A Bayesian Approach to Graphical Record Linkage and De-duplication

Rebecca C. Steorts[1], Rob Hall[2], and Stephen E. Fienberg[3]

Department of Statistical Science, Duke University, Durham NC 27707
[2]Etsy, Inc, Brooklyn, NY, [3]Department of Statistics,
Carnegie Mellon University, Pittsburgh, PA 15213
beka@stat.duke.edu, rhall@etsy.com, fienberg@andrew.cmu.edu

November 3, 2015

Abstract

We propose an unsupervised approach for linking records across arbitrarily many files, while simultaneously detecting duplicate records within files. Our key innovation involves the representation of the pattern of links between records as a *bipartite* graph, in which records are directly linked to latent true individuals, and only indirectly linked to other records. This flexible representation of the linkage structure naturally allows us to estimate the attributes of the unique observable people in the population, calculate transitive linkage probabilities across records (and represent this visually), and propagate the uncertainty of record linkage into later analyses. Our method makes it particularly easy to integrate record linkage with post-processing procedures such as logistic regression, capture-recapture, etc. Our linkage structure lends itself to an efficient, linear-time, hybrid Markov chain Monte Carlo algorithm, which overcomes many obstacles encountered by previously record linkage approaches, despite the high-dimensional parameter space. We illustrate our method using longitudinal data from the National Long Term Care Survey and with data from the Italian Survey on Household and Wealth, where we assess the accuracy of our method and show it to be better in terms of error rates and empirical scalability than other approaches in the literature.

1   Introduction

When data about individuals comes from multiple sources, it is often desirable to match, or link, records from different files that correspond to the same individual. Other names associated with record linkage are entity disambiguation, entity resolution, and coreference resolution, meaning that records which are *linked* or *co-referent* can be thought of as corresponding to the same underlying *entity* (Christen, 2012). Solving this problem is not just important as a preliminary step to statistical analysis; the noise and distortions in typical data files make it a difficult, and intrinsically

1

high-dimensional, problem (Herzog, Scheuren and Winkler, 2007; Lahiri and Larsen, 2005; Winkler, 1999, 2000).

Our methodological advances are in our unified representation of record linkage and de-duplication, via the linkage structure. This lends itself to the use of a large family of models. The particular one we put forward in this paper is the most basic and minimal member of this family. We study it not for its realism but for its simplicity and to show what even such a simple model of the family can do. Thus, we propose a Bayesian approach to the record linkage problem based on a parametric model for categorical data that addresses matching k files simultaneously and includes duplicate records within lists. We represent the pattern of matches and non-matches as a bipartite graph, in which records are directly linked to the true but latent individuals which they represent, and only indirectly linked to other records. Such *linkage structures* allow us to simultaneously address three problems: record linkage, de-duplication, and estimation of unique observable population attributes. The Bayesian paradigm naturally handles uncertainty about linkage, which poses a difficult challenge to frequentist record linkage techniques. (Liseo and Tancredi (2013) review Bayesian contributions to record linkage). A Bayesian approach permits valid inference regarding posterior matching probabilities of records and propagation of errors, as we discuss in Section 4.

To estimate our model, we develop a hybrid MCMC algorithm, in the spirit of Jain and Neal (2004), which runs in linear time in the number of records and the number of MCMC iterations, even in high-dimensional parameter spaces. Our algorithm permits duplication across and within lists but runs faster if there are known to be no duplicates within lists. We achieve further gains in speed using standard record linkage blocking techniques (Christen, 2012).

We apply our method to data from the National Long Term Care Survey (NLTCS), which tracked and surveyed approximately 20,000 people at five-year intervals. At each wave of the survey, some individuals had died and were replaced by a new cohort, so the files contain overlapping but not identical sets of individuals, with no within-file duplicates. We also apply our method to data from the Italian Survey on Household and Wealth (FWIW), a sample survey 383 households conducted by the Bank of Italy every two years. We introduce this application to compare our method to that of Tancredi and Liseo (2011), whereas applying the latter to the NLTCS study

is not computationally feasible in a reasonable amount of time using the competitor's method as it would take roughly 1500 hours (62 days), whereas our method takes 3 hours. We explore the validity of our method using simulated data.

