# Bayesian Parametric and Nonparametric Inference for Multiple Record Linkage

**Rob Hall**
Department of Machine Learning and Statistics
Carnegie Mellon University
Pittsburgh, PA 15213
rjhall@cs.cmu.edu

**Rebecca C. Steorts**
Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213
beka@cmu.edu

**Stephen E. Fienberg**
Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213
fienberg@stat.cmu.edu

## Abstract

Record linkage is an historically important statistical problem arising when data about some population of individuals is spread over several files. As kids, we grew up with the game "Where in the world is Carmen San Diego"? Nowadays, the name of the game for the U.S. Census Bureau and other organizations is who's the *real Steve Fienberg*, where they are dealing with deciding if someone named Steve Fienberg is the same person across multiple lists. This problem has been looked at in the past in the two-file setting and more recently for multiple files. We propose Bayesian parametric and nonparametric methodology for multiples files in which the fields are regarded as independent. A linkage structure represents whether or not individuals $i$ and $j$ in particular lists are matches. We estimate the posterior distribution of this linkage structure via a hybrid Markov chain Monte Carlo (MCMC) algorithm. We then compute associated matching probabilities that individuals are captured using the National Long-Term Care Survey (NLTCS) dataset, where the truth is known since patients are tracked by a unique patient identifier.

## 1 Introduction

Record linkage is a historically important statistical problem arising when data about some population of individuals is spread over several files. The goal of record linkage is to determine whether a record from one file corresponds to a record of a second file, in the sense that the two records describe the same individual. Another goal is then to integrate multiple sources of data into a single file, which can later be used for further statistical analysis. The classical method [1] considered the simple setting of two files, and this work was recently generalized to the case of $k > 2$ files [4, 5]. In the original two-file case, nonparametric techniques were provided by [2], while [6] retained a parametric formulation but approached the problem from a Bayesian standpoint. A key benefit of the Bayesian paradigm is its natural ability to correctly handle uncertainty in the linkage structure, which poses a difficult challenge to frequentist record linkage techniques. Moreover, a Bayesian approach allows the utilization of prior information if such information is available or if certain types of linkage structures are believed to naturally be more likely than others.

We also take a Bayesian approach to the record linkage problem. However, unlike [6], we propose both parametric and nonparametric methodology, and our work addresses the more complicated case of more than two data files. Furthermore, we present a hybrid MCMC approach to sample from the posterior distribution of the linkage structure,

which allows us to state the approximate posterior probability of a match between two or more records, as opposed to simply a binary match/non-match decision. We demonstrate our approach using three of the five data files of the NLTCS, which tracked and surveyed approximately 20,000 individuals at five-year intervals. At each wave of the survey, some individuals had died and were replaced by a new cohort, thus the files contain overlapping but different sets of individuals.

## 2 Model Construction and Notation

Suppose we have three files in which each record corresponds to an individual, and where all files have the same set of "matching" variables, or identifying information. We assume that each file has at most one record for each individual, i.e., we assume that duplicates have been removed prior to the record linkage process. We also assume that there is no missing data, apart from the obvious fact that many individuals will be "missing" their entire records from one or more files.

Consider the restriction of each record in each file to the so-called "matching" variables present in all files. Denoting the records by $\boldsymbol{X}$, we assume that the observed records are generated by "distorting" some unobserved variables which characterize the underlying individual. For instance, an individual who is a male born in 1984 may appear in two files, one reporting the correct information and the other specifying the birth year as 1985 due to some kind of transcription error in the data collecting process. Further sources of distortions might be measurement errors when the data contain biometric information like weights, or when the characteristics of the underlying individual change over time (for instance body weight, indicators of disabilities, and others). In the interest of simplicity, we assume the probability of these kinds of distortions are equal for all the files. Also, we impose the requirement that each individual is composed of at most one record from each file. We model the probability of a linkage structure $\boldsymbol{\Lambda}$ of the data $\boldsymbol{X}$ via

