \propto [n_1 - n_{12}(\mathbf{Z}_{-j})] \frac{n_2 - n_{12}(\mathbf{Z}_{-j}) - 1 + \beta_\pi}{n_{12}(\mathbf{Z}_{-j}) + \alpha_\pi}.$$

3.3.3 Joint Method with Imputation (JMI)

The sampler for JMI is similar to the sampler for JM, except we impute $\mathbf{X}$ for non-matches in $\mathbf{A}_2$. Thus, we require a model for $\mathbf{X}$, which we also use to compute $f(\mathbf{Y})$. In accordance with the simulation set-up, we present the sampler with $X \sim N(\mu, \tau^2)$.

I.1 Sample $\mathbf{m}_f^{[t+1]}$ and $\mathbf{u}_f^{[t+1]}$ using **J.1**.
I.2 Sample $\mu^{[t+1]}$ and $(\tau^2)^{[t+1]}$ using **J.2**.
I.3 Impute $\mathbf{X}_{mis}$ for those records in $\mathbf{A}_2$ without a matched X. For the sampler in the simulation study, the predictive distribution is

$$X_{mis} \sim N\Big(\mu + \frac{\beta_1 \tau^2}{\sigma^2 + \beta_1^2 \tau^2}(Y - \beta_0 - \beta_1 \mu), \tau^2 - \frac{\beta_1^2 \tau^4}{\sigma^2 + \beta_1^2 \tau^2}\Big). \quad (18)$$

In JMI, we use the values of $(\beta^{[t]}, (\sigma^2)^{[t]}, (\tau^2)^{[t+1]})$ in (18). Once we have $\mathbf{X}_{mis}^{[t+1]}$, in the full conditional distributions for $(\beta^{[t+1]}, (\sigma^2)^{[t+1]})$ we use both the matched and imputed data for all records in $\mathbf{A}_2$, with the priors in Sect. 3.3.1. As a result, we draw $\beta^{[t+1]}$ and $(\sigma^2)^{[t+1]}$ based on (13) and (14), but let $(\widetilde{\mathbf{X}}, \widetilde{\mathbf{Y}})$ include both the linked pairs and imputed pairs in $\mathbf{A}_2$.
I.4 Sample $\mathbf{Z}^{[t+1]}$ sequentially using **J.4**.

3.3.4 Joint Method with Imputation and Reduced Feedback (JMIF)

The sampler for JMIF is like the sampler for JMI, but we use a different predictive model for $\mathbf{X}_{mis}$. We again present the sampler with $X \sim N(\mu, \tau^2)$.

F.1 Sample $\mathbf{m}_f^{[t+1]}$ and $\mathbf{u}_f^{[t+1]}$ using **J.1**.
F.2 Sample $\mu^{[t+1]}$ and $(\tau^2)^{[t+1]}$ using **J.2**.
F.3 Given $\mathbf{Z}^{[t]}$, take a draw (β^*, σ^{2*}) from the full conditional distributions in (13) and (14), using only the linked cases at iteration t. We impute $\mathbf{X}_{mis}$ for those records in $\mathbf{A}_2$ without a matched X using (β^*, σ^{2*}). For the sampler in the simulation study, we use (18) with (β^*, σ^{2*}) and $(\tau^2)^{[t+1]}$. Once we have $\mathbf{X}_{mis}^{[t+1]}$, in the full conditional distributions for $(\beta^{[t+1]}, \sigma^{[t+1]})$ we use both the matched and imputed data for all records in $\mathbf{A}_2$, with the priors in Sect. 3.3.1. As a result, we draw $\beta^{[t+1]}$ and $(\sigma^2)^{[t+1]}$ based on (13) and (14), but let $(\widetilde{\mathbf{X}}, \widetilde{\mathbf{Y}})$ include both the linked pairs and imputed pairs in $\mathbf{A}_2$.
F.4 Sample $\mathbf{Z}^{[t+1]}$ sequentially using (**J.4**).

