Scalable Bayesian Record Linkage

Brendan S. McVeigh
Carnegie Mellon University
bmcveigh@stat.cmu.edu

Jared S. Murray
University of Texas at Austin
jared.murray@mccombs.utexas.edu

Abstract

Probabilistic record linkage (PRL) is the process of determining which records in two databases correspond to the same underlying entity without a unique identifier. Bayesian methods provide a powerful mechanism for characterizing uncertainty in links between records (via the posterior distribution). The computational complexity of existing Bayesian approaches to PRL severely limits the size of problems to which these techniques can be successfully applied. We propose a new computationally efficient and scalable approach, providing both a fast algorithm for generating a point estimate of the linkage structure that properly accounts for one-to-one matching and a restricted MCMC algorithm that samples from an approximate posterior distribution. These advances make it possible to perform Bayesian PRL for larger problems, and to perform sensitivity analysis with respect to different prior specifications. We demonstrate the methods on simulated and real data.

1 Introduction

When two or more databases need to be merged without a unique identifier, record linkage must be estimated based on only incomplete or imperfect information and is therefore uncertain (probabilistic). Bayesian methods provide a straightforward approach to account for uncertainty in the link structure via the posterior distribution.

Early applications of probabilistic record linkage (PRL) include linking files from surveys and censuses to estimate the number of records in common and estimate the total population size in capture-recapture studies (Neter et al., 1965; Fellegi and Sunter, 1969; Winkler and Thibaudeau, 1991). Similar methods have been applied for estimating casualty counts in conflict regions (Steorts et al., 2016; Ventura and Nugent, 2014; Sadinle et al., 2014; Sadinle, 2017). PRL is also used to gather more complete about individuals, such as when covariates and a response are recorded on different files (Judson, 2007; Gutman et al., 2013; Gu and Gutman, 2016; Dalzell and Reiter, 2016). Here we consider the common case of merging two de-duplicated files, so that each record in one file matches at most one record in the other ("one-to-one" matching).

2 A Model for Record Linkage

Consider two sets of records, denoted A and B, containing n_A and n_B records. Records $a \in A$ and $b \in B$ are said to be "matched" or "linked" if they refer to the same entity, which we denote $a \sim b$. In PRL we would like to infer the latent links, which can be conveniently represented in matrix form:

$$C_{ab} = \begin{cases} 1 & a \sim b \\ 0 & a \nsim b. \end{cases} \tag{1}$$

For each record we observe a set of features, such as names, addresses, and demographic information. We estimate the unobserved link structure C using *comparisons* between the features observed for pairs of records. Examples of comparisons between records include the absolute differences in two

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

ages, or string similarities between two names; see e.g. Christen (2012) for a detailed account of generating comparisons. For each record pair (a, b) we observe a vector of d comparisons:

$$\gamma_{ab} = (\gamma_{ab}^1, \gamma_{ab}^2, \ldots, \gamma_{ab}^d), \tag{2}$$

where each entry γ_{ab}^j is an ordinal level of agreement between fields. We model these comparison vectors as arising from a two component mixture, with observed comparison vectors generated by either true matching or non-matching record pairs:

$$m(g) = \Pr\left(\gamma_{ab} = g \mid C_{ab} = 1\right); \qquad u(g) = \Pr\left(\gamma_{ab} = g \mid C_{ab} = 0\right). \tag{3}$$

To reduce the number of parameters in the model we make the common assumption of conditional independence between comparisons. Define:

$$m_{jh} = \Pr\left(\gamma_{ab}^j = h | C_{ab} = 1\right); \qquad u_{jh} = \Pr\left(\gamma_{ab}^j = h | C_{ab} = 0\right), \tag{4}$$

for $1 \leq j \leq d$ and $1 \leq h \leq k_j$, where comparison j has k_j possible levels. Under conditional independence we have

$$m(\gamma_{ab}) = \prod_{j=1}^{d} \prod_{h=1}^{k_i} m_{jh}^{\mathbb{1}(\gamma_{ab}^j = h)}; \qquad u(\gamma_{ab}) = \prod_{j=1}^{d} \prod_{h=1}^{k_i} u_{jh}^{\mathbb{1}(\gamma_{ab}^j = h)}. \tag{5}$$

In our examples here the prior over C is $p(C) \propto \exp(-\theta L)$ where $L = \sum_a \sum_b C_{ab}$. (Green and Mardia, 2006). Finally, we assume that

$$(m_{j1}, \ldots, m_{jk_j}) \sim Dir(\alpha_{mj1}, \ldots, \alpha_{mjk_j}); \;\; (u_{j1}, \ldots, u_{jk_j}) \sim Dir(\alpha_{uj1}, \ldots, \alpha_{ujk_j}) \tag{6}$$

independently. We use these models and prior distributions here for exhibition; our restricted MCMC algorithm is not specific to a particular model or prior distribution.