$$P(\boldsymbol{X}|\boldsymbol{\Lambda}) = \int \left[ \prod_{i=1}^{k} \prod_{j=1}^{n_i} P(\boldsymbol{x}_{ij}|\boldsymbol{y}, \boldsymbol{\Lambda}) \right] dP(\boldsymbol{y})$$

where $\boldsymbol{\Lambda} = (\lambda_{11}, \ldots, \lambda_{kn_k})$ and $\lambda_{ij}$ represents the linkage of the the $j$th record within the $i$th file. Thus, we treat each individual as being generated independently, by first generating some underlying "true" or "canonical" variables $\boldsymbol{y}$ (from $P(\boldsymbol{y})$ —the underlying population), and then by distorting these measurements via $P(\boldsymbol{x}|\boldsymbol{y})$ for each record. In principle, we may designate a parametric or nonparametric form for each component $P(\boldsymbol{x}|\boldsymbol{y})$ and $P(\boldsymbol{y})$ in the above formulation.

Before specifying the proposed models, we formalize our notation used throughout the paper. Let $\boldsymbol{x}_{ij}$ be the data for the $j$th individual in dataset $i$, where $i = 1, \ldots, k,\ j = 1, \ldots, n_i$, and $n_i$ is the number of individuals in dataset $i$. Let $\boldsymbol{y}_{j'}$ be the latent variable for the $j'$th individual in the population, where $j' = 1, \ldots, N$. Note that $N$ could be as small as $N = n_1 = n_2 = \cdots = n_k$ if every individual has a record in all $k$ datasets or as large as $N = n_1 + n_2 + \cdots + n_k$ if no datasets have any individuals in common. Now define $\boldsymbol{\Lambda} = \{\lambda_{ij}\ ;\ i = 1, \ldots, k\ ;\ j = 1, \ldots, n_i\}$ where $\lambda_{ij}$ is an integer from 1 to $N$ such that the data for the $j$th individual in dataset $i$ ($\boldsymbol{x}_{ij}$) corresponds to the latent variable for the $\lambda_{ij}$th individual in the population ($\boldsymbol{y}_{\lambda_{ij}}$). Let $\delta_{\boldsymbol{y}_{\lambda_{ij}\ell}}$ denote the distribution of a point mass at $\boldsymbol{y}_{\lambda_{ij}\ell}$. So when we write $U \sim \delta_{\boldsymbol{y}_{\lambda_{ij}\ell}}$ we mean that $U = \boldsymbol{y}_{\lambda_{ij}\ell}$ almost surely.

### 2.1 A Bayesian Parametric Model Assuming Independent Fields

We first formulate a simple Bayesian parametric model, where we are able to write the joint posterior in closed form and sample from the full conditionals using a hybrid MCMC algorithm that employs split-merge proposals.

The model can be written as

$$\boldsymbol{x}_{ij\ell} \mid \lambda_{ij}, \boldsymbol{y}_{\lambda_{ij}\ell}, z_{ij\ell}, \boldsymbol{\theta}_\ell \overset{\text{ind}}{\sim} \begin{cases} \delta_{\boldsymbol{y}_{\lambda_{ij}\ell}} & \text{if } z_{ij\ell} = 0 \\ \text{Multinomial}(1, \boldsymbol{\theta}_\ell) & \text{if } z_{ij\ell} = 1 \end{cases}$$

$$z_{ij\ell} \overset{\text{ind}}{\sim} \text{Bernoulli}(\beta_\ell)$$

$$\boldsymbol{y}_{j'\ell} \mid \boldsymbol{\theta}_{j\ell} \overset{\text{ind}}{\sim} \text{Multinomial}(1, \boldsymbol{\theta}_\ell)$$

$$\boldsymbol{\theta}_\ell \overset{\text{ind}}{\sim} \text{Dirichlet}(\boldsymbol{\mu}_\ell)$$

$$\beta_\ell \overset{\text{ind}}{\sim} \text{Beta}(a_\ell, b_\ell)$$

where $i = 1, \ldots, k;\ j = 1, \ldots, n_i;\ j' = 1, \ldots, N;\ \ell = 1, \ldots, p$.
Also, let $m = 1, \ldots, M_\ell$ index the possible categories of the $\ell$th field, and let $\delta_{y_{\lambda_{ij}\ell m}}(x_{ij\ell m})$ denote the Dirac delta function with a "spike" at $y_{\lambda_{ij}\ell m}$ evaluated at $x_{ij\ell m}$.