3.4 MCMC Starting Values

Sadinle [17] starts the MCMC sampler by assuming none of the records in file $\mathbf{A}_2$ have a link in file $\mathbf{A}_1$. We do not recommend this starting point for the hierarchical model, as it is beneficial to specify an initial set of links to determine sensible starting values for the linear regression parameters. Instead, we employ a standard FS algorithm—implemented using the `RecordLinkage` package in `R`—to determine a set of links to use as a starting point.

4 Simulation Studies

We generate simulated data sets using the `RLdata10000` data set from the `RecordLinkage` package in `R`. The `RLdata10000` contains 10,000 records; 10% of these records are duplicates belonging to 1,000 individuals. The `RLdata10000` includes linking variables, which the developers of the data set have distorted to create uncertainty in the RL task. To create $\mathbf{A}_2$, we first randomly sample $n_{12} = 750$ individuals from the 1,000 individuals with duplicates. We then sample 250 individuals from the remaining 8,000 individuals in `RLdata10000`. This ensures that each record in $\mathbf{A}_2$ belongs to one and only one individual. To create $\mathbf{A}_1$, we first take the duplicates for the 750 records in $\mathbf{A}_2$; these are true matches. Next, we sample another 250 records from the individuals in `RLdata10000` but not in $\mathbf{A}_2$. Thus, we have 750 records that are true links, and 250 records in each file that do not have a match in the other file. We repeat this process independently in each simulation run.

In both files, in each simulation run, we generate the response and explanatory variables, as none are available in the `RLdata10000`. For each sampled record i, we generate $x_i \sim N(0, 1)$ and $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$ in each simulation run. We set $\beta_0 = 3$ and $\sigma^2 = 1$. We consider $\beta_1 \in \{.4, .65, .9\}$, to study how the correlation between X and Y affects performance of the methods.

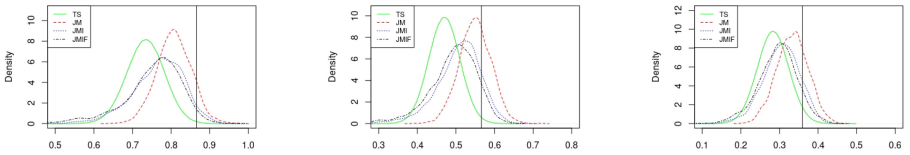
We use four linking variables: the first name and last name, and two constructed binary variables based on birth year and birth day. For the constructed variables, we create an indicator of whether the individual in the record was born before or after 1974, and another indicator of whether the individual in the record was born before or after the 16th day of the month.

To compare first and last name, we use the Levenshtein edit distance (LD), defined as the minimum number of insertions, deletions, or substitutions required to change one string into the other. We divide this distance by the length of the longest string to standardize it. The final measure is in the range of $[0, 1]$, where 0 represents total agreement and 1 total disagreement. Following [17], we categorize the LD into four levels of agreement. We set $f = 1$ and $\gamma_{ij}^f = 3$ when the first names for record i and j match perfectly ($LD = 0$); $\gamma_{ij}^f = 2$ when these names show mild disagreement ($0 < LD \leq .25$); $\gamma_{ij}^f = 1$ when these names show moderate disagreement ($.25 < LD \leq .5$); and, $\gamma_{ij}^f = 0$ when these names show extreme disagreement ($LD \geq .5$). The same is true for last names with $f = 2$. For the constructed binary variables based on birth day and year, we set $\gamma_{ij}^f = 1$ when the values for record i and j agree with each other, and $\gamma_{ij}^f = 0$ otherwise.

We create two scenarios to represent different strengths of the information available for linking. The strong linkage scenario uses all four linking variables, and the weak linkage scenario uses first and last name only.

