Record Linkage Modeling in Federal Statistical Databases

Michael D. Larsen

George Washington University, Department of Statistics
Biostatistics Center, 6110 Executive Blvd., Rockville, MD 20850, mlarsen@bsc.gwu.edu

1 Introduction

Record linkage (e.g., Fellegi and Sunter 1969, Newcombe *et al.* 1959) involves comparing two or more files on the same population for purposes of unduplication of records and merging files. Record linkage is used in many applications, including population size estimation at the U.S. Census Bureau (Winkler 1994, 1995, and Jaro 1989, 1995), epidemiology and medical studies (Newcombe 1988, Gill 1997), sociological studies (Belin *et al.* 2004), survey frame improvement, and, more recently, counterterrorism (Gomatam and Larsen 2004). See also Alvey and Jamerson (1997) and references therein.

Latent class (McCutcheon 1987) and mixture models (McLachlan and Peel 2000) have been used to model the data arising from comparing records in two files (Larsen and Rubin 2001, Winkler 1988, 1994, 1995, Jaro 1989, 1995). Although successful in many applications (Alvey and Jamerson 1997), the models used in these applications have not accounted for all restrictions in the data. In particular, forcing each record on one file to have at most a single, matching record on the other file ("one-to-one matching") has been implemented post-hoc with a one-to-one, linear-sum assignment procedure (Burkard and Derigs 1980, Jaro 1989) to choose individual links. The one-to-one assignment procedure can effectively eliminate many candidate links that have some degree of similarity, but actually are nonlinks.

Experience from previous record linkage operations has been used informally to select models (Larsen and Rubin 2001) and restrict parameters (Winkler 1989, 1994). Bayesian approaches to record linkage have been suggested by Larsen (1999a, 2002, 2004, 2005), Fortini *et al.* (2002, 2000), and McGlinchy (2004). A procedure is described here that explicitly uses the one-to-one matching assumption and allows parameter values to vary by file block, which is a subset of the data being linked. The approach will necessarily be Bayesian, because of the relatively small sample sizes within blocks and the difficulty of calculating expectations under complex restrictions on unobserved data.

This article is organized as follows. Record linkage is introduced in section 2. Bayesian record linkage algorithms are presented in section 3. Section 4 is a conclusion and discusses future work. Computational procedures are presented in appendices Appendix A and Appendix B. References are given in section 5.

2 Record Linkage

Suppose that there are two files, A and B, on a single population. Consider record a in file A and record b in B. Do records a and b correspond to the same person or entity? If files A and B ***do not*** have unique, accurately recorded identification numbers for every unit in both files, then it is necessary to consider the information recorded in a and b in order to answer the question.

Define agreement for each piece of information common to both files. Decennial census applications at U.S. Census Bureau record variables including last name and first name, street number and name, age, sex, race, and relation to head of household. Files are extensively preprocessed before linkage is attempted. For example, names are standardized and coded according to Soundex codes or other scheme. Names and address fields are parsed and standardized. In the case of simple comparisons, for each pair of records (a, b) being considered, a vector of 1's and 0's indicating agreement and disagreement on K comparison fields is recorded. That is, for $a \in A$ and $b \in B$, define

$$\gamma(a,b) = \{\gamma(a,b)_1, \gamma(a,b)_2, \ldots, \gamma(a,b)_K\}$$

where $\gamma(a,b)_k$ equals 1 (agreeement) or 0 (disagreement) on field k, $k = 1, \ldots, K$. Many agreements ($\gamma(a,b)$ mostly 1's) are typical of matches. Many disagreements ($\gamma(a,b)$ mostly 0's) are typical of nonmatches. Some variables (e.g., race) are informative in some locations regarding matches and nonmatches, but not in others. Disagreement on sex suggests a nonmatch, whereas agreement on sex is not persuasive by itself for being a match.

2.1 Data with One-to-One Restrictions and Blocking

In some cases, it is assumed that the two data files do not contain duplicate records for any person or entity. In the case of the census, records are organized by geographical location, each household should have only one form, and efforts are made at unduplicating records. Insurance companies and medical records systems are updated continuously

Table 1: Illustration of possible matching structure within a household.

Housing unit i		File B			
		Person 1	Person 2	...	Person n_{B_i}
File A	Person 1	**same**	different	...	different
	Person 2	different	**same**	...	different
	...	...	...	...	
	Person n_{A_i}	different	different		

and efforts are made to avoid duplicate records. Record linkage could be of interest in these cases. Census follow-up operations are conducted independently in large areas. The National Death Index is matched to existing insurance, medical, and other databases for studies such as Livingston and Ko (2005), Rauscher and Sandler (2005), Thompson *et al.* (2005), and many others.

In census and other operations, the files are divided geographically into groups of records or 'blocks' that do not overlap. Blocking is used in other applications as well in order to reduce the number of record pairs being compared. It is assumed that there are no (or very few) matches across different blocks. Other operations use first letter of last name (individuals) or industry code (businesses) or state as blocking variables.

Let blocks be indexed by $s = 1, \ldots, S$. Suppose that file A has n_{a_s} records and file B has n_{b_s} records, respectively, in block s. For blocks $s = 1, \ldots, S$, $a_s = 1, \ldots, n_{a_s}$ and $b_s = 1, \ldots, n_{b_s}$, define

$$I(a_s, b_s) = \begin{cases} 1 & a \text{ and } b \text{ are matches} \\ 0 & a \text{ and } b \text{ are nonmatches} \end{cases}$$

The set of match-nonmatch indicators in block s is $I_s = \{I(a_s, b_s)\}$. The one-to-one restrictions and blocking assumptions mean that $\sum_{b_s} I(a_s, b_s) \leq 1$, $\sum_{a_s} I(a_s, b_s) \leq 1$, and $\sum_{a_s} \sum_{b_{s'}} I(a_s, b_{s'}) = 0$ for $s \neq s'$. The number of matches in block s, n_{m_s} is defined and restricted under one-to-one matching as follows:

$$\sum_{a_s} \sum_{b_s} I(a_s, b_s) = n_{m_s} \leq \min(n_{a_s}, n_{b_s}).$$

In census applications, there could be further structure in the data. Within a housing unit, if there are no duplicate listings, there should be unique matches. The situation is illustrated in Table 1. Such an explicit structure is not imposed here, because missing values and recording errors might make exact household linkages problematic, but blocking might be less of a problem. Further, blocks at the U. S. Census Bureau use first letter of last name as a blocking criterion, in addition to geography, thereby splitting some households. Information on individuals within households is critical for differentiating similar records, such as father-son, mother-daughter, husband-wife, brother-brother, sister-sister pairs, etc., across files, and actual matching pairs. Another reason for not blocking on household is that address listings, especially in multiple household dwelling, might not be unique or reported values might be nonunique.

2.2 Prior beliefs and logical relationships

Prior experience and data often are available from previous record linkage operations and sites. In previous record linkage studies, clerks at the U.S. Census Bureau looked at record pairs and determined whether or not they truly were nonmatches or matches. Belin (1993, 1995), Larsen (1999b), and Larsen and Rubin (2001) found that in some U.S. Census Bureau record linkage applications characteristics of populations being studied varied by area in ways that made a significant impact on estimates of parameters needed for record linkage. There were, however, consistent patterns across areas. The percentage of record pairs, one record from each of two files, under consideration that actually are matches corresponding to the same person is roughly similar across sites. The probability of agreeing on some key fields of information among matches and nonmatches are similar across sites. The probability of agreements are higher among matches than among nonmatches. There is, however, variability across sites in these and many other characteristics.

It is expected that the probability of agreeing on an individual field of comparison is higher for matches than for nonmatches:

$$P(\gamma_k(a, b) = 1 | (a, b) \in \text{Match}) > P(\gamma_k(a, b) = 1 | (a, b) \in \text{Nonmatch}).$$

Logically, the number of matches in block s, n_{m_s}, is smaller than the smaller of the number of records in file A (n_{a_s}) and in file B (n_{b_s}). So the probability of a match in block s, p_{sM} is less than or equal to the minimum size ($\min(n_{a_s}, n_{b_s})$) divided by the number of pairs in block s: $n_{a_s} n_{b_s}$.