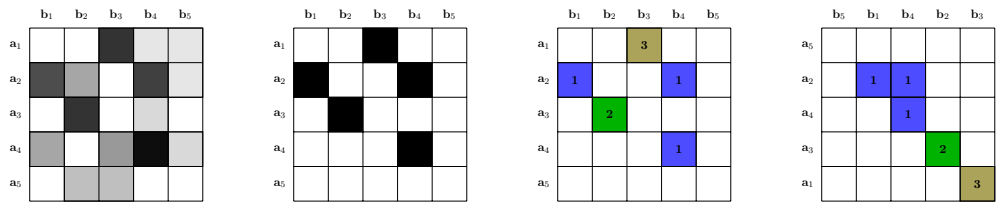
3 Post-hoc Blocking for Posterior Sampling

In large-scale record linkage problems it is necessary to reduce the number of candidate record pairs, regardless of the approach to PRL that will be employed. Typically a *blocking* scheme is used during pre-processing; for example, we might only consider record pairs from the same geographical area as possible matches. These areas form *blocks* of records such that links are only allowed between records in the same block. Due to their computational complexity, Bayesian approaches typically require stricter blocking than competing methods. Stricter blocking schemes will tend to have larger false-negative rates (for example, due to excluding matches when people move outside the original geographical area, or when the blocking variable is recorded with error).

We propose a data-driven approach to defining a fine-scale, accurate blocking scheme, which we call *post-hoc blocking*. (In large-scale problems, post-hoc blocking will necessarily follow an application of "traditional" blocking.) In post-hoc blocking we use rough estimates of the m and u parameters to identify record pairs that likely have a non-trivial posterior match probability – i.e., pairs with estimated weights $\hat{w}_{ab} = \log[\hat{m}(\gamma_{ab})/\hat{u}(\gamma_{ab})]$ above a conservative threshold w_0. In Appendix A we outline how to compute a MAP estimate of these parameters under the model above.

Algorithm 1 summarizes our approach. We use the estimated weights to define a bipartite graph G, where the nodes are records and an edge exists between a and b if $\hat{w}_{ab} > w_0$. Then we find the connected components of G, which can be done efficiently (in linear time (Tarjan, 1972; Gazit, 1986)). All the links must occur within these connected components, which we call *post-hoc blocks*. Figures 1 demonstrates post-hoc blocking in a small example.

With the post-hoc blocks in hand, we can implement an efficient restricted MCMC algorithm that only proposes linking records a and b if they lie in the same post-hoc block (and therefore have $\hat{w}_{ab} > w_0$). In addition to reducing the number of record pairs under consideration, post-hoc blocking also makes the MCMC embarrassingly parallel. Conditional on the model parameters, the blocks of C corresponding to the post-hoc blocks are conditionally independent and can be updated in parallel. In our examples C is updated using Metropolis-Hastings add/drop/swap link updates(see e.g. Green and Mardia (2006)). If the blocks are small we have the option of jointly sampling all links within a block by enumerating all possibilities and doing a simple Gibbs update, providing further computational gains. (Here we use Metropolis steps throughout.)



(a) (b) (c) (d)

Figure 1: An example of post-hoc blocking (a) is a heatmap of estimated weights, with dark cells corresponding to larger weights. (b) The matrix of weights converted to an binary matrix by thresholding the weights at w_0. This defines the adjacency matrix of a bipartite graph G. In (c) we identify the connected components of this graph, which is reordered in (d). Each connected component defines a post-hoc block, and our restricted MCMC will only consider linking records in the same block.

Algorithm 1 Post-hoc Blocking

Input: Comparison vectors Γ, weight threshold w_0

Output: Penalized likeihood estimates of m, u, and C. An $n_A \times n_B$ adjacency matrix G, and its connected components defining post-hoc blocks.