4.1 Results

Table 1 displays averages of the estimated regression coefficients over 100 simulation runs. Across all scenarios, on average the point estimates of β_1 from the



(a) $\beta_1 = 0.9$. (b) $\beta_1 = 0.65$. (c) $\beta_1 = 0.4$.

Fig. 1. Posterior density of β_1 in one arbitrarily chosen data set for each value of β_1 under the strong linking information scenario. Posterior distribution for the two-step approach is in solid green, for JM is in dashed red, for JMI is in dotted blue, and for JMIF is in dotdash black. Vertical lines are estimates of β_1 when using all the correct links.

Bayesian hierarchical model, regardless of the MCMC algorithm, are at least as close to the true β_1 as the point estimates from the two-step approach. The hierarchical model offers the largest improvements in accuracy over the two-step approach when the correlation between X and Y is strongest and the information in the linking variables is weakest. In this scenario, the model takes advantage of information in the relationship between the study variables that the two-step approach cannot. In contrast, when the correlation between X and Y is weakest and the information in the linking variables is strongest, there is little difference in the performances of the hierarchical model and two-step approach. These patterns are illustrated in Fig. 1.

Generally, all the algorithms tend to underestimate β_1 in these simulations. It is practically impossible to identify all the true links. Therefore, the regression is estimated with some invalid pairs of (x_i, y_i). This attenuates the estimates of β_1. The hierarchical model tends to overestimate β_0 slightly. The difference is most noticeable when the correlation between X and Y is strong. Generally, on average the two-step approach offers more accurate estimates of β_0, although the differences are practically irrelevant.

Among the hierarchical models, JM outperforms JMI and JMIF, with JMIF slightly better than JMI. This is because inaccuracies in the estimated distribution of $\mathbf{X}_{mis}$ in JMI and JMIF are propagated to the estimated distributions of (β_0, β_1). To illustrate this, suppose in a particular MCMC iteration the value of β_1 is somewhat attenuated, which in turn leads to inaccuracy in the parameters of the imputation model for $X|Y$. As a result, the imputed values of $\mathbf{X}_{mis}$ are not samples from an accurate representation of $f(X|Y)$. Thus, the full conditional distribution of (β_0, β_1) is estimated from completed-data that do not follow the relationship between X and Y. The inaccurate samples of (β_0, β_1) then create inaccurate imputations, and the cycle continues. In contrast, in any iteration, JM samples coefficients using only the records deemed to be links (in that iteration), thereby reducing the effects of feedback from imprecise imputations. This also explains why JMIF yields slightly more accurate estimates of β_1 than JMI when the correlation between X and Y is strong.

Table 1. Summary of simulation results for regression coefficients. Results based on 100 runs per scenario. The true $\beta_0 = 3$ in all scenarios. For each reported average, the Monte Carlo standard errors are smaller than .01. "Strong" refers to scenarios where we use all four comparison fields, and "Weak" refers to scenarios where we use only two comparison field's.

	Results for β_1				Results for β_0			
	TS	JM	JMIF	JMI	TS	JM	JMIF	JMI
Strong								
$\beta_1 = .90$	.73	.85	.80	.79	3.01	3.05	3.04	3.04
$\beta_1 = .65$	.52	.60	.56	.55	3.00	3.02	3.01	3.01
$\beta_1 = .40$	.32	.36	.33	.32	2.99	3.00	3.00	3.00
Weak								
$\beta_1 = .90$	.60	.82	.78	.76	3.00	3.05	3.04	3.03
$\beta_1 = .65$	.42	.57	.53	.53	3.00	3.03	3.02	3.02
$\beta_1 = .40$	.27	.34	.32	.32	3.00	3.01	3.01	3.01

Table 2 displays averages across the 100 simulation runs of six standard metrics for the quality of the record linkages. These include the average numbers of correct links (CL), correct non-links (CNL), false negatives (FN), and false positives (FP), as well as the false negative rate (FNR) and false discovery rate (FDR). These are formalized in Appendix A. The results in Table 2 indicate that the hierarchical model offers improved linkage quality over the two-step approach, regardless of the estimation algorithm. In particular, the hierarchical model tends to have smaller FP and larger CNL than the two-step approach. The difference in CNL is most apparent when the information in the linking variables is weak and when the correlation between X and Y is strong. The hierarchical model tends to have higher CL than the two-step approach, but the difference is practically important only when the linkage information is weak and the correlation is relatively strong ($\beta_1 = .9, \beta_1 = .65$). Overall, the hierarchical model has lower FDR compared to the two step approach.