3 The Bayesian Record Linkage Model and Computations of Posterior Distributions

Section 3.1 presents a Bayesian version of record linkage for the mixture model approach of Fellegi and Sunter (1969), Larsen and Rubin (2001), and others, as described previously by Larsen (1999a). Section 3.2 contains the details of a hierarhical Bayesian model for record linkage suggested in part by Larsen (2004, 2005). Section 3.3 discusses the outline of the algorithm for simulating the posterior distribution of unknown parameters and unobserved matching indicators. Sections 3.4 and 3.5 present models and algorithms for incorporating one-to-one matching. Appendices Appendix A and Appendix B provide details of the iterative simulation algorithm.

3.1 Bayesian Approach to Latent Class Record Linkage Models

The mixture model approach to record linkage models the probability of a comparison vector γ as arising from a mixture distribution:

$$\Pr(\gamma) \quad = \quad \Pr(\gamma|M)p_M + \Pr(\gamma|U)p_U, \tag{1}$$

where $\Pr(\gamma|M)$ and $\Pr(\gamma|U)$ are the probabilities of the pattern γ among the matches (M) and nonmatches (U), respectively, and p_M and $p_U = 1 - p_M$ are marginal probabilities of matches and unmatched pairs. In practice at the U.S. Census and Statistics Canada, models using three classes often are useful when matching individuals because estimates based on them reflect household structure (see, e.g., Larsen and Rubin 2001, Armstrong and Mayda 1993, and Winkler 1995). Databases on businesses in general would not reflect the household grouping typical of people. We will consider the situation with two classes and comment on extensions in the discussion.

The conditional independence assumption simplifies the model by reducing the dimension within each mixture class from $2^K - 1$ parameters to K:

$$\Pr(\gamma|C) \quad = \quad \prod_{k=1}^{K} \Pr(\gamma_k|C)^{\gamma_k}(1 - \Pr(\gamma_k|C))^{1-\gamma_k}, \tag{2}$$

with $C \in \{M, U\}$. Interactions between comparison fields have been allowed in Larsen and Rubin (2001), Armstrong and Mayda (1993), Thibaudeau (1993), Winkler (1989), and others. Here we consider only the conditional independence model and extensions of it to a hierarchical framework.

Previous approaches have not directly enforced one-to-one linkage in the likelihood and have used the following likelihood function:

$$\prod_{s=1}^{S} \prod_{a \in A_s, b \in B_s} \Pr(\gamma(a, b)), \tag{3}$$

where $\Pr(\gamma(a, b))$ is a comparison vector modeled using the mixture assumption (1). When the parameters determining $\Pr(\gamma|M)$ and $\Pr(\gamma|U)$ do not depend on the block from which the pairs originate and n_γ is the number of pairs of records with comparison pattern γ, the simple likelihood can be written as

$$\prod_{\gamma \in \Gamma} \Pr(\gamma)^{n_\gamma}.$$

Assuming the conditional independence model (2) and global parameters that do not vary by block, a prior distribution on parameters can be specified conveniently as the product of independent Beta distributions as follows:

$$p_M \sim \text{Beta}(\alpha_M, \beta_M),$$

$$\Pr(\gamma_k(a, b) = 1|M) \sim \text{Beta}(\alpha_{Mk}, \beta_{Mk}), k = 1, \ldots, K,$$

and

$$\Pr(\gamma_k(a, b) = 1|U) \sim \text{Beta}(\alpha_{Uk}, \beta_{Uk}), k = 1, \ldots, K.$$

Instead of specifying the prior distribution in this manner, it would conceptually be possible to specify a prior distribution on the whole of the probability vector associated with the set of comparison vectors γ as two Dirichlet distributions. That is, independent prior distributions $\Pr(\gamma|M) \sim \text{Dirichlet}(\delta_M)$ and $\Pr(\gamma|U) \sim \text{Dirichlet}(\delta_U)$ could be specified. This option is not explored in this paper. It is noted, however, that pairs of records with known match status could be used as "training data" (as in Belin and Rubin 1995) for the purposes of specifying a prior distribution. The prior parameter values, δ_M and δ_U, could be considered as 'prior counts' by agreement vector pattern in the matches and nonmatches.

The match/nonmatch indicators $\mathbf{I} = \{I(a,b), a \in A_s, b \in B_s, s = 1, \ldots, S\}$ are unobserved. By Bayes' theorem, if the parameters were known and one does not consider restrictions from one-to-one matching, one could calculate for a pair (a,b) the probability that a and b match:

$$\Pr(I(a,b) = 1|\gamma(a,b)) \;=\; \Pr(M|\gamma(a,b)) \quad = \quad p_M \Pr(\gamma(a,b)|M)/\Pr(\gamma(a,b)) \tag{4}$$

with the denominator given by (1).

If the match indicators $\mathbf{I}$ were known, the posterior distributions of individual parameters given values of the other parameters would be as follows:

$$p_M|\mathbf{I} \quad \sim \quad \text{Beta}(\alpha_M + \sum_{(a,b)} I(a,b),\; \beta_M + \sum_{(a,b)} (1 - I(a,b)), \tag{5}$$

$$\Pr(\gamma_k(a,b) = 1|M, \mathbf{I}) \quad \sim \quad \text{Beta}(\alpha_{Mk} + \sum I_{ab}\gamma_k(a,b), \beta_{Mk} + \sum I_{ab}(1 - \gamma_k(a,b))) \tag{6}$$

for $k = 1, \ldots, K$, and

$$\Pr(\gamma_k(a,b) = 1|U, \mathbf{I}) \quad \sim \quad \text{Beta}(\alpha_{Uk} + \sum (1 - I_{ab})\gamma_k(a,b), \beta_{Uk} + \sum (1 - I_{ab})(1 - \gamma_k(a,b))) \tag{7}$$

for $k = 1, \ldots, K$, where $I_{ab} = I(a,b)$ and sums are over all pairs allowed within the blocking structure.

The posterior distribution of parameters is simulated by sampling from alternating conditional distributions (Gibbs sampling; Geman and Geman 1984, Geland and Smith 1990) as follows.

1. Specify parameters for the prior distributions. Choose initial values of unknown parameters.
2. Repeat the following four steps numerous times until the distribution of draws has converged to the posterior distribution of interest.
   (a) Draw values for the components of I independently from Bernoulli distributions with the probability that $I(a,b) = 1$ given by formula (4).
   (b) Draw a value of p_M from the distribution specified in formula (5) and calculate $p_U = 1 - p_M$.
   (c) Draw values of $\Pr(\gamma_k(a,b) = 1|M, \mathbf{I})$ independently for $k = 1, \ldots, K$ from distributions specified in formula (6).
   (d) Draw values of $\Pr(\gamma_k(a,b) = 1|U, \mathbf{I})$ independently for $k = 1, \ldots, K$ from distributions specified in formula (7).
3. Stop once the algorithm has converged. Convergence of the algorithm can be monitored by comparing distributions from multiple independent series as suggested by Gelman and Rubin (1992) and Brooks and Gelman (1998).

Once the algorithm has converged, it is necessary to decide which pairs of records to designate links and nonlinks and which to send to clerical review or leave undecided. If one-to-one restrictions are not enforced, then one could calculate the proportion of times that a record pair (a,b) has $I(a,b) = 1$ and for each record in file A assign the record in file B that has the largest proportion. If one-to-one matching is desired, the simulated probabilities (4) of matching could be supplied to a linear-sum-assignment algorithm.

There are some *restrictions on parameters* that could improve the performance of this model for record linkage. First, the range of p_M logically should be restricted to be less than or equal to the smaller of the two file sizes divied by the number of pairs under the blocking structure. When p_M is drawn in the Gibbs sampling algorithm from its conditional distribution, values of p_M greater than the cutoff should not be used. Alternatively, if $p_M = c_M p'_M$ where p'_M has the Beta distribution given above and $c_M < 1$ is a scale factor appropriate for transforming p'_M to the allowable range of p_M, one can sample p'_M and scale it by c_M. Second, logically the probability of a record pair agreeing on a comparison field should be larger among matches than among nonmatches. That is, $\Pr(\gamma_k|M) > \Pr(\gamma_k|U)$, for $k = 1, \ldots, K$. Such a restriction can be added to the Gibbs sampling algorithm by simply ignoring sampled pairs of these probabilities that do not satisfy the constraint. Alternatively, one can draw one value, say $\Pr(\gamma_k|M)$, and scale the value of $\Pr(\gamma_k|U)$ to be in the range $(0, \Pr(\gamma_k|M))$. That is, after drawing a value of $\Pr(\gamma_k|M)$, draw a value from the Beta distribution specified in the algorithm and multiply it by $\Pr(\gamma_k|M)$.