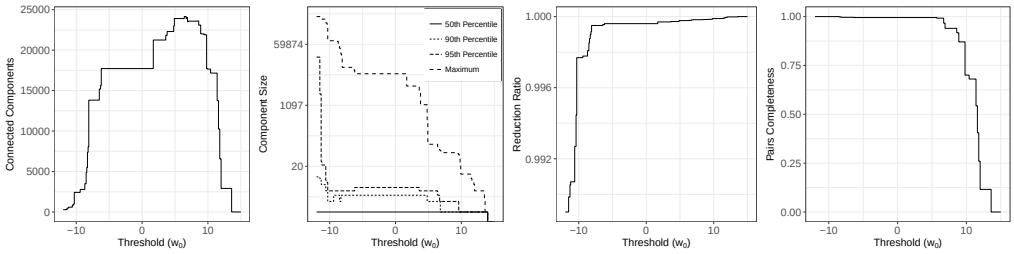
1. Generate $\hat{m}$, $\hat{u}$ and $\hat{C}$ via penalized maximum likelihood (Appendix A)
2. Set $G_{ab} = \mathbb{1}(\log[\hat{m}(\gamma_{ab})/\hat{u}(\gamma_{ab})] > w_0)$
3. Find the connected components of the bipartite graph with adjacency matrix G; these are the *post-hoc blocks*

4 Experiments

Italian Census Data - Bayesian estimation and sensitivity analysis In Appendix C we use a small-scale example to show that 1) the restricted MCMC algorithm gives very similar results to a full MCMC algorithm, 2) that posterior inferences about linkage structure can be highly sensitive to prior specification, and 3) the advantages of estimating a full posterior over a simple point estimate of the link structure. The second point suggests that we should routinely assess the sensitivity of our results to different prior distributions, which is prohibitively difficult computationally without post-hoc blocking and restricted MCMC.

Large-scale synthetic example We use synthetic data provided by Sadinle (2017) to illustrate the large-scale behavior of post-hoc blocking and restricted MCMC. The data contains 100 pairs of datasets to be linked, with each individual data set containing 500 records. We stacked the original synthetic records to create two datasets of 50,000 records ($n_A = 50,000$, $n_B = 50,00$). Each record contains four fields: given name, family name, age and occupation categories. Errors are introduced into 2 of the 4 fields for each record and the share of records which are linked is 50%. As in Sadinle (2017), given name and family name are compared based on discretized Levenstein dissimilarity, with levels (0.5, 1.0], (0.25, 0.50], (0, 0.25], and 0 (ranging from strong dissimilarity to exact agreement). Age and occupation were compared using exact agreement. We take $m_j \sim Dir(1, 2, 5, 10)$ for the discretized string comparisons and $m_j \sim Beta(4, 1)$ for exact agreement comparisons. For all comparisons, $p(u_j) \propto 1$.

It is infeasible to compute comparison vectors for 2.5 billion record pairs, so as an initial blocking step we exclude any record pairs that have different postal codes. The 29 resulting blocks contain about 160 million record pairs in total and exclude no truly matching pairs (because no errors are introduced into the postal codes). For MCMC sampling, we implemented post-hoc blocking with $w_0 = 4.9$. This threshold was selected to limit the maximum size of the post-hoc blocks to no more than 50 records from either file, yielding 23,903 total blocks. (In practice we suggest choosing the largest value for w_0 that yields a computationally feasible problem.) Finding the post-hoc blocks – computing penalized likelihood estimates and computing the connected components – took 312 seconds to run on a laptop with a 2.60 GHz processor.



(a) (b) (c) (d)

Figure 2: Post-hoc blocks resulting from varying threshold value w_0. (a) Number of blocks. (b) Block size quantiles. (c) Reduction ratio (proportion of record pairs excluded by post-hoc blocking). (d) Pairs completeness (proportion of truly matching record pairs retained by post-hoc blocking).

After post-hoc blocking the number of candidate pairs falls to 36,356 from over 160 million, for a reduction ratio (proportion of record pairs excluded by post-hoc blocking) of about 99.98%. Since we have ground truth here we can assess the effect of post-hoc blocking on false non-match rates: Of the 25,000 truly matching record pairs, 24,877 are retained in the same post-hoc block for a pairs completeness (proportion of truly matching record pairs retained by post-hoc blocking) of $24,877/25,000 \approx 99.5\%$. A majority of the post-hoc blocks (19,901) contain only a single possible record pair while the largest contains 32 and 31 records from file A and B, respectively. The effect of w_0 on the size of the post-hoc blocks, the reduction ratio, and the pairs completeness metric are shown in Figure 2. There is a range of values for w_0 (roughly ± 5) that yields a reasonable tradeoff between computational gains and bias due to false non-matches.