5 Discussion

The simulation results suggest that the Bayesian hierarchical model for simultaneous regression and RL can offer more accurate coefficient estimates than the two-step approach in which one first performs RL then runs regression on linked data. The hierarchical model is most effective when the correlation between the response and explanatory variable is strong. The hierarchical model also can improve linkage quality, in particular by identifying more non-links. This is especially the case when the information in the linking variables is not strong. In all scenarios, the relationship between the response and explanatory variable complements the information from the comparison vectors, which helps us declare record pairs more accurately.

Table 2. Summary of linkage quality across 100 simulation runs. Averages in first four columns have standard errors less than 3. Averages in the last two columns have Monte Carlo standard errors less than .002.

		CL	CNL	FN	FP	FNR	FDeR
Strong							
$\beta_1 = .90$	JM	702	152	47	128	.06	.15
	JMIF	702	155	48	125	.06	.15
	JMI	702	155	48	125	.06	.15
	TS	697	123	53	167	.07	.19
$\beta_1 = .65$	JM	702	139	48	144	.06	.17
	JMIF	701	143	49	140	.06	.16
	JMI	701	143	49	141	.06	.17
	TS	698	119	52	172	.07	.20
$\beta_1 = .40$	JM	698	131	52	158	.07	.18
	JMIF	698	133	52	155	.07	.18
	JMI	698	133	52	155	.07	.18
	TS	697	121	53	170	.07	.20
Weak							
$\beta_1 = .90$	JM	636	114	114	223	.15	.26
	JMIF	636	116	114	222	.15	.26
	JMI	636	115	114	223	.15	.26
	TS	620	53	130	315	.17	.34
$\beta_1 = .65$	JM	632	93	118	256	.16	.29
	JMIF	631	94	119	255	.16	.29
	JMI	631	92	119	256	.16	.30
	TS	621	51	129	317	.17	.34
$\beta_1 = .40$	JM	625	69	125	291	.17	.32
	JMIF	624	69	126	291	.17	.32
	JMI	625	69	125	291	.17	.32
	TS	620	50	130	319	.17	.34

As with any simulation study, we investigate only a limited set of scenarios. Our simulations have 75% of the individuals in the target file as true matches. Future studies could test whether the hierarchical model continues to offer gains with lower overlap rates, as well as different values of other simulation parameters. We used a correctly specified linear regression with only one predictor. Appendix B presents a simulation where the linear regression is mis-specified; the model continues to perform well. We note that the hierarchical model without imputation for missing $\mathbf{X}$ extends readily to multivariate $\mathbf{X}$. When outcomes are

binary, analysts can use probit regression in the hierarchical model. The model also can be modified for scenarios where one links to a smaller file containing explanatory variables. In this case, we use the marginal distribution of $\mathbf{X}$ and conditional distribution of $\mathbf{X}|\mathbf{Y}$ rather than those for $\mathbf{Y}$ and $\mathbf{Y}|\mathbf{X}$ in (17).