There are several significant limitations to this model beyond the logical restrictions already noted. There is no explicit one-to-one matching in the likelihood (3) and without subsequent processing some records could be involved

in more than one designated link. As a consequence, it was not necessary to model the number of matches overall or within individual blocks. In many applications, some records in file A and some in file B might not have any matches. One-to-one matching then is the assumption that records have at most one match in the other file. The parameters are global and do not vary across blocks despite the fact that populations can vary greatly across blocks. The conditional independence assumption was made for convenience and is not realistic. It has been relaxed in the case of maximum likelihood estimation (see Larsen and Rubin 2001 and references therein). Interactions between comparison fields within the matches and nonmatches could be allowed in the Bayesian approach as well. It is the belief of the author, however, that explicitly modeling (at most) one-to-one matching and allowing parameters to vary by block will be more beneficial than modeling interactions for entire operations.

3.2 A Hierachical Bayesian Model

A hierarchical model for record linkage will specify distributions of parameters within blocks $s = 1, \ldots, S$. In this section, the specification of the model is given. The outline of the iterative simulation algorithm is given in Section 3.3. Details of the Metropolis-Hastings step can be found in Appendix Appendix A. Consideration of a Bayesian model incorporating the one-to-one restriction begins in Section 3.4. That is, the likelihood used in this section is given by likelihood (3) with parameters varying by block.

The probabilities of agreeing on fields of information are allowed to vary by block as follows:

$$p_{sMk} = \Pr(\gamma_k = 1|M, s) \sim \text{Beta}(\alpha_{sMk}, \beta_{sMk})$$

and

$$p_{sUk} = \Pr(\gamma_k = 1|U, s) \sim \text{Beta}(\alpha_{sUk}, \beta_{sUk})$$

independently across blocks, fields, and classes (M and U). The restriction that $p_{sMk} \geq p_{sUk}$ will be assumed.

Hyperpriors distributions are placed on transformed versions of the Beta parameters. The distributions are independent across blocks, fields, and groups. These transformations were suggested by Andrew Gelman (2004) and incorporated into Larsen (2004):

$$\theta_{sMk} = \text{logit}(\frac{\alpha_{sMk}}{\alpha_{sMk} + \beta_{sMk}}) \sim N(\mu_{\theta Mk}, \sigma^2_{\theta Mk}),$$

$$\theta_{sUk} = \text{logit}(\frac{\alpha_{sUk}}{\alpha_{sUk} + \beta_{sUk}}) \sim N(\mu_{\theta Uk}, \sigma^2_{\theta Uk}),$$

$$\tau_{sMk} = \log(\alpha_{sMk} + \beta_{sMk}) \sim N(\mu_{\tau Mk}, \sigma^2_{\tau Mk}),$$

and

$$\tau_{sUk} = \log(\alpha_{sUk} + \beta_{sUk}) \sim N(\mu_{\tau Uk}, \sigma^2_{\tau Uk}).$$

Note that there is a unique bivariate inverse transformation: $\alpha_{sCk} = e^{\tau_{sCk}}\text{logit}^{-1}(\theta_{sCk})$ and $\beta_{sCk} = e^{\tau_{sCk}}\text{logit}^{-1}(1 - \theta_{sCk})$ for $C = M, U$. One could also consider correlation in the hyper parameter distribution for block s among the θ's, among the τ's, and between the θ's and τ's.

The restriction noted in the previous paragraph *does not* mean that, for $k = 1, \ldots, K$, $\theta_{sMk} \geq \theta_{sUk}$; the restriction only constrains the parameters p_{sMk} and p_{sUk}. It would be possible to use a prior distribution with the constraint that $\theta_{sMk} \geq \theta_{sUk}$ as well.

The probability of belonging to class M in block s, p_{sM}, is given a Beta$(\alpha_{sM}, \beta_{sM})$ prior distribution. The hyperprior distrbutions are

$$\theta_{sM} = \text{logit}(\frac{\alpha_{sM}}{\alpha_{sM} + \beta_{sM}}) \sim N(\mu_{\theta M}, \sigma^2_{\theta M})$$

and

$$\tau_{sM} = \log(\alpha_{sM} + \beta_{sM}) \sim N(\mu_{\tau M}, \sigma^2_{\tau M}),$$

and are independent of the other hyperpriors. The restriction that p_{sM} is smaller than the minimum of n_{A_s} and n_{B_s} divided by the number of pairs $n_{A_s}n_{B_s}$ is enforced in this model. If it were not, the small sample size and great variability across blocks would surely produce poor results for some blocks. Note that $\alpha_{sM} = e^{\tau_{sM}}\text{logit}^{-1}(\theta_{sM})$ and $\beta_{sM} = e^{\tau_{sM}}\text{logit}^{-1}(1 - \theta_{sM})$.

3.3 Simulating the Hierachical Model Posterior Distribution

The posterior distribution of parameters and unobserved match/nonmatch indicators will be simulated using Gibbs sampling. The conditional distributions for the hyperparameters will be sampled using the Metropolis-Hastings (MH) algorithm (Hastings 1970) within the Gibbs sampling framework. The procedure iterates through draws of full conditional distributions as described below.

1. Choose hyperparameter distributions. That is, specify $(\mu_{\theta M}, \sigma^2_{\theta M})$ and, for $k = 1, \ldots, K$, specify $(\mu_{\theta Mk}, \sigma^2_{\theta Mk})$, $(\mu_{\theta Uk}, \sigma^2_{\theta Uk})$, $(\mu_{\tau Mk}, \sigma^2_{\tau Mk})$, and $(\mu_{\tau Uk}, \sigma^2_{\tau Uk})$.

2. Generate initial values of $(\alpha_{sM}, \beta_{sM})$ and, for $k = 1, \ldots, K$, $(\alpha_{sMk}, \beta_{sMk})$ and $(\alpha_{sUk}, \beta_{sUk})$ from their prior distributions.

3. Assign an initial match/nonmatch configuration $\mathbf{I}$. Since one-to-one matching is not being forced, but constraints on the parameters and proportion of matches are, the algorithm of section 3.1 with analogous parameter constraints could be run for several iterations. An alternative is to randomly generate a matrix of 1's and 0's in each block to represent match/nonmatch status.

4. Cycle through the following steps numerous times until the distribution of drawn values converges to the target posterior distribution. Let I_{ab} denote $I(a, b)$.

   (a) For $s = 1, \ldots, S$, draw p_{sM} from its conditional distribution given the current indicators $\mathbf{I}_s$ and values of $(\alpha_{sM}, \beta_{sM})$. Specifically,

   $$p_{sM} | I_s, \alpha_{sM}, \beta_{sM} \sim \text{Beta}(\alpha_{sM} + \sum I_{ab}, \beta_{sM} + n_{a_s} n_{b_s} - \sum I_{ab}),$$

   where the sum is over all pairs (a, b) in block s. Enforce the constraint

   $$p_{sM} \leq \min(n_{a_s}, n_{b_s})/(n_{a_s} n_{b_s}).$$

   (b) For $s = 1, \ldots, S$ and $k = 1, \ldots, K$, draw p_{sMk} and p_{sUk} from their conditional distribution given the current indicators $\mathbf{I}_s$, the comparison vectors γ_s in block s, and values of $(\alpha_{sCk}, \beta_{sCk}), C \in \{M, U\}$. Specifically,

   $$p_{sMk} | I_s, \gamma_s, \alpha_{sMk}, \beta_{sMk} \sim \text{Beta}(\alpha_{sMk} + \sum_s I_{ab} \gamma_k(a, b), \beta_{sMk} + \sum_s I_{ab}(1 - \gamma_k(a, b))),$$

   $$p_{sUk} | I_s, \gamma_s, \alpha_{sUk}, \beta_{sUk} \sim \text{Beta}(\alpha_{sUk} + \sum_s (1 - I_{ab}) \gamma_k(a, b), \beta_{sUk} + \sum_s (1 - I_{ab})(1 - \gamma_k(a, b))),$$

   and $p_{sMk} \geq p_{sUk}$, where sums are over all pairs (a, b) in block s.

   (c) For $s = 1, \ldots, S$, use the Metropolis-Hastings algorithm (Hastings 1970; see also Gelman 1992 and Gelman *et al.* 2004, chapter 11) to draw values of hyperparameters θ_{sM} and τ_{sM} from their full conditional distributions. See appendix Appendix A for details of this and the next two steps.