Finally we ran the restricted MCMC algorithm, performing 2,500 iterations (where each iteration performed an add/delete/swap move within each of the $23,903$ post-hoc blocks) after 1,000 iterations of burn-in. The resulting Bayes estimate has recall of 95.5% and precision of 94.6%. We also attempted to run a MCMC algorithm using only the traditional blocking step. After running that chain over 200 times longer than the restricted MCMC, the unrestricted chain had clearly not converged. To our knowledge this is the largest PRL example where MCMC has been successfully applied; existing algorithms are simply too inefficient to conceivably mix over an unrestricted space for C.

5 Discussion

Post-hoc blocking and restricted MCMC represents a significant advance in approximate Bayesian inference for PRL. Much existing work has been done on other models for Bayesian record linkage, including those that model record features directly rather than via comparisons (see e.g. Gutman et al. (2013); Fortini et al. (2002); Tancredi et al. (2011); Steorts et al. (2015, 2016)). Post-hoc blocking is readily applied to any record linkage model; the weights from the penalized likelihood procedure in Appendix A will generally be a reasonable filter, even if the Bayesian model of interest does not match the penalized likelihood. Generalizing post-hoc blocking and restricted MCMC to settings with multiple files and duplicate records is conceptually straightforward but poses new modeling and computational challenges; this is an important area for future work.

References

Christen, P. (2012). *Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection*. Springer Science & Business Media.

Dalzell, N. M. and Reiter, J. P. (2016). Regression modeling and file matching using possibly erroneous matching variables. *arXiv preprint arXiv:1608.06309*.

Fellegi, I. P. and Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64(328):1183–1210.

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

Gazit, H. (1986). An optimal randomized parallel algorithm for finding connected components in a graph. In *Foundations of Computer Science, 1986., 27th Annual Symposium on*, pages 492–501. IEEE.

Green, P. J. (2015). Mad-bayes matching and alignment for labelled and unlabelled configurations. *Geometry Driven Statistics*, 121:377.

Green, P. J. and Mardia, K. V. (2006). Bayesian alignment using hierarchical models, with applications in protein bioinformatics. *Biometrika*, 93(2):235–254.

Gu, C. and Gutman, R. (2016). Combining item response theory with multiple imputation to equate health assessment questionnaires. *Biometrics*.

Gutman, R., Afendulis, C. C., and Zaslavsky, A. M. (2013). A bayesian procedure for file linking to analyze end-of-life medical costs. *Journal of the American Statistical Association*, 108(501):34–47.

Jonker, R. and Volgenant, T. (1986). Improving the hungarian assignment algorithm. *Operations Research Letters*, 5(4):171–175.

Judson, D. (2007). Information integration for constructing social statistics: history, theory and ideas towards a research programme. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(2):483–501.

Kuhn, H. W. (1955). The hungarian method for the assignment problem. *Naval Research Logistics (NRL)*, 2(1-2):83–97.

Lawler, E. L. (1976). *Combinatorial optimization: networks and matroids*. Courier Corporation.

Neter, J., Maynes, E. S., and Ramanathan, R. (1965). The effect of mismatching on the measurement of response errors. *Journal of the American Statistical Association*, 60(312):1005–1027.

Sadinle, M. (2017). Bayesian estimation of bipartite matchings for record linkage. *Journal of the American Statistical Association*, 112(518):600–612.

Sadinle, M. et al. (2014). Detecting duplicates in a homicide registry using a bayesian partitioning approach. *The Annals of Applied Statistics*, 8(4):2404–2434.

Steorts, R. C. et al. (2015). Entity resolution with empirically motivated priors. *Bayesian Analysis*, 10(4):849–875.

Steorts, R. C., Hall, R., and Fienberg, S. E. (2016). A bayesian approach to graphical record linkage and deduplication. *Journal of the American Statistical Association*, 111(516):1660–1672.

Tancredi, A., Liseo, B., et al. (2011). A hierarchical bayesian approach to record linkage and population size problems. *The Annals of Applied Statistics*, 5(2B):1553–1585.

Tarjan, R. (1972). Depth-first search and linear graph algorithms. *SIAM journal on computing*, 1(2):146–160.

Ventura, S. L. and Nugent, R. (2014). Hierarchical linkage clustering with distributions of distances for large-scale record linkage. In *International Conference on Privacy in Statistical Databases*, pages 283–298. Springer.