Section 2 provides a motivating example of the record linkage problem. In Section 3.1, we introduce the notation and model, and describe the algorithm in Section 3.2. Sections 3.3 and 3.4 introduce posterior matching sets which upholding transitivity, and taking functions of the linkage structure. In Sections 4 and 5 we apply the method to the the NLTCS under two algorithms (SMERE and SMERED). We also do comparisons, showing that SMERED beats the method of Tancredi and Liseo (2011) for every region of Italy from the FWIW. In Section 6 we review the strengths and limitations of our method, while providing a user's guide for performing record linkage. We evaluate each method compared to a simple baseline and explore the validity of our method under simulation studies. We discuss future directions in Section 7.

1.1 Related Work

The classical work of Fellegi and Sunter (1969) considered linking two files in terms of Neyman-Pearson hypothesis testing. Compared to this baseline, our approach is distinctive in that it handles multiple files, models distortion explicitly, offers a Bayesian treatment of uncertainty and error propagation, and employs a sophisticated graphical data structure for inference to latent individuals. Methods based upon Fellegi and Sunter (1969) can extend to $k > 2$ files (Sadinle and Fienberg, 2013), but they break down for even moderately large k or complex data sets. Moreover, they provide little information about uncertainty in matches, or about the true values of noise-distorted records. Copas and Hilton (1990) describe the idea of modeling the distortion process using what they call the "Hit-Miss Model," which anticipates part of our model in Section 3.1. The specific distortion model we use is, however, closer to that introduced in Hall and Fienberg (2012), as part of a nonparametric frequentist technique for matching $k = 2$ files that allows for distorted data. Thus, their work is related to ours as we view the records as noisy, distorted entities, that we model using parameters and latent individuals.

There has been much work done in clustering and latent variable modeling in statistics, but

also in machine learning and computer science, where of these applications are focused toward author disambiguation. For example, Bhattacharya and Getoor (2006) proposed a record linkage method based on latent Dirichlet allocation, which infers the total number of unobserved entities (authors). Their method assumes labeled data such that co-authorship groups can be estimated. In similar work, Dai and Storkey (2011) used a a non-parametric Dirichlet Process (DP) model, where groups of authors are associated with topics instead of individual authors. This works well for these specific types of applications such as author disambiguation. However, when clustering records to a hypothesized latent individual, the number of latents typically does not grow as the size of the records does. Hence, a DP process or any non-parametric process tends to over-cluster since we have what we call a *small clustering problem.* Since our goal is to handle a variety of applications including author disambiguation, extensions of our model should be able to handle applications like author disambiguation either using parametric or non-parametric models in sound and principled ways.

Within the Bayesian paradigm, most work has focused on specialized approaches related to linking two files, which propagate uncertainty (Belin and Rubin, 1995; Fienberg, Makov and Sanil, 1997; Gutman, Afendulis and Zaslavsky, 2013; Larsen and Rubin, 2001; Tancredi and Liseo, 2011). These contributions, while valuable, do not easily generalize to multiple files and to duplicate detection. Three recent papers (Domingos and Domingos, 2004; Gutman et al., 2013; Sadinle, 2014) are most relevant to the novelty of our work, namely the linkage structure. To aid in the recovery of information about the population from distorted records, Gutman et al. (2013) called for developing "more sophisticated network data structures." Our linkage graphs are one such data structure with the added benefit of permitting de-duplication and handling multiple files. Moreover, due to exact error propagation, we can easily integrate our methods with other analytic procedures. Algorithmically, the closest approach to our linkage structure is the graphical representation in Domingos and Domingos (2004), for de-duplication within one file. Their representation is a unipartite graph, where records are linked to each other. Our use of a bipartite graph with latent individuals naturally fits in the Bayesian paradigm along with distortion. Our method is the first to handle record linkage and de-duplication, while also modeling distortion and running in linear

time. Finally, Sadinle (2014) recently extended our linkage structure in the representation of a conference matrix, or rather partitioning approaching, where they deviate from our methods using comparison data as to use both categorical and non-categorical data. However, one advantage we maintain is that our methods are more scalable when the data is categorical since a Gibbs sampler does not explore the parameter space as efficiently as a hybrid MCMC approach.