   (d) For $s = 1, \ldots, S$ and $k = 1, \ldots, K$, use the Metropolis-Hastings algorithm to draw values of hyperparameters θ_{sMk} and τ_{sMk}.

   (e) For $s = 1, \ldots, S$ and $k = 1, \ldots, K$, use the Metropolis-Hastings algorithm to draw values of hyperparameters θ_{sUk} and τ_{sUk}.

   (f) For $s = 1, \ldots, S$, $a = 1, \ldots, n_{a_s}$, and $b = 1, \ldots, n_{b_s}$, given values of p_{sM} and, for $k = 1, \ldots, K$, p_{sMk} and p_{sUk}, draw a value of $I(a, b)$ from a Bernoulli distribution with the following probability:

   $$\frac{p_{sM} \prod_{k=1}^K \left[p_{sMk}^{\gamma_k(a,b)} (1 - p_{sMk})^{1 - \gamma_k(a,b)} \right]}{\left\{ p_{sM} \prod_{k=1}^K \left[p_{sMk}^{\gamma_k(a,b)} (1 - p_{sMk})^{1 - \gamma_k(a,b)} \right] + (1 - p_{sM}) \prod_{k=1}^K \left[p_{sUk}^{\gamma_k(a,b)} (1 - p_{sUk})^{1 - \gamma_k(a,b)} \right] \right\}}.$$

5. Stop once the algorithm has converged.

Note that one-to-one restrictions are not imposed on the $\mathbf{I}$ matrix. The size of the candidate match class in each block is controlled in 4(a) by keeping p_{sM} small. Once the algorithm has converged, it is necessary to decide which pairs of records to designate links and nonlinks and which to send to clerical review or leave undecided. Suggestions were made at the end of section 3.1. Metropolis-Hastings and algorithm details are in appendix Appendix A.

3.4 A Hierachical Bayesian Model with One-to-One Restrictions

In this section, the one-to-one linkage assumption will be enforced in the set of indicators **I**. The simulation for this model is described in section 3.5. The hierarchical specification of section 3.2 will continue to be used. In order to use the non-hierarchical model with one-to-one restrictions, one would need to combine the appropriate modeling assumptions and prior distributions from section 3.1 and this section. Algorithms for sampling from the posterior distribution would then combine appropriate steps from section 3.3 and the next section.

Define n_{m_s} to be the number of matches in block s, $s = 1, \ldots, S$. By definition, $n_{m_s} \leq \min(n_{a_s}, n_{b_s})$. The prior distribution for n_{m_s}, independently for each s, is taken to be

$$n_{m_s} \sim \text{Binomial}(\min(n_{a_s}, n_{b_s}), p_s), \tag{8}$$

where $p_s \sim \text{Beta}(\alpha_p, \beta_p)$. If $\alpha_p = 4$ and $\beta_p = 1$, then $Ep_s = 0.8$, $SDp_s = 0.16$, and the distribution is skewed strongly left. If $\alpha_p = 4.5$ and $\beta_p = 1.5$, then $Ep_s = 0.75$, $SDp_s = 0.16$, and the distribution is skewed left, but not quite so strongly. The parameters p_{sM} do not play a role in this model.

Let $I_s = \{I(a,b), a \in A_s, b \in B_s\}$ for $s = 1, \ldots, S$. The prior distribution for I_s is taken to be uniform on the space of possible matching configurations:

$$P(I_s|n_{m_s}) = \left[\begin{pmatrix} n_{a_s} \\ n_{m_s} \end{pmatrix} \begin{pmatrix} n_{b_s} \\ n_{m_s} \end{pmatrix} n_{m_s}! \right]^{-1}. \tag{9}$$

Without examining the data to some degree, it would not be possible to assign another prior distribution. In the census application, it would be reasonable if records are grouped by household to place higher probability on records in the same household within blocks. One would probably gain by placing higher probabilty of matching on the first record listed in each household and continuing with descreasing probabilty through the household list. The task of enumerating such a prior distribution, however, would be quite cumbersome. As was mentioned earlier, address information might have its limitations and blocking by first letter of last name at census sometimes separates household members, so placing too much reliance on exact address correspondence might have a downside in some neighborhoods.

Given values for the components of I, the likelihood for parameters based on the comparisons of recorded information is

$$\begin{aligned} \Pr(\gamma|I) &= \prod_{s=1}^{S} \left[\prod_{a \in A_s, b \in B_s} \left(\prod_{k=1}^{K} p_{sMk}^{\gamma_k(a,b)} (1 - p_{sMk})^{1-\gamma_k(a,b)} \right)^{I(a,b)} \right. \\ &\qquad \left. \left(\prod_{k=1}^{K} p_{sUk}^{\gamma_k(a,b)} (1 - p_{sUk})^{1-\gamma_k(a,b)} \right)^{1-I(a,b)} \right] \\ &= \prod_{s=1}^{S} \left[\prod_{a \in A_s, b \in B_s, (a,b) \in M} \prod_{k=1}^{K} p_{sMk}^{\gamma_k(a,b)} (1 - p_{sMk})^{1-\gamma_k(a,b)} \right. \\ &\qquad \left. \prod_{a \in A_s, b \in B_s, (a,b) \in U} \prod_{k=1}^{K} p_{sUk}^{\gamma_k(a,b)} (1 - p_{sUk})^{1-\gamma_k(a,b)} \right] \end{aligned} \tag{10}$$

As mentioned before, the parameters p_{sM} are not used in this model.

Let the prior distributions for p_{sMk} and p_{sUk}, $s = 1, \ldots, S$, $k = 1, \ldots, K$ and their associated hyperprior distributions be the same as in section 3.2.

3.5 Simulating the Hierarchical One-to-one Model Posterior Distribution

The posterior distribution of parameters and unobserved match/nonmatch indicators will be simulated using Gibbs sampling with Metropolis-Hastings steps. The procedure iterates through draws of full conditional distributions as described below.

1. Choose hyperparameter distributions. That is, specify α_p and β_p and, for $k = 1, \ldots, K$, means and variances $(\mu_{\theta Mk}, \sigma^2_{\theta Mk})$, $(\mu_{\theta Uk}, \sigma^2_{\theta Uk})$, $(\mu_{\tau Mk}, \sigma^2_{\tau Mk})$, and $(\mu_{\tau Uk}, \sigma^2_{\tau Uk})$.

2. Generate initial values in blocks $s = 1, \ldots, S$ for matching variables $k = 1, \ldots, K$ of $(\alpha_{sMk}, \beta_{sMk})$ and $(\alpha_{sUk}, \beta_{sUk})$ from their prior distributions.

3. Assign an initial match/nonmatch configuration $\mathbf{I}$. Since one-to-one matching is being forced, the algorithms of sections 3.1 and 3.2 with appropriate constraints on parameters followed by a linear sum assignment procedure (Burkard and Derigs 1980) could be used to produce an initial $\mathbf{I}$. An alternative approach is to randomly generate a matrix of 1's and 0's with at most one '1' in each row and column in each block to represent match/nonmatch status. In block s, $n_{m_s} = \sum_{a \in A_s} \sum_{b \in B_s} I(a,b)$.

4. Cycle through the following steps numerous times until the distribution of drawn values converges to the target posterior distribution.

   (a) For $s = 1, \ldots, S$, draw p_s from its conditional distribution given the current indicators $\mathbf{I}_s$ (and hence n_{m_s}) and values of (α_p, β_p). Specifically,

   $$p_s | I_s, \alpha_p, \beta_p \sim \text{Beta}(\alpha_p + n_{m_s}, \beta_p + \min(n_{a_s}, n_{b_s}) - n_{m_s}).$$

   (b) For $s = 1, \ldots, S$ and $k = 1, \ldots, K$ draw p_{sMk} and p_{sUk} from their conditional distribution as described in step (4b) of section 3.3.

   (c) For $s = 1, \ldots, S$ and $k = 1, \ldots, K$, use the Metropolis-Hastings algorithm to draw values of hyperparameters θ_{sMk} and τ_{sMk} as described in appendix Appendix A step (d).

   (d) For $s = 1, \ldots, S$ and $k = 1, \ldots, K$, use the Metropolis-Hastings algorithm to draw values of hyperparameters θ_{sUk} and τ_{sUk} as described in appendix Appendix A step (e).

   (e) For $s = 1, \ldots, S$, use the Metropolis-Hastings algorithm to draw values of $\mathbf{I}_s$ and n_{m_s} from their full conditional distributions. See appendix Appendix B for details of this step.