Winkler, W. E. and Thibaudeau, Y. (1991). An application of the fellegi-sunter model of record linkage to the 1990 us decennial census. *US Bureau of the Census*, pages 1–22.

A Penalized Likelihood Estimation

The unnormalized log-posterior for the model in Section 2 is

$$\ell(C, m, u|\Gamma) = \sum_{a,b \in A \times B} [C_{ab} \log(m(\gamma_{ab})) + (1 - C_{ab}) \log(u(\gamma_{ab}))] - \theta \sum_{ab} C_{ab} \tag{7}$$

Here θ is the penalty parameter representing the cost of each additional link. The penalty term is better understood by rearranging (7) as

$$\sum_{a,b \in A \times B} \log(u(\gamma_{ab})) + C_{ab}(w_{ab} - \theta) \tag{8}$$

where the weight

$$w_{ab} = \log\left(\frac{m(\gamma_{ab})}{u(\gamma_{ab})}\right) \tag{9}$$

summarizes information about the *relative* likelihood of a record pair being a link versus non-link.

Here θ represents the cost of each additional link, so only pairs with $w_{ab} > \theta$ can be linked without decreasing the log-likelihood. In our penalized likelihood approach this constraint is enforced during the initial estimation.

A local mode of (8) is readily obtained by alternating maximization steps: (1) Holding m, u constant, maximizing with respect to C and then maximizing with respect to m, u holding C constant. Maximization of (8) with respect to C equivalent to solving the following optimization problem:

$$\begin{aligned} \max_{C} \quad & \sum_{a,b \in A \times B} C_{ab}(w_{ab} - \theta) \\ \text{subject to} \quad & C_{ab} \in \{0, 1\} \\ & \sum_{b \in B} C_{ab} \leq 1 \quad \forall a \in A \\ & \sum_{a \in A} C_{ab} \leq 1 \quad \forall b \in B \end{aligned} \tag{10}$$

The solution can be found by solving an LSAP and then simply deleting links where $w_{ab} \leq \theta$, details on the LSAP are given in Appendix B. Efficient algorithms exist for solving LSAPs, (e.g. the Hungarian algorithm (Kuhn, 1955)) and have a worst case complexity of $O(n^3)$ where $n = \max(n_A, n_B)$ (Jonker and Volgenant, 1986; Lawler, 1976). (Green (2015) proposes a similar penalized likelihood approach for alignment problems under a different class of models.)

Holding C constant, (7) separates into distinct factors for each m_{jh} and u_{jh}, which are maximized by setting

$$m_{jh} = \frac{n_{mjh} + \sum_{ab} C_{ab} \mathbb{1}(\gamma^j_{ab} = h)}{\sum_h n_{mjh} + \sum_{ab} C_{ab}} \qquad u_{jh} = \frac{n_{ujh} + \sum_{ab} (1 - C_{ab}) \mathbb{1}(\gamma^j_{ab} = h)}{\sum_h n_{ujh} + \sum_{ab} (1 - C_{ab})}. \tag{11}$$

where the n's are optional pseudocounts used to regularize the estimates (omitted from (7)). In our implementation the n's are simply the hyperparameters of the prior distributions of the m and u parameters.

B LSAP

The optimization problem in (10) of Appendix A is solved using a LSAP based on modified weights

$$\tilde{w}_{ab} = \begin{cases} w_{ab} - \theta & w_{ab} - \theta > 0 \\ 0 & w_{ab} \leq \theta \end{cases} \tag{12}$$

the result of applying soft-thresholding to the weights defined in (9). We then solve the optimization problem:

$$\begin{aligned} \max_{C} \quad & \sum_{a,b \in A \times B} C_{ab} \tilde{w}_{ab} \\ \text{subject to} \quad & C_{ab} \in \{0,1\} \\ & \sum_{b \in B} C_{ab} \leq 1 \quad \forall a \in A \\ & \sum_{a \in A} C_{ab} \leq 1 \quad \forall b \in B \end{aligned} \tag{13}$$