Finally, we specify $\pi(\mathbf{\Lambda}) \propto 1$ for the remainder of the paper, i.e., we specify that every legal configuration of the $\lambda_{ij}$ is equally likely a priori. However, this obviously corresponds to a non-uniform prior on other related quantities of interest, such as the number of individuals in the data.

### 2.2 Hybrid Metropolis-Hastings Algorithm for Parametric Model

The joint posterior and the full conditional distributions can be derived for the proposed parametric model, however, Gibbs sampling here is not optimal in terms of convergence. Instead we propose a hybrid MCMC algorithm that performs split-merges using a Metropolis-Hastings approach to ensure that the resulting Markov chain maintains ergodicity. At each step, the algorithm proposes altering only a small number of elements of the linkage structure and latent variables, making the computation of the Metropolis-Hastings acceptance probabilities reasonably simple.

The algorithm proceeds as follows: Beginning with a linkage structure in which each record corresponds to a different individual, we choose two records uniformly at random at each iteration. If they belong to the same individual then all records assigned to that individual are randomly reassigned, either to the same individual or to a new individual currently not associated with any records, subject to the restriction that exactly one of the two particular records we chose must be assigned to the new individual. Otherwise, if the two records chosen correspond to different individuals, then all records associated with these individuals are reassigned to a single individual. However, if such a "merge" would result in two records in the same file being assigned to the same individual, then instead we take no action. Furthermore, since there are multiple latent variables associated with each individual (the field contents as well as the per-record distortion indicators), these are Gibbs sampled within each iteration of the Metropolis-Hastings sampler. The basis for this is that the use of a fixed proposal distribution for these variables would result in the rejection of many proposed steps.

Finally, we note that the above algorithm may be slightly generalized to allow more than one split or merge operation to be performed at each step. This allows the algorithm to make somewhat larger "jumps" through the space of linkage structures with each step, at the expense of tending to reduce each step's acceptance probability. Thus, the number of split or merge operations per step functions as a sort of tuning parameter that provides somewhat finer control over the algorithm's behavior and that can be specified to suit the problem at hand.

### 2.3 Application to the National Long-Term Care Survey

We test the parametric model on the NLTCS, which consists of five files, corresponding to survey responses from approximately 20,000 individuals who were tracked and surveyed at five-year intervals. At each wave of the survey, some individuals had died and were replaced by a new cohort, thus the files contain overlapping but different sets of individuals. We perform the record matching on three of the five files using our hybrid MCMC algorithm over 100,000 split-merge proposals per iteration (where each iteration is run 1,000 times). This results in

| | true match | false match |
|---|---|---|
| parametric model | 18,635 | 3,381 |
| ground truth | 28,246 | |

Table 1: True and false matches for the parametric model versus the ground truth

with regard to the "true model," since we can track each patient in the study, we know there are 9,904 two-way matches and 6,114 three-way matches (we are not counting individuals). Although this table presents only the results of a binary match/non-match decision, it should be noted that our actual results actually provide posterior match probabilities between each pair of records, a much richer collection of information.