5. Stop once the algorithm has converged.

Note that one-to-one restrictions are imposed on the $\mathbf{I}$ matrix. The size of the match class in block s is explicitly controlled by the fact that $n_{m_s} \leq \min(n_{a_s}, n_{b_s})$; $0 < p_s < 1$. Once the algorithm has converged, it is necessary to decide which pairs of records to designate as links and nonlinks and which to send to clerical review or leave undecided. Suggestions in this regard were made at the end of section 3.1. Appendix Appendix B contains details for step 4(e) above and the Metropolis-Hastings implementation.

4 Conclusions and Future Work

A novel hierarchical Bayesian model for record linkage has been presented. The model allows probabilities to vary by block and reflect local information. One-to-one matching restrictions are imposed in the likelihood. Indicators of match status are sampled using Gibbs sampling and the Metropolis-Hastings algorithm. Simulations could be used to evaluate the performance of the proposed methods.

Several areas can be identified for future work. Many of these will be important in actual applications. It will be interesting to apply these methods to data from the U.S. Census Bureau, the U.S. National Center for Health Statistics, and other sources. An automated system for applying these models to new sets of files would be useful in this regard. In a real application, one could consider better specifications of prior distributions for the record linkage model parameters. In particular, if data are available from another record linkage site and the site differs in some ways from the current application, then one must decide the degree to which data from the previous site should be discounted or down weighted when analyzing the new site. In some applications, the size of the files will be a challenge. In order to speed computations, one might consider parallel computations; for example, many computations are performed separately in each block.

The algorithm's perfomance could be improved by studying tuning parameters and the order of sampling cycles within Metropolis-Hastings and Gibbs sampling algorithms. One could study the sensitivity of results to the specification of hyperprior distributions. If some Metropolis-Hastings draws for some parameters and elements of $\mathbf{I}$ infrequently lead to changes in the values, then one could examine methods for increaseing the frequency of accepting Metropolis-Hastings' moves. In particular, one could consider combining two or more attempted moves into one step.

Two extensions related to the record linkage model can be studied. First, one can consider expanded definitions of the agreement/disagreement comparisons for the matching variables. That is, one could allow partial agreement, missing values, and string comparator metrics (Winkler 1993, 1994). These comparisons probably would use matching information more effeciently. Second, in some applications, one could consider more fully using household structure. In some applications at the U.S. Census Bureau, household structure is reflected in part by the use of three latent classes in the record linkage mixture model (Larsen and Rubin 2001 and references therein).

Another direction for development in the future is the Bayesian analysis of files that are created through record linkage operations. Lahiri and Larsen (2005) extended Scheuren and Winkler (1993) on adjusting for the bias that arises due to errors in matching. One could image a feed-back loop, as in Scheuren and Winkler (1997), where points with large residuals in a linear regression model are more likely than their agreement patterns alone suggest to be nonmatches and points that are very certain to be matches have more influence on a linear regression fit.

5 References

Alvey, W., and Jamerson, B. (1997), *Record Linkage Techniques – 1997*, Proceedings of an International Workshop and Exposition. Federal Committee on Statistical Methodology, Office of Management of the Budget.

Armstrong, J. B., and Mayda, J. E. (1993). Model-Based Estimation of Record Linkage Error Rates. *Survey Methodology*, 19, 137-147.

Belin, T. R. (1993). Evaluation of sources of variation in record linkage through a factorial experiment. *Survey Methodology* **19**, 13-29.

Belin, T. R., Ishwaran, H., Duan, N., Berry, S., and Kanouse, D. (2004). Identifying likely duplicates by record linkage in a survey of prostitutes. *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*. Gelman, A., and Meng, X. L., editors. New York: Wiley.

Belin, T. R., and Rubin, D. B. (1995). A method for calibrating false match rates in record linkage. *Journal of the American Statistical Association* **90**, 694- 707.

Brooks, S. P. , and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7 , 434-455

Burkard, R.E., and Derigs, U. (1980). Assignment and Matching Problems: Solution Methods with FORTRAN-Programs. *Lecture Notes in Economics and Mathematical Systems, No. 184*, Springer-Verlag: Berlin, Heidelberg, New York, pp. 1-11.

Fellegi, I. P., and Sunter, A. B. (1969), "A Theory for Record Linkage," *Journal of the American Statistical Association*, 64, 1183-1210.

Fortini, M., Liseo, B., Nuccitelli, A., and Scanu, M. (2000). "On Bayesian Record Linkage," *Bayesian Methods with Applications to Science, Policy, and Official Statistics*: Selected Papers from ISBA 2000: The Sixth World Meeting of the International Society for Bayesian Analysis. Editor E. I. George, 155-164.

Fortini, M., Nuccitelli, A., Liseo, B., and Scanu, M. (2002). "Modelling issues in record linkage: A Bayesian perspective," *Proceedings of the American Statistical Association, Survey Research Methods Section*, 1008-1013.

Gelfand, Alan E. , and Smith, Adrian F. M. (1990), "Sampling-based approaches to calculating marginal densities", Journal of the American Statistical Association, 85 , 398-409

Gelman, A. (1992). Iterative and non-iterative simulation algorithms. *Computing Science and Statistics* **24**, 433-438.

Gelman, A. (2004). Personal communication.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*, 2nd edition. Chapman & Hall/CRC.

Gelman, A., and Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science* 7, 457- 472.

Geman, S., and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6 , 721-741

Gill, L. E. (1997), "OX-LINK: The Oxford Medical Record Linkage System Demonstration of the PC Version," in *Record Linkage Techniques – 1997*, Proceedings of an International Workshop and Exposition. Federal Committee on Statistical Methodology, Office of Management of the Budget, page 491.

Gomatam, S., and Larsen, M.D. (2004), "Record Linkage and Counterterrorism," *Chance*, 17(1): 25-29.

Hastings, W. K. (1970). Monte Carlo sapling methods using Markov chains and their applications, *Biometrika*, 57, 97-109.

Jaro, M. A. (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.

Jaro, M. A. (1995), "Probabilistic Linkage of Large Public Health Data Files, " *Statistics in Medicine*, 14, 491-498.

Lahiri, P., and Larsen, M.D. (2005). Regression Analysis with Linked Data. *Journal of the American Statistical Association*, 100, 222-230.

Larsen, M.D. (1999a), "Multiple Imputation Analysis of Records Linkage Using Mixture Models,," *Proceedings of the Statistical Society of Canada, Survey Methods Section*, 65-71.

Larsen, M.D. (1999b). "Predicting the Residency Status for Administrative Records that Do Not Match Census Records," *Administrative Records Research Memorandum Series*, #20, Bureau of the Census, U.S. Department of Commerce.

Larsen, M.D. (2002), "Comment on Hierarchical Bayesian Record Linkage," *Proceedings of the Section on Bayesian Statistical Science*, American Statistical Association meeting in New York City. CDROM: 1995-2000.

Larsen, M.D. (2004), Record linkage using finite mixture models, *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*. Gelman, A., and Meng, X. L., editors. New York: Wiley, 309-318.

Larsen, M.D. (2005), Hierarchical Bayesian record linkage. Technical report, Iowa State University, Department of Statistics.

Larsen, M. D., and Rubin, D. B. (2001), "Iterative automated record linkage using mixture models," *Journal of the American Statistical Association*, 96, 32-41.

Livingston, E. H., and Ko, C. Y. (2005). Effect of diabetes and hypertension on obesity-related mortality. *Surgery*, 137 (1): 16-25.

McCutcheon, A. L. (1987). *Latent class analysis*. Sage Publications, Inc.: Newbury Park, CA; London.

McGlinchy, M. (2004). A Bayesian record linkage methodology for multiple imputation of missing links. *Proceedings of the American Statistical Association, Section on Survey Research Methods*. Alexandria, VA: CDROM.

McLachlan, G.J., and Peel, D. (2000). *Finite Mixture Models*. New York: Wiley.

Newcombe, H. B. (1988), *Handbook of record linkage: Methods for health and statistical studies, administration, and business*, Oxford University Press: Oxford.

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

Rauscher, G. H., and Sandler, D. P. (2005). Validating cancer histories in deceased relatives. *Epidemiology*, 16 (2): 262-265.

Scheuren, F., and Winkler, W. E. (1993), "Regression analysis of data files that are computer matched," *Survey Methodology*, 19, 39-58.