2   Motivating Example

The databases (files) contain records regarding individuals that are distorted versions of their unobserved true attributes (fields). We assume that each record corresponds to only one unobserved latent individual. These distortions have various causes—measurement errors, transcription errors, lies, etc.—which we do not model. We do, however, assume that the probability of distortion is the same for all files (and we do so for computational convenience). Such distortions, and the lack of unique identifiers shared across files, make it ambiguous when records refer to the same individuals. This ambiguity can be reduced by increasing the amount of information available, either by adding informative fields to each record, or, sometimes, by increasing the number of files.

We illustrate this issue with a motivating example of real world distortion and noise (see Table 1). With gender and state alone, these records could refer to (i) a single individual with a data entry error (or old address, etc.) in one file; (ii) one individual correctly recorded as living in SC and another correctly recorded in WV; (iii) two individuals with errors in at least one address; (iv) three distinct individuals with correct addresses; (v) three individuals with errors in addresses. (There are still further possibilities if the gender field might contain errors.) The goal is to determine whether distinct records refer to the same individual, or to distinct individuals.

Table 1 illustrates a scenario where there is considerable uncertainty about whether two records correspond to the same individual (under just gender and state). As we mentioned earlier, the identities of the true individuals to which the records correspond are not entirely clear due to the limited information available. Suppose we expand the field information by adding date of birth (DOB) and race. We still have the same host of possibilities as before, but the addition of DOB *may* let us make better decisions about matches. It is not clear if File 1 and File 2 are the same

person or different people (who just happen to have the same birthdate). However, the introduction of DOB does make it more likely that File 3 is not the same person as in File 1 and File 2. The method we propose in Section 3.1 deals with this type of noise; however, it proposes to deal with noisier records in that they traditionally do not have identifying information such as name and address, making the matching problem inherently difficult, even indeterminate.

	Gender	State	DOB	Race
File 1	F	SC	04/15/83	White
File 2	F	WV	04/15/83	White
File 3	F	SC	07/25/43	White

Table 1: Three files with year of birth and race.

3 Notation, Assumptions, and Linkage Structure

We begin by defining some notation, for k files or lists. For simplicity, we assume that all files contain p fields in common, which are all categorical, field ℓ having M_ℓ levels. Thus, we do not handle missingness of fields across databases, however when we do have p fields in common, some information could be missing. Handling missing-at-random fields within records is a minor extension within the Bayesian framework (Reiter and Raghunathan, 2007). Let $\boldsymbol{x}_{ij}$ be the data for the jth record in file i, where $i = 1, \ldots, k$, $j = 1, \ldots, n_i$, and n_i is the number of records in file i; $\boldsymbol{x}_{ij}$ is a categorical vector of length p. Let $\boldsymbol{y}_{j'}$ be the latent vector of true field values for the j'th individual in the population (or rather aggregate sample), where $j' = 1, \ldots, N$, N being the total number of *observed* individuals from the population. N could be as small as 1 if every record in every file refers to the same individual or as large as $N_{\max} \equiv \sum_{i=1}^k n_i$ if no datasets share any individuals.

Next, we define the linkage structure $\boldsymbol{\Lambda} = \{\lambda_{ij} \; ; \; i = 1, \ldots, k \; ; \; j = 1, \ldots, n_i\}$ where λ_{ij} is an integer from 1 to $N_{\max}$ indicating which latent individual the jth record in file i refers to, i.e., $\boldsymbol{x}_{ij}$ is a possibly-distorted measurement of $\boldsymbol{y}_{\lambda_{ij}}$. Finally, $z_{ij\ell}$ is 1 or 0 according to whether or not a particular field ℓ is distorted in $\boldsymbol{x}_{ij}$.