A Record Linkage Evaluation Metrics

Here, we review the definitions of the average numbers of correct links (CL), correct non-links (CNL), false negatives (FN), and false positives (FP). These allow one to calculate the false negative rate (FNR) and false discovery rate (FDR) [19]. For any MCMC iteration t, we define $\text{CL}^{[t]}$ as the number of record pairs with $Z_j \leq n_1$ and that are true links. We define $\text{CNL}^{[t]}$ as the number of record pairs with $Z_j > n_1$ that also are not true links. We define $\text{FN}^{[t]}$ as the number of record pairs that are true links but have $Z_j > n_1$. We define $\text{FP}^{[t]}$ as the number of record pairs that are not true links but have $Z_j \leq n_1$. In the simulations, the true number of true links is $\text{CL}^{[t]}+\text{FN}^{[t]} = 750$, and the estimated number of links is $\text{CL}^{[t]}+\text{FP}^{[t]}$. Thus, $\text{FNR}^{[t]} = \text{is FN}^{[t]}/(\text{CL}^{[t]}+\text{FN}^{[t]})$. The $\text{FDR}^{[t]} = \text{FP}^{[t]}/(\text{CL}^{[t]}+\text{FP}^{[t]})$, where by convention we take $\text{FDR}^{[t]} = 0$ when both the numerator and denominator are 0. We report the FDR instead of the FPR, as an algorithm that does not link any records has a small FPR, but this does not mean that it is a good algorithm. Finally, for each metric, we compute the posterior means across all MCMC iterations, which we average across all simulations.

B Additional Simulations with a Mis-specified Regression

As an additional simulation, we examine the performance of the hierarchical model in terms of linkage quality when we use a mis-specified regression. The true data generating model is $\log(\mathbf{Y})|\mathbf{X}, \mathbf{V}, \mathbf{Z} \sim N(\mathbf{X}\beta, \sigma^2\mathbf{I})$, but we incorrectly assume $\mathbf{Y}|\mathbf{X}, \mathbf{V}, \mathbf{Z} \sim N(\mathbf{X}\beta, \sigma^2\mathbf{I})$ in the hierarchical model. Table 3 summarizes the measures of linkage quality when the linkage variables have weak information. Even though the regression component of the hierarchical model is mis-specified, the hierarchical model still identifies more correct non-matches than the two-step approach identifies, although the difference is less obvious than when we use the correctly specified regression. We see a similar trend when the information in the linking variables is strong, albeit with smaller differences between the two-step approach and the hierarchical model.

Table 3. Results for simulation with mis-specified regression component in the hierarchical model. Entries summarize the linkage quality across 100 simulation runs. Averages in first four columns have standard errors less than 3. Averages in the last two columns have Monte Carlo standard errors less than .002.

		CL	CNL	FN	FP	FNR	FDeR
$\beta_1 = .90$	JM	3625	69	125	292	.17	.32
	JMIF	624	70	126	291	.17	.32
	JMI	624	69	126	292	.17	.32
	TS	619	51	131	318	.17	.34
$\beta_1 = .65$	JM	626	62	124	299	.17	.32
	JMIF	626	62	124	299	.17	.32
	JMI	626	62	124	299	.17	.32
	TS	622	49	128	319	.17	.34
$\beta_1 = .40$	JM	623	56	127	309	.17	.33
	JMIF	623	56	127	309	.17	.33
	JMI	623	57	127	309	.17	.33
	TS	622	50	128	318	.17	.34