Scheuren, F., and Winkler, W. E. (1997), "Regression analysis of data files that are computer matched – Part II," *Survey Methodology*, 23, 157- 165.

Thibaudeau, Y. (1993), "The Discrimination Power of Dependency Structures in Record Linkage," *Survey Methodology*, 19, 31-38.

Thompson, D., Kriebel, D., Quinn, M. M., Wegman, D. H., and Eisen, E. A. (2005). Occupational exposure to metalworking fluids and risk of breast cancer among female autoworkers. *American Journal of Industrial Medicine*, 47 (2): 153-160.

Winkler, W. E. (1988), "Using the EM Algorithm for Weight Computation in the Fellegi-Sunter Model of Record Linkage," *American Statistical Association Proceedings of Survey Research Methods Section*, pp. 667- 671.

Winkler, W. E. (1989), "Near automatic weight computation in the Fellegi-Sunter model of record linkage", *Proceedings of the Bureau of the Census Annual Research Conference*, 5, 145-155.

Winkler, W. E. (1993), "Improved decision rules in the Fellegi-Sunter model of record linkage," in *American Statistical Association Proceedings of Survey Research Methods Section*, pp. 274-279.

Winkler, W. E. (1994), "Advanced Methods for Record Linkage, " in *American Statistical Association Proceedings of Survey Research Methods Section*, pp. 467-472.

Winkler, W. E. (1995), "Matching and Record Linkage," in *Business Survey Methods*, ed. Cox, B. G., Binder, D. A., Chinnappa, B. N., Christianson, A., Colledge, M. J., and Kott, P. S., New York: Wiley Publications, pp. 355-384.

Appendix A Metropolis-Hastings Sampling Steps for the Hierarchical Record Linkage Model

Details of the three Metropolis-Hastings (Hastings 1970) steps in the simulation procedure of section 3.3 are presented below.

(c). For $s = 1, \ldots, S$, use the Metropolis-Hastings algorithm (Hastings 1970; see also Gelman 1992 and Gelman *et al.* 2004 chapter 11) to draw values of hyperparameters θ_{sM} and τ_{sM} from their full conditional distributions. Specifically, given current values of θ_{sM} and τ_{sM} (and hence α_{sM} and β_{sM}), $\mathbf{I}_s$, and other parameters,

(i) Define tuning constants $h_{\theta M} > 0$ and $h_{\tau M} > 0$.

(ii) Draw $u \sim \text{Uniform}(0, 1)$,

$$\theta^* \sim N(\theta_{sM}, \sigma^2_{\theta M}/h_{\theta M}),$$

and

$$\tau^* \sim N(\tau_{sM}, \sigma^2_{\tau M}/h_{\tau M}).$$

(iii) Calculate $\alpha^* = e^{\tau^*}\text{logit}^{-1}(\theta^*)$ and $\beta^* = e^{\tau^*}\text{logit}^{-1}(1-\theta^*)$.

(iv) Calculate

$$\begin{aligned} r &= \min\Big(1,\ p_{sM}^{\alpha^*-\alpha_{sM}}(1-p_{sM})^{\beta^*-\beta_{sM}} \\ &\quad \times \exp\big(-\frac{h_{\theta M}}{\sigma^2_{\theta M}}(\theta_{sM}-\theta^*)^2\big)\exp\big(-\frac{h_{\tau M}}{\sigma^2_{\tau M}}(\tau_{sM}-\tau^*)^2\big)\Big) \end{aligned}$$

(v) If $u \le r$, let $\theta_{sM} = \theta^*$ and $\tau_{sM} = \tau^*$.
Otherwise, let θ_{sM} and τ_{sM} remain the same.

(d). For $s = 1, \ldots, S$ and $k = 1, \ldots, K$, use the Metropolis-Hastings algorithm to draw values of hyperparameters θ_{sMk} and τ_{sMk}. Specifically, given current values of θ_{sMk} and τ_{sMk} (and hence α_{sMk} and β_{sMk}), $\mathbf{I}_s$, and other parameters, follow the steps outlined in step (c) above but with all M indexes replaced by Mk's.

(e). For $s = 1, \ldots, S$ and $k = 1, \ldots, K$, use the Metropolis-Hastings algorithm to draw values of hyperparameters θ_{sUk} and τ_{sUk}. Specifically, given current values of θ_{sUk} and τ_{sUk} (and hence α_{sUk} and β_{sUk}), $\mathbf{I}_s$, and other parameters, follow the steps outlined in step (c) above but with all M indexes replaced by Uk's.

The tuning parameters $h_{\theta M}$ and $h_{\tau M}$ are chosen so that the drawn values of the parameters are accepted approximately 23-44% of the time (Gelman *et al.* 2004 chapter 11.9). Thus the algorithm could be run for several iterations to assess the acceptance rate, adapting the tuning paramters as necessary. A second phase then could be initiated with fixed values for tuning parameters.

Appendix B Metropolis-Hastings Steps for Sampling from the Hierarchical Record Linkage Model with One-to-One Restrictions

In this section, the updating step for the number of matches, n_{m_s}, and the configuration of matches and nonmatches, $\mathbf{I}_s$, for blocks $s = 1, \ldots, S$ is described. It is assumed that current values of parameters and hyperparameters are given. Each block is updated separately. Given the value of a match/nonmatch configuration $\mathbf{I}_s$, the unknown parameters of the model are drawn as described in section 3.5.

Let $\gamma_s = \{\gamma(a,b), a \in A_s, b \in B_s\}$ be the collection of comparison vectors for all pairs in block s. For notational convenience, let $\alpha_s = (\alpha_{sMk}, \alpha_{sUk}, k = 1, \ldots, K)$, $\beta_s = (\beta_{sMk}, \beta_{sUk}, k = 1, \ldots, K)$, $\mu = (\mu_{\theta Mk}, \mu_{\theta Uk}, \mu_{\tau Mk}, \mu_{\tau Uk}, k = 1, \ldots, K)$, and $\sigma^2 = (\sigma^2_{\theta Mk}, \sigma^2_{\theta Uk}, \sigma^2_{\tau Mk}, \sigma^2_{\tau Uk}, k = 1, \ldots, K)$ be collections of hyperparameters. For block s, the full conditional distribution of $(n_{m_s}, \mathbf{I}_s)$ is

$$\begin{aligned}\Pr(n_{m_s}, \mathbf{I}_s | \gamma_s, \{p_{sMk}, p_{sUk}, k = 1, \ldots, K\}, p_s, \alpha_s, \beta_s, \mu, \sigma^2) \quad \propto \\ \Pr(n_{m_s}|p_s)\Pr(\mathbf{I}_s|n_{m_s})\Pr(\gamma_s|\mathbf{I}_s, \{p_{sMk}, p_{sUk}, k = 1, \ldots, K\}),\end{aligned} \tag{11}$$

which is non-zero if and only if the one-to-one and match class size restrictions of section 2.1 are fulfilled. The distributions listed in (11) are discrete.

One way to implement a Gibbs sampling step to draw a new value is $(n^*_{m_s}, \mathbf{I}^*_s)$ as follows. Compute the right-hand side of (11) for all possibilities $(n_{m_s}, \mathbf{I}_s)$, ignoring the constant of proportionality. Then add all the terms together, and divide them by their total to normalize them to sum to one. Order the possibilities $(n_{m_s}, \mathbf{I}_s)$ for block s and compute the cumulative probability distribution for a given order by accumulating the normalized likelihood values. Draw a random deviate $u \sim \text{Uniform}(0,1)$. Let $(n^*_{m_s}, \mathbf{I}^*_s)$ be the match/nonmatch configuration with cumulative probability for the given order closest to but not less than u. The problem with this implementation is the enormous computation needed to compute the (unnormalized) density (11) for all possibilities of $(n_{m_s}, \mathbf{I}_s)$. For example, when $n_{a_s} = n_{b_s} = 10$, there are 234,662,231 possibilities.