As usual, we use I for indicator functions (e.g., $I(x_{ij\ell} = m)$ is 1 when the ℓth field in record j in file i has the value m), and δ_a for the distribution of a point mass at a (e.g., $\delta_{y_{\lambda_{ij}\ell}}$). The

vector $\boldsymbol{\theta}_\ell$ of length M_ℓ denotes the multinomial probabilities. For clarity, we always index as follows: $i = 1, \ldots, k;\, j = 1, \ldots, n_i;\, j' = 1, \ldots, N;\, \ell = 1, \ldots, p;\, m = 1, \ldots, M_\ell$. We provide an example of the linkage structure and a graphical representation in Appendix A.

3.1 Independent Fields Model

We assume the k files are conditionally independent, given the latent individuals, and that fields are independent within individuals (this is done for computational simplicity as is the motivation for our Bayesian model). We formulate the following Bayesian parametric model:

$$\boldsymbol{x}_{ij\ell} \mid \lambda_{ij}, \boldsymbol{y}_{\lambda_{ij}\ell}, z_{ij\ell}, \boldsymbol{\theta}_\ell \stackrel{\text{ind}}{\sim} \begin{cases} \delta_{\boldsymbol{y}_{\lambda_{ij}\ell}} & \text{if } z_{ij\ell} = 0 \\[2ex] \mathrm{MN}(1, \boldsymbol{\theta}_\ell) & \text{if } z_{ij\ell} = 1 \end{cases}$$

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

$$\boldsymbol{y}_{j'\ell} \mid \boldsymbol{\theta}_\ell \stackrel{\text{ind}}{\sim} \mathrm{MN}(1, \boldsymbol{\theta}_\ell)$$

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

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

$$\pi(\boldsymbol{\Lambda}) \propto 1,$$

where a_ℓ, b_ℓ, and $\boldsymbol{\mu}_\ell$ are all known.

Remark 3.1: We assume that every legitimate configuration of the λ_{ij} is equally likely a priori. This implies a non-uniform prior on related quantities, such as the number of individuals in the data. The uniform prior on $\boldsymbol{\Lambda}$ is convenient, since it simplifies computation of the posterior. Devising non-uniform priors over linkage structures remains a challenging problem both computationally and statistically, as sensible priors must remain invariant when permuting the labels of latent individuals, and cannot use covariate information about records.

Deriving the joint posterior and conditional distributions is mostly straightforward. One subtlety, however, is that $\boldsymbol{y}$, $\boldsymbol{z}$ and $\boldsymbol{\Lambda}$ are all related, since if $z_{ij\ell} = 0$, then it must be the case that

$y_{\lambda_{ij}\ell} = x_{ij\ell}$. Taking this into account, the joint posterior is

$$\pi(\mathbf{\Lambda}, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\beta} \mid \boldsymbol{x}) \propto \prod_{i,j,\ell,m} \left[(1 - z_{ij\ell}) \delta_{y_{\lambda_{ij}\ell}}(x_{ij\ell}) + z_{ij\ell} \theta_{\ell m}^{I(x_{ij\ell}=m)} \right]$$

$$\times \prod_{\ell,m} \theta_{\ell m}^{\mu_{\ell m} + \sum_{j'=1}^{N} I(y_{j'l}=m)} \times \prod_{\ell} \beta_{\ell}^{a_\ell - 1 + \sum_{i=1}^{k} \sum_{j=1}^{n_i} z_{ij\ell}}$$

$$\times (1 - \beta_\ell)^{b_\ell - 1 + \sum_{i=1}^{k} \sum_{j=1}^{n_i} (1 - z_{ij\ell})}.$$