## 3 A Bayesian Non-Parametric Model with Clustering

One problem with our parametric model in the previous section is that it does not take into account dependencies between fields. Recall that combined fields could be (Male, 84, 116 Walton Street). Given that we know a resident lives in Florida and is 64 years of age, there are certain features or dependencies that this individual shares with other 64-year-olds living in Florida. This information should be incorporated into the model, and this is why we suggest putting a Dirichlet Process on the unobserved latent record $\boldsymbol{y}_j$. String-valued variables, such as address or name, provide another motivation for this approach. We first write the model out assuming that the data is only categorical in nature, without strings. The linkage structure is taken to be the same as in the parametric model. Note that we can either cluster the latent records together or we can cluster individuals instead. Clustering the former will most likely result in too much dependence occurring or rather over-clustering and the latter will have the opposite effect. Finding a compromise between the two would be optimal. We refer to the model that clusters the latent records as $M_L$ and we define the model that clusters at the individual field level to be $M_I$.

Then the new part of $M_L$ that takes into account the clustering can be written as

$$\begin{aligned} \boldsymbol{y}_{j'} \mid G &\sim G && (j' = 1, \ldots, N) \\ G \mid \boldsymbol{\theta} &\sim \text{DP}(\alpha, G_0) \\ G_0 &= \prod_{\ell=1}^{p} \text{MN}(1, \boldsymbol{\theta}_l) \\ \boldsymbol{\theta}_\ell &\overset{\text{iid}}{\sim} \text{Dirichlet}(\boldsymbol{\mu}_i). \end{aligned}$$

However, if we propose to cluster at individual field levels, we require a different model. This model proposes to have the value of $\boldsymbol{\theta}$ vary from individual to individual, rather than just from field to field. We then cluster these $\boldsymbol{\theta}_{j''}$ values instead of the $\boldsymbol{y}_{j'}$ values. We can write the new component due to the clustering of individuals of model $M_I$ as

$$\begin{aligned} \boldsymbol{y}_{j'} \mid \boldsymbol{\theta}_{j'} &\overset{\text{ind}}{\sim} \prod_{\ell=1}^{p} \text{MN}(1, \boldsymbol{\theta}_{j'}) && (j' = 1, \ldots, N) \\ \boldsymbol{\theta}_{j'} \mid G &\sim G \\ G &\sim \text{DP}(\alpha, G_0) \\ G_0 &= \prod_{\ell=1}^{p} \text{Dirichlet}(1, \boldsymbol{\mu}_l). \end{aligned}$$

It can be shown that for each proposed model each joint posterior is difficult to sample from and hence, and we once again propose a hybrid MCMC approach with split-merge proposals. This is done in the spirit of [3] along with the standard record linkage block heuristic mentioned previously.

## 4 Summary and Future Work

We have provided Bayesian parametric and nonparametric models for model record linkage in a three-file setting, which is a new development in the literature as previously explained. Moreover, we have performed a preliminary analysis on our Bayesian parametric model as discussed in Section 2.3 with application to the NLTCS. In future work, we will report our results on the Bayesian nonparametric model for the NLTCS. We propose extending these methods to $k$ files as well as specifying a more desirable prior on the linkage structure. We then propose to extend this work to more extensive parametric and/or nonparametric models to improve our ability to find true matches for more complicated and larger datasets. Finally, we propose to then test our methodology on the 2010 Census, 2010 Census Coverage Measurement (CCM) Program, and the American Community Survey (ACS).

## References


[1] I. Fellegi and A. Sunter. A theory for record linkage. *Journal of the American Statistical Association*, 64(328):1183–1210, 2011.

[2] Rob Hall and Stephen E. Fienberg. Valid statistical inference on automatically matched files. *PSD 2012*, 2012.

[3] Sonia Jain and Radford Neal. A split-merge markov chain monte carlo procedure for the dirichlet process mixture model. *Journal of Computational and Graphical Statistics*, 13:158–182, 2000.

[4] Mauricio Sadinle and Stephen E. Fienberg. A generalized Fellegi-Sunter framework for multiple record linkage with application to homicide record-systems, 2012. pre-print.

[5] Mauricio Sadinle, Rob Hall, and Stephen E. Fienberg. Approaches to multiple record linkage, 2011. ISI invited paper.

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