One Metropolis-Hastings sampling procedure would randomly pick a new candidate configuration $(n^*_{m_s}, \mathbf{I}^*_s)$ from the product of the prior distributions given in (8) and (9) and the current value of p_s: $\Pr(n_{m_s}|p_s)\Pr(I_s|n_{m_s})$. The move to the candidate configuration would be accepted with the probability

$$\min\Big(\quad 1, \quad \frac{\Pr(n^*_{m_s}, \mathbf{I}^*_s | \text{all current parameter values})\ \Pr(n_{m_s}, \mathbf{I}_s|p_s)}{\Pr(n_{m_s}, \mathbf{I}_s | \text{all current parameter values})\ \Pr(n^*_{m_s}, \mathbf{I}^*_s|p_s)}\ \Big)\,. \tag{12}$$

The normalizing constants from the ratio of the conditional density (11) evaluated at the two sample points cancel. Although it would be simple to sample from the prior distribution and to evaluate the ratio in (12), the procedure is not recommended due to the fact that it would rarely move from a reasonable configuration to a randomly selected new configuration. In practice, each record in a file has few potential matches in the other file. Thus, sampling candidate values from the prior distribution will be very inefficient.

Here we propose incremental ways of modifying n_{m_s} and I_s to cover the space of possible configurations and to produce higher probabilities of change across iterations. Three basic "moves" or modifications of n_{m_s} and I_s will be considered. First, one matching pair can be turned into a nonmatching pair: $n^*_{m_s} = n_{m_s} - 1$ and $I(a,b)$ changes from one to zero for some (a,b) in block s. Second, one nonmatching pair is grouped with the matches: $n^*_{m_s} = n_{m_s} + 1$ and $I(a,b)$ changes from zero to one for some (a,b) such that, before changing the indicator to one, $\sum_{a \in A_s} I(a,b) = 0$ and $\sum_{b \in B_s} I(a,b) = 0$. Third, $n^*_{m_s} = n_{m_s}$ is unchanged, but I^*_s is different from I_s. The changes in I_s that will be considered will involve at most two records from A_s and two from B_s. That is, a record in file A block s that has a match will switch to a new match in file B block s, a record with a match in file B will switch to a new match in file A, or two pairs of records in block s will switch matches.

B.1 Move 1: $n^*_{m_s} = n_{m_s} - 1$

In this movement, one pair currently designated to be a match is changed to a nonmatch designation. Option 1 below chooses a matched pair from block s with uniform probability. Option 2 below chooses a matched pair based on the probability that the pair is a nonmatch given that one among the matches is a nonmatch.

Option 1: Pick a pair $(a, b), a \in A_s, b \in B_s$ that is designated to be a match, $I(a, b) = 1$, at random from the set of all matches in block s with equal probabilities. Suppose the chosen pair is (a_i, b_j), for some index values i and j. For the selected pair, set $I(a_i, b_j) = 0$. Probability of picking a pair (a_i, b_j) is $(n_{m_s})^{-1}$.

The inverse move is to add the deleted pair of records to the set of designated matches (see Move 2 below). If a uniform selection probability is used to add a nonmatching pair to the set of matches, the probability of selecting the dropped match is $((n_{a_s} - n_{m_s})(n_{b_s} - n_{m_s}))^{-1}$. The acceptance probability for the MH algorithm is

$$\min\left(\quad 1, \quad \frac{\Pr(n^*_{m_s}, \mathbf{I}^*_s | \text{all current parameter values})\ n_{m_s}}{\Pr(n_{m_s}, \mathbf{I}_s | \text{all current parameter values})\ (n_{a_s} - n_{m_s})(n_{b_s} - n_{m_s})}\ \right).$$

Option 2: Pick a matched pair (a_i, b_j) at random with probabilities given below and set $I(a_i, b_j) = 0$. In the formulas below, products and summations are over pairs that are designated matches in block s. The pair $(a_i, b_j), a_i \in A_s, b_j \in B_s$ has comparison vector γ_{ij}.

$$\begin{aligned}
\Pr(\text{drop pair}(a_i, b_j)) &= \Pr(\gamma_{ij}|U, s) \prod_{(k,l)\neq(i,j)} \Pr(\gamma_{kl}, (k, l) \neq (i, j)|M, s)/ \\
&\qquad \left(\sum_{i'} \sum_{j'} \Pr(\gamma_{i'j'}|U, s) \prod_{(k,l)\neq(i',j')} \Pr(\gamma_{kl}, (k, l) \neq (i', j')|M, s) \right) \\
&= \Pr(\gamma_{ij}|U, s)/ \\
&\qquad \left(\Pr(\gamma_{ij}|U, s) + \Pr(\gamma_{ij}|M, s) \sum_{i'} \sum_{j'} \Pr(\gamma_{i'j'}|U, s)/\Pr(\gamma_{i'j'}|M, s) \right)
\end{aligned}$$

Given that n_{m_s} in some blocks might not be too large, the computation of probabilities above in some applications might be reasonable. Pairs of records that agree on all or almost all comparisons and that have low levels of agreement with other potential matches likely would not be selected to be dropped. Pairs of records that have more disagreements and that have alternative matches should be dropped more readily.

As for option 1, the inverse move is to add the deleted pair of records to the set of designated matches (see Move 2 below). Let $\Pr(\text{drop pair}(a_i, b_j))$ be the probability of dropping pair (a_i, b_j) from the match set to decrease n_{m_s} by one. Let $\Pr(\text{add pair}(a_i, b_j))$ be the probability of adding pair (a_i, b_j) to the match set to increase n_{m_s} by one. The acceptance probability for the MH algorithm is

$$\min\left(\quad 1, \quad \frac{\Pr(n^*_{m_s}, \mathbf{I}^*_s | \text{all current parameter values})\ \Pr(\text{add pair}(a_i, b_j))}{\Pr(n_{m_s}, \mathbf{I}_s | \text{all current parameter values})\ \Pr(\text{drop pair}(a_i, b_j))}\ \right).$$

B.2 Move 2: $n^*_{m_s} = n_{m_s} + 1$

In this movement, one pair currently designated to be a nonmatch is changed to a match designation. Option 1 below chooses a nonmatched pair from block s with uniform probability. Option 2 below chooses a nonmatched pair based on the probability that the pair is a match given that one among the nonmatches is a match.

Option 1: Pick a pair $(a, b), a \in A_s, b \in B_s$ that is designated to be a nonmatch, $I(a, b) = 0$, at random from the set of all nonmatches in block s with equal probabilities. For the selected pair, set $I(a_i, b_j) = 1$. Probability of picking a pair (a_i, b_j) is $((n_{a_s} - n_{m_s}(n_{b_s} - n_{m_s}))^{-1}$.

The inverse move is to deleted a pair of records from the set of designated matches (see Move 1 below). If a uniform selection probability is used to delete a matching pair, the acceptance probability for the MH algorithm is

$$\min\left(\quad 1, \quad \frac{\Pr(n^*_{m_s}, \mathbf{I}^*_s | \text{all current parameter values})\ (n_{a_s} - n_{m_s})(n_{b_s} - n_{m_s})}{\Pr(n_{m_s}, \mathbf{I}_s | \text{all current parameter values})\ n_{m_s}}\ \right).$$

Option 2: Pick a nonmatched pair (a_i, b_j) at random with probabilities given below and set $I(a_i, b_j) = 1$. In the formulas below, products and summations are over pairs that are designated nonmatches in block s.

$$
\begin{aligned}
\Pr(\text{add pair}(a_i, b_j)) &= \Pr(\gamma_{ij}|M,s) \prod_{(k,l)\neq(i,j)} \Pr(\gamma_{kl}, (k,l)\neq(i,j)|U,s)/ \\
&\qquad \left(\sum_{i'} \sum_{j'} \Pr(\gamma_{i'j'}|M,s) \prod_{(k,l)\neq(i',j')} \Pr(\gamma_{kl}, (k,l)\neq(i',j')|U,s) \right) \\
&= \Pr(\gamma_{ij}|M,s)/ \\
&\qquad \left(\Pr(\gamma_{ij}|M,s) + \Pr(\gamma_{ij}|U,s) \sum_{i'} \sum_{j'} \Pr(\gamma_{i'j'}|M,s)/\Pr(\gamma_{i'j'}|U,s) \right)
\end{aligned}
$$

Pairs of records that disagree on all or almost all comparisons are not likely to be added. Pairs of records that are current nonmatches but agree on many fields are likely to be added. As for option 1, the inverse move is to delete the added pair of records from the set of designated nonmatches (see Move 1 above). The acceptance probability for the MH algorithm is

$$
\min\left(\quad 1, \quad \frac{\Pr(n^*_{m_s}, \mathbf{I}^*_s|\text{all current parameter values})\ \Pr(\text{drop pair}(a_i, b_j))}{\Pr(n_{m_s}, \mathbf{I}_s|\text{all current parameter values})\ \Pr(\text{add pair}(a_i, b_j))}\ \right).
$$

B.3 Move 3: n_{m_s} unchanged but I_s altered

In this movement, three things can happen: two matches can switch pairs, a matched pair can replace one of its units with an unmatched pair, or a matched pair can be dropped and replaced with another matched pair.