Now we consider the conditional distribution of $\boldsymbol{y}$. Here, the part of the posterior involving $\boldsymbol{x}$ only matters for the conditional of $\boldsymbol{y}$ when $z_{ij\ell} = 0$. Specifically, when $z_{ij\ell} = 0$, we know that $y_{\lambda_{ij}\ell} = x_{ij\ell}$. Next, for each $j' = 1, \ldots, N$, let $R_{ij'} = \{j : \lambda_{ij} = j'\}$, so that $\boldsymbol{x}_{ij}$ and $\boldsymbol{y}_{j'}$ refer to the same individual if and only if $j \in R_{ij'}$.

Remark 3.2: This notation allows the consideration of duplication within lists, i.e., distinct records within a list that correspond to the same individual. In particular, two records j_1 and j_2 in the same list i correspond to the same individual if and only if $\lambda_{ij_1} = \lambda_{ij_2}$. Implementing this in our hybrid MCMC is *simpler* than assuming the lists are already de-duplicated, since de-duplication implies that certain linkages are undefined.

From the joint posterior above, we can write down the full conditional distributions of $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ directly as

$$\beta_\ell \mid \mathbf{\Lambda}, \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{y}, \boldsymbol{x} \sim \text{Beta}\left(a_\ell + \sum_{i=1}^{k} \sum_{j=1}^{n_i} z_{ij\ell}, b_\ell + \sum_{i=1}^{k} \sum_{j=1}^{n_i} (1 - z_{ij\ell}) \right)$$

for all ℓ and

$$\theta_{\ell m} \mid \mathbf{\Lambda}, \boldsymbol{z}, \boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{x} \sim \text{Dirichlet}\left(\mu_{\ell m} + \sum_{j'=1}^{N} y_{j'\ell} + \sum_{i=1}^{k} \sum_{j=1}^{n_i} z_{ij\ell}\, x_{ij\ell} + 1 \right),$$

8

for all ℓ and for each $m = 1, \ldots, M_\ell$. Then

$$
\boldsymbol{y}_{j'l} \mid \boldsymbol{\Lambda}, \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{x} \sim \begin{cases} \delta_{x_{ij\ell}} & \text{if there exist } i, j \in R_{ij'} \text{ such that } z_{ij\ell} = 0, \\ \text{Multinomial}(1, \boldsymbol{\theta}_l) & \text{otherwise.} \end{cases}
$$

In other words, the linkage structure $\boldsymbol{\Lambda}$ tells us that $\boldsymbol{y}_{j'}$ corresponds to some $\boldsymbol{x}_{ij}$ when there is no distortion. Now consider that $\boldsymbol{z}$ represents the indicator of whether or not there is a distortion. When we condition on $\boldsymbol{x}, \boldsymbol{y}$, and $\boldsymbol{\Lambda}$, there are times when a distortion is certain. Specifically, if $x_{ij\ell} \neq y_{\lambda_{ij}\ell}$, then we know there must be a distortion, so $z_{ij\ell} = 1$. However, if $x_{ij\ell} = y_{\lambda_{ij}\ell}$, then $z_{ij\ell}$ may or may not equal 0. Therefore, we can show

$$
\begin{aligned}
& P(z_{ij\ell} = 1 \mid \boldsymbol{\Lambda}, \boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{x}) \\
&= \frac{P(\boldsymbol{\Lambda}, \boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{x} \mid z_{ij\ell} = 1)P(z_{ij\ell} = 1)}{P(\boldsymbol{\Lambda}, \boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{x} \mid z_{ij\ell} = 1)P(z_{ij\ell} = 1) + P(\boldsymbol{\Lambda}, \boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{x} \mid z_{ij\ell} = 0)P(z_{ij\ell} = 0)} \\
&= \frac{\beta_\ell \prod_{m=1}^{M_\ell} \theta_{\ell m}^{x_{ij\ell}}}{\beta_\ell \prod_{m=1}^{M_\ell} \theta_{\ell m}^{x_{ij\ell}} + (1 - \beta_\ell)}.
\end{aligned}
$$