References

1. Christen, P.: A survey of indexing techniques for scalable record linkage and deduplication. IEEE Trans. Knowl. Data Eng. **24**(9), 1537–1555 (2012)
2. Christen, P.: Data linkage: the big picture. Harv. Data Sci. Rev. **1**(2) (2019)
3. Dalzell, N.M., Reiter, J.P.: Regression modeling and file matching using possibly erroneous matching variables. J. Comput. Graph. Stat. **27**, 728–738 (2018)
4. Fellegi, I.P., Sunter, A.B.: A theory for record linkage. J. Am. Stat. Assoc. **64**(328), 1183–1210 (1969)
5. Fortini, M., Liseo, B., Nuccitelli, A., Scanu, M.: On Bayesian record linkage. Res. Off. Stat. **4**, 185–198 (2001)
6. Gutman, R., Afendulis, C.C., Zaslavsky, A.M.: A Bayesian procedure for file linking to analyze end-of-life medical costs. J. Am. Stat. Assoc. **108**(501), 34–47 (2013)
7. Hof, M.H., Ravelli, A.C., To, A.H.Z.: A probabilistic record linkage model for survival data. J. Am. Stat. Assoc. **112**(520), 1504–1515 (2017)
8. Jaro, M.A.: Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. J. Am. Stat. Assoc. **84**, 414–420 (1989)
9. Larsen, M.D.: Iterative automated record linkage using mixture models. J. Am. Stat. Assoc. **96**, 32–41 (2001)
10. Larsen, M.D.: Comments on hierarchical Bayesian record linkage. In: Proceedings of the Section on Survey Research Methods. ASA, Alexandria, VA, 1995–2000 (2002)
11. Larsen, M.D.: Advances in record linkage theory: hierarchical Bayesian record linkage theory. In: Proceedings of the Section on Survey Research Methods. ASA, Alexandria, VA, pp. 3277–3284 (2005)
12. Larsen, M.D., Rubin, D.B.: Iterative automated record linkage using mixture models. J. Am. Stat. Assoc. **96**(453), 32–41 (2001)

13. Levenshtein, V.I.: Binary codes capable of correcting deletions, insertions, and reversals. Soviet Phys. Doklady **10**, 707–710 (1965)
14. Marchant, N.G., Steorts, R.C., Kaplan, A., Rubinstein, B.I., Elazar, D.N.: d-blink: distributed end-to-end Bayesian entity resolution (2019). arXiv preprint arXiv:1909.06039
15. Newcombe, H.B., Kennedy, J.M., Axford, S.J., James, A.P.: Automatic linkage of vital records: computers can be used to extract "follow-up" statistics of families from files of routine records. Science **130**(3381), 954–959 (1959)
16. Sadinle, M.: Detecting duplicates in a homicide registry using a Bayesian partitioning approach. Ann. Appl. Stat. **8**(4), 2404–2434 (2014). MR3292503
17. Sadinle, M.: Bayesian estimation of bipartite matchings for record linkage. J. Am. Stat. Assoc. **112**, 600–612 (2017)
18. Sadinle, M., Fienberg, S.E.: A generalized Fellegi-Sunter framework for multiple record linkage with application to homicide record systems. J. Am. Stat. Assoc. **108**(502), 385–397 (2013)
19. Steorts, R.C.: Entity resolution with empirically motivated priors. Bayesian Anal. **10**(4), 849–875 (2015). MR3432242
20. Steorts, R.C., Hall, R., Fienberg, S.E.: A Bayesian approach to graphical record linkage and deduplication. J. Am. Stat. Assoc. **111**(516), 1660–1672 (2016)
21. Tancredi, A., Liseo, B.: A hierarchical Bayesian approach to record linkage and population size problems. Ann. Appl. Stat. **5**(2B), 1553–1585 (2011). MR2849786
22. Winkler, W.E.: String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage. In: Proceedings of the Section on Survey Research Methods. ASA, Alexandria, VA, pp. 354–359 (1990)
23. Winkler, W.E.: Overview of record linkage and current research directions. Technical report. Statistics #2006-2, U.S. Bureau of the Census (2006)
24. Winkler, W.E.: Matching and record linkage. Wiley Interdisc. Rev.: Comput. Stat. **6**(5), 313–325 (2014)
25. Zanella, G., Betancourt, B., Wallach, H., Miller, J., Zaidi, A., Steorts, R.C.: Flexible models for microclustering with application to entity resolution. In: Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS 2016, NY, USA. Curran Associates Inc., pp. 1425–1433 (2016)
26. Zhao, B., Rubinstein, B.I.P., Gemmell, J., Han, J.: A Bayesian approach to discovering truth from conflicting sources for data integration. Proc. VLDB Endow. **5**(6), 550–561 (2012)