Variation 1: Two matches switch pairings: Randomly select two matched pairs, (a_i, b_j) and (a_k, b_l), with probability $2/(n_{m_s}(n_{m_s}-1))$ and switch the pairings: (a_i, b_l) and (a_k, b_j). That is, change $I(a_i, b_j)$ and $I(a_k, b_l)$ from one to zero and $I(a_i, b_l)$ and $I(a_k, b_j)$ from zero to one. The reverse move is to undo the switch. The acceptance probability of the MH algorithm is the minimum of one and

$$
(P(\gamma_{il}|M,s)P(\gamma_{kj}|M,s)P(\gamma_{ij}|U,s)P(\gamma_{kl}|U,s)) \,/\, (P(\gamma_{ij}|M,s)P(\gamma_{kl}|M,s)P(\gamma_{il}|U,s)P(\gamma_{kj}|U,s)).
$$

It would be possible to select two matched pairs with non uniform probabilities, but doing so could be computationally expensive. Given that two pairs need to be switched, the probability that pairs (a_i, b_j) and (a_k, b_l) are to be switched is $P(\gamma_{il}|M,s)P(\gamma_{kj}|M,s)P(\gamma_{ij}|U,s)P(\gamma_{kl}|U,s)$ divided by the sum of products of this sort over $n_{m_s}(n_{m_s}-1)/2-1$ sets of two pairs (all sets of two pairs except (i,j) and (k,l)). If n_{m_s} is large, this could be a large number of computations. The MH acceptance probability is

$$
\min\left(\quad 1, \quad \frac{\sum\sum_{(i',j')\neq(i,j),(k',l')\neq(k,l)} P(\gamma_{i'l'}|M,s)P(\gamma_{k'j'}|M,s)P(\gamma_{i'j'}|U,s)P(\gamma_{k'l'}|U,s)}{\sum\sum_{(i',l')\neq(i,l),(k',j')\neq(k,j)} P(\gamma_{i'j'}|M,s)P(\gamma_{k'l'}|M,s)P(\gamma_{i'l'}|U,s)P(\gamma_{k'j'}|U,s)}\ \right).
$$

A less computationally intense approach would randomly choose one matched pair, say (a_i, b_j), with uniform probability $(1/n_{m_s})$ and a second matched pair with non-uniform probability. Given that pair (a_i, b_j) is going to be broken and switched with another pair from the current matches, one could select the pair (a_k, b_l) with probability

$$
\frac{P(\gamma_{il}|M,s)P(\gamma_{kj}|M,s)P(\gamma_{ij}|U,s)P(\gamma_{kl}|U,s)}{\sum_{(k',l')\neq(i,j)} P(\gamma_{il'}|M,s)P(\gamma_{k'j}|M,s)P(\gamma_{ij}|U,s)P(\gamma_{k'l'}|U,s)}.
$$

If a similar reverse move is considered, then the MH acceptance probability is

$$
\min\left(\quad 1, \quad \frac{\sum_{(k',l')\neq(i,j)} P(\gamma_{il'}|M,s)P(\gamma_{k'j}|M,s)P(\gamma_{ij}|U,s)P(\gamma_{k'l'}|U,s)}{\sum_{(i'l')\neq(k,j)} P(\gamma_{i'j}|M,s)P(\gamma_{kl'}|M,s)P(\gamma_{i'l'}|U,s)P(\gamma_{kj}|U,s)}\ \right).
$$

Variation 2: A matched pair replaces one of its matching records with a nonmatching record: In this move, a matched pair of records is randomly chosen and one of its component records is replaced with a record from the same file in the same block that does not have a designated match. That is, suppose $I(a_i, b_j) = 1$ and the matched pair (a_i, b_j) is chosen. One of the matched pairs can be chosen with uniform probability: $1/n_{m_s}$. A records a_k in file A without a match satisfies $\sum_{j'} I(a_k, b_{j'}) = 0$. A record b_l in file B without a match satisfies $\sum_{i'} I(a_{i'}, b_l) = 0$. There are $n_{a_s} + n_{b_s} - 2n_{m_s}$ nonmatched records in block s. One option is to choose a nonmatched record randomly. The

reverse move would involve switching to the initial pairings. If the A-record a_i is replaced through random selection with A-record a_k, the MH acceptance probability is the minimum of one and

$$P(\gamma_{kj}|M,s)P(\gamma_{ij}|U,s)/\left(P(\gamma_{ij}|M,s)P(\gamma_{kj}|U,s)\right).$$

If the B-record b_j is replaced through random selection with B-record b_l, the MH acceptance probability is the minimum of one and

$$P(\gamma_{il}|M,s)P(\gamma_{ij}|U,s)/\left(P(\gamma_{ij}|M,s)P(\gamma_{il}|U,s)\right).$$

Another way to choose the replacement record is to compute the probability given current parameter values that a particular nonmatching record is a match, assuming that pair (a_i, b_j) is a nonmatching pair. The probability the matching pair is (a_k, b_j), where record a_k currently does not have a match, is

$$P(\gamma_{kj}|M,s)/\left(\sum_{k'} P(\gamma_{k'j}|M,s) + \sum_{l'} P(\gamma_{il'}|M,s)\right),$$

where the sums are over A-records without a match (k') and B-records without a match (l'). The MH algorithm acceptance probability is the minimum of one and

$$\frac{\sum_{k'} P(\gamma_{k'j}|M,s) + \sum_{l'} P\gamma_{il'}|M,s)}{\sum_{i'} P(\gamma_{i'j}|M,s) + \sum_{j'} P(\gamma_{kj'}|M,s)}.$$

The summations in the denominator are over records that would not have matches if (a_k, b_j) were a match.

The probability that the matching pair is (a_i, b_l), where record b_l currently does not have a match, is

$$P(\gamma_{il}|M,s)/\left(\sum_{l'} P(\gamma_{il'}|M,s) + \sum_{k'} P(\gamma_{k'j}|M,s)\right),$$

where the sums are over A-records without a match (k') and B-records without a match (l'). The MH algorithm acceptance probability is the minimum of one and

$$\frac{\sum_{l'} P(\gamma_{il'}|M,s) + \sum_{k'} P\gamma_{k'j}|M,s)}{\sum_{i'} P(\gamma_{i'l}|M,s) + \sum_{j'} P\gamma_{ij'}|M,s)}.$$

The summations in the denominator are over records that would not have matches if (a_i, b_l) were a match.

Variation 3: A matched pair is deleted and two unmatched records are paired: The last move that will be contemplated is the deletion of a matched pair and the joining of two unmatched records. If (a_i, b_j) is a match and a_k and b_l are unmatched records, the move entails setting $I(a_i, b_j) = 0$ and $I(a_k, b_l) = 1$. This is in effect almost the combination of the first two moves: removal of a match and addition of a new match other than the one that was removed. The match to be removed could be chosen using uniform probabilities or with probabilities as described for Move 1. The nonmatching pairs to be linked together can be chosen using uniform probabilities or with probabilities similar to those described in Move 2. An acceptance probability for the MH algorithm can be computed as the product of appropriately modified probabilities associated with Moves 1 and 2.

The choice of how often to consider various moves: If there is good matching information in most blocks and drawn parameter values are in ranges that are appropriate for identifying matches, then it is possible that accepted changes in $\mathbf{I}$ and $n_{m_s}, s = 1, \ldots, S$ will oscillate among likely match/nonmatch configurations with an occasional unlikely configuration now and then. In any particular block, members of the set of likely configuations could differ from each other in several ways. The Metropolis-Hastings/Gibbs sampling algorithm should explore the posterior distribution of parameters and match/nonmatch configurations if moves that correspond to differences between likely configuations in a block are used more often in that block. Thus, an adaptive procedure can be recommended. In block s, start the iterative simulation and try each type of move each iteration for the first several iterations. After a specified number of iterations, increase the frequency with which moves that result in accepted changes and decrease the frequency with which moves that do not result in accepted changes are attempted. An alternative that would be simpler to implement would be to attempt one move, randomly selected from the possible types of moves, each iteration of the algorithm. That is, it is not necessary to try each type of move every iteration. Practical experience and monitoring of the algorithm in a particular application will be necessary before any further recommendations can be made.