Then we can write the conditional as

$$
z_{ij\ell} \mid \boldsymbol{\Lambda}, \boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{x} \stackrel{\text{ind}}{\sim} \text{Bernoulli}(p_{ij\ell}), \text{ where}
$$

$$
p_{ij\ell} = \begin{cases} 1 & \text{if } x_{ij\ell} \neq y_{\lambda_{ij}\ell} \\ \dfrac{\beta_\ell \prod_{m=1}^{M_\ell} \theta_{\ell m}^{x_{ij\ell}}}{\beta_\ell \prod_{m=1}^{M_\ell} \theta_{\ell m}^{x_{ij\ell}} + (1 - \beta_\ell)} & \text{if } x_{ij\ell} = y_{\lambda_{ij}\ell}, \end{cases} \quad \text{for all } \ell.
$$

Finally, we derive the conditional distribution of $\boldsymbol{\Lambda}$. This is the only part of the model which changes if we allow duplication within lists. Conditional on $\boldsymbol{x}, \boldsymbol{y}$, and $\boldsymbol{z}$, we can rule out there are many linkage structures , i.e., ones that have probability zero. Specifically, for any i, j, ℓ such that $z_{ij\ell} = 0$, there is no distortion. This means that if $z_{ij\ell} = 0$, then for any c such that $x_{ij\ell} \neq y_{cl}$, we know that $\lambda_{ij} = c$ is impossible. On the other hand, if $z_{ij\ell} = 1$, then $\boldsymbol{x}_{ij\ell}$ simply comes from a multinomial, in which case the linkage structure is totally irrelevant. If we assume that no duplication is allowed within each list, then $j_1 \neq j_2$ implies that $\lambda_{ij_1} \neq \lambda_{ij_2}$. Additionally, we note

that for each file i, the part of the linkage structure corresponding to that dataset $(\lambda_{i1}, \ldots, \lambda_{in_i})$ is independent of the linkage structures for the other datasets conditional on everything else. Then we can write the following (assuming no duplicates):

$$
P(\lambda_{i1} = c_1, \ldots, \lambda_{in_i} = c_{n_i} \mid \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{x}) \overset{\text{ind}}{\propto}
\begin{cases}
0 & \text{if there exist } j, \ell \text{ such that} \\
& \quad z_{ij\ell} = 0 \text{ and } x_{ij\ell} \neq y_{c_j\ell}, \\
& \quad \text{or if } c_{j_1} = c_{j_2} \text{ for any } j_1 \neq j_2 \\
1 & \text{otherwise,}
\end{cases}
$$

where the somewhat nonstandard notation $\overset{\text{ind}}{\propto}$ simply denotes that distributions for different i are independent.

Allowing duplicates within lists, we find that the conditional distribution lifts a restriction from the one just derived. That is,

$$
P(\lambda_{i1} = c_1, \ldots, \lambda_{in_i} = c_{n_i} \mid \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{x}) \overset{\text{ind}}{\propto}
\begin{cases}
0 & \text{if there exist } j, \ell \text{ such that} \\
& \quad z_{ij\ell} = 0 \text{ and } x_{ij\ell} \neq y_{c_j\ell}, \\
1 & \text{otherwise.}
\end{cases}
$$

3.2 Split and MErge REcord linkage and De-duplication (SMERED) Algorithm

Our main goal is estimating the posterior distribution of the linkage (i.e., the clustering of records into latent individuals). The simplest way of accomplishing this is via Gibbs sampling. We could iterate through the records, and for each record, sample a new assignment to an individual (from among the individuals represented in the remaining records, plus an individual comprising only that record). However, this requires the quadratic-time checking of proposed linkages for every record. Thus, instead of Gibbs sampling, we use a hybrid MCMC algorithm to explore the space of possible linkage structures, which allows our algorithm to run in linear time.