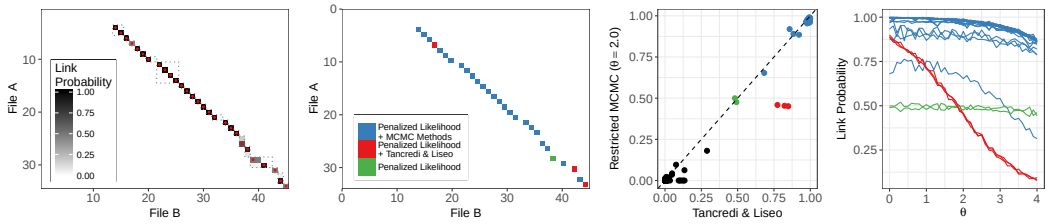
If C^* is the value of C which maximizes (10) then C^* also maximizes (13) as the coefficients of all entries of C_{ab} for which $\tilde{w}_{ab} > 0$ are unaffected by the transformation. However, the same objective value will be achieved by any value of C where $C_{ab} = 1$ for all a, b where $C^*_{ab} = 1$ but there may also exist a, b where $C_{ab} = 1$ and $C^*_{ab} = 0$. It is this equivalence which allow the optimal solution to be found using by solving a LSAP, which assigns all rows to a column. The soft-thresholding ensures that optimal objective value of the LSAP will be identical to that of (10) allowing C^* to be recovered by simply deleting links where $\tilde{w}_{ab} = 0$.

C Small-scale Example: Italian Census

We reanalyze data from the 2001 Italian census and a post-enumeration survey from Tancredi et al. (2011). The data come from a small geographic area; there are 34 records from the census (file A) and 45 records from the post-enumeration survey (file B). Each record includes three categorical variables: the first two consonants of the family name (339 categories), sex (2 categories), and education level (17 categories). We generate comparisons as binary indicators of an exact match between each field. We assume that $m_i \sim Beta(20, 3)$ for $i = 1, 2, 3$ and $u_j \sim Beta(3, 20)$ for $j = 1, 2, 3$.

We first compute estimates of the weights using the penalized likelihood method described in Appendix A. A value of $\theta = 2.0$ in the prior was chosen based on the implied marginal prior over the number of links L. We then employ post-hoc blocking using a threshold of $w_0 = 0.0$ before running our restricted MCMC algorithm. Finally, we run our restricted MCMC algorithm for 10,000 steps with the first 1,000 discarded as burn-in. Trace plots indicate that the restricteed MCMC algorithm convergences quickly. Figure 3a shows the posterior probabilities of a match between each record pair with the links identified by the penalized likelihood approach indicated with red dots. The post-hoc blocks are indicated by the dashed lines; entries outside the blocks are fixed at 0 by the restricted MCMC. Only 46 (3%) of the 1,530 record pairs are contained in a post-hoc block, significantly reducing the scale of the problem. We also note that the largest of the 23 post-hoc blocks contains only 4 records from each file, only slightly more than 1% of all record pairs.

Figure 3b provides a comparison of different point estimates. The 'MCMC Methods' designation indicates agreement between our restricted MCMC and the unrestricted version (run for 200,000 steps), as well as the results of Tancredi et al. (2011). For MCMC methods we use the Bayes estimate



(a) (b) (c) (d)

Figure 3: (a) Restricted MCMC posterior match probabilities, with penalized likelihood estimate shown in red. (b) Comparison of point estimates for C, 'MCMC Methods' includes both a restricted and unrestricted MCMC algorithm as well as an estimate from Tancredi et al. (2011) (c) Posterior link probabilities from the restricted MCMC algorithm compared to those of Tancredi et al. (2011). (d) Prior sensitivity analysis of restricted MCMC posterior link probability estimate.

for C under balanced misclassification loss functions, which is $\hat{C}_{ab} = \mathbb{1}(\Pr(C_{ab} = 1 \mid \Gamma) > 0.5)$ (Tancredi et al., 2011).

The single record pair for which the penalized likelihood estimate disagrees with that of Tancredi et al. (2011) highlights the importance of estimating a full posterior distribution. Here one record in file A has two exact matches in file B, so the posterior match probability for each record pair is just under 0.5 (as can been seen in green in Figure 3c, up to Monte Carlo error). The penalized likelihood estimator calls one of these pairs a match more or less at random, since there are (at least) two equivalent posterior modes.

The posterior link probabilities from the restricted MCMC are similar to those obtained using the significantly more complicated model introduced by Tancredi et al. (2011) (Figure 3c). However, the points shown in red are notable exceptions. Figure 3d shows that as we increase θ in the prior, the marginal probability of linking these records drops precipitously. In contrast, the posterior probability of linking record pairs shown in blue and green generally show less sensitivity to the choice of prior. This kind of sensitivity analysis is only possible with efficient posterior computation; one run of Tancredi et al. (2011)'s MCMC algorithm takes about 20 minutes even in this small problem.