Our hybrid MCMC takes advantage of split-merge moves, as done in Jain and Neal (2004), which avoids the problems associated with Gibbs sampling, even though the number of parameters

10

grows with the number of records. This is accomplished via proposals that can traverse the state space quickly and frequently visit high-probability modes, since the algorithm splits or merges records in each update, and hence, frequent updates of the Gibbs sampler are not necessary.

Blocking is a common technique that creates places similar records into partitions or blocks based on some rule or automated process. For a review of cladssical and newer blocking methods see (Steorts, Ventura, Sadinle and Fienberg, 2014; Winkler, 2000). The form of blocking that we illustrate in our examples requires an exact match in certain fields (e.g., birth year) if records are to be linked. *Blocking* can greatly reduce the number of possible links between records. In our application, we use an approximate blocking procedure, not an exact one. Since blocking gives up on finding truly co-referent records which disagree on those fields, it is best to block on fields that have little or no distortion. We block on the fairly reliable fields of sex and birth year in our application to the NLTCS below. A strength of our model is that it incorporates blocking organically. Setting $b_\ell = \infty$ for a particular field ℓ forces the distortion probability for that field to zero. This requires matching records to agree on the ℓth field, just like blocking.

We now discuss how the split-merge process links records to records, which it does by assigning records to latent individuals. Instead of sampling assignments at the record level, we do so at the individual level. Initially, each record is assigned to a unique individual. On each iteration, we choose two records at random. If the pair belong to *distinct* latent individuals, then we propose merging those individuals to form a single new latent individual (i.e., we propose that those records are co-referent). On the other hand, if the two records belong to the *same* latent individual, then we propose splitting it into two new latent individuals, each seeded with one of the two chosen records, and the other records randomly divided between the two. Proposed splits and merges are accepted based on the Metropolis-Hastings ratio and rejected otherwise.

Sampling from all possible pairs of records will sometimes lead to proposals to merge records in the same list. If we permit duplication within lists, then this is not a problem. However, if we know (or assume) there are no duplicates within lists, we should avoid wasting time on such pairs. The no-duplication version of our algorithm does precisely this. (See Appendix B for the algorithm and pseudocode.) When there are no duplicates within files, we call this the SMERE (Split and

MErge REcord linkage) algorithm, which enforces the restriction that $R_{ij'}$ must be either $\varnothing$ or a single record. This is done through limiting the proposal of record pairs to those in distinct files; the algorithm otherwise matches SMERED.

3.3 Posterior Matching Sets and Linkage Probabilities

In a Bayesian framework, the output of record linkage is not a deterministic set of matches between records, but a probabilistic description of how likely records are to be co-referent, based on the observed data. Since we are linking multiple files at once, we propose a range of posterior matching probabilities: the posterior probability of linkage between two arbitrary records and more generally among k records, the posterior probability that a given set of records is linked, and the posterior probability that a given set of records is a maximal matching set (which will be defined later). Furthermore, we make a connection to Tancredi and Liseo (2011) of why our proposed posterior matching probabilities our *optimal*.

Two records (i_1, j_1) and (i_2, j_2) *match* if they point to the same latent individual, i.e., if $\lambda_{i_1 j_1} = \lambda_{i_2 j_2}$. The posterior probability of a match can be computed from the S_G MCMC samples:

$$P(\lambda_{i_1 j_1} = \lambda_{i_2 j_2} | \boldsymbol{X}) = \frac{1}{S_G} \sum_{h=1}^{S_G} I(\lambda_{i_1 j_1}^{(h)} = \lambda_{i_1 j_2}^{(h)}).$$

A one-way match occurs when an individual appears in only one of the k files, while a two-way match is when an individual appears in exactly two of the k files, and so on (up to k-way matches). We approximate the posterior probability of arbitrary one-way, two-way, ..., k-way matches as the ratio of frequency of those matches in the posterior sample to S_G.

Although probabilistic results and interpretations provided by the Bayesian paradigm are useful both quantitatively and conceptually, we often need to report a point estimate of the linkage structure. Thus, we face the question of how to condense the overall posterior distribution of $\boldsymbol{\Lambda}$ into a single estimated linkage structure.

Perhaps the most obvious approach is to set some threshold v, where $0 < v < 1$, and to declare (i.e., estimate) that two records match if and only if their posterior matching probability exceeds v.

This strategy is useful if only a few specific pairs of records are of interest, but its flaws are exposed when we consider the coherence of the overall estimated linkage structure implied by such a thresholding strategy. Note that the true linkage structure is *transitive* in the following sense: if records A and B are the same individual, and records B and C are the same individual, then records A and C must be the same individual as well. This requirement of transitivity, however, is in no way enforced by the simple thresholding strategy described above. Thus, a more sophisticated approach is required if the goal is to produce an estimated linkage structure that preserves transitivity.

To this end, it is useful to define a new concept. A set of records $\mathcal{A}$ is a *maximal matching* set (MMS) if every record in the set has the same value of λ_{ij} and no record outside the set has that value of λ_{ij}. Define $\boldsymbol{\Omega}(\mathcal{A}, \boldsymbol{\Lambda})$ to be 1 if $\mathcal{A}$ is an MMS in $\boldsymbol{\Lambda}$ and 0 otherwise:

$$\boldsymbol{\Omega}(\mathcal{A}, \boldsymbol{\Lambda}) = \sum_{j'} \left(\prod_{(i,j) \in \mathcal{A}} I(\lambda_{ij} = j') \prod_{(i,j) \notin \mathcal{A}} I(\lambda_{ij} \neq j') \right).$$

Essentially, records are in the same maximal matching set if and only if they match the same latent individual, though which individual is irrelevant. Given a set of records $\mathcal{A}$, the posterior probability that it is an MMS in $\boldsymbol{\Lambda}$ is simply

$$P(\boldsymbol{\Omega}(\mathcal{A}, \boldsymbol{\Lambda}) = 1) = \frac{1}{S_G} \sum_{h=1}^{S_G} \boldsymbol{\Omega}(\mathcal{A}, \boldsymbol{\Lambda}^{(h)}).$$

The MMSs allow a more sophisticated method of preserving transitivity when estimating a single overall linkage structure. For any record (i, j), its *most probable MMS* $\mathcal{M}_{ij}$ is the set containing (i, j) with the highest posterior probability of being an MMS, i.e., $\mathcal{M}_{ij} := \arg \max_{\mathcal{A}:(i,j) \in \mathcal{A}} P(\boldsymbol{\Omega}(\mathcal{A}, \boldsymbol{\Lambda}) = 1)$. Note, however, that there are still problems with a strategy of linking each record to exactly those other records in its most probable maximal matching set. Specifically, it is possible for different records' most probable maximal matching sets to contradict each other. For example, Record A may be in Record B's most probable maximal matching set, but Record B may not be in Record A's most probable maximal matching set. To solve this problem, we define a *shared most probable MMS* to be a set that is the most probable MMS for each of its members. We then

estimate the overall linkage structure by linking records if and only if they are in the same shared most probable MMS. The resulting linkage structure is by construction transitive. We illustrate examples of MMSs and shared most probable MMSs in Section 4.2.

Finally, we can say that our shared most probably MMSs are optimal under an optimal decision G^*, i.e., the one that minimizes the posterior expected loss $G^* = \arg\min_{G \in \mathcal{G}} W(G)$, where $W(G) = E[L(\mathbf{\Lambda}, G) \mid \mathbf{X}]$. This was considered in Tancredi and Liseo (2011) under the conference matrix for two files and for the loss functions squared error, false match rate, and absolute number of errors. For our situation, it immediately follows due to Tancredi and Liseo (2011) (Theorem 4.1) that under squared error loss and absolute number of errors, the optimal Bayesian solution is simply for any record (i, j), its *most probable maximal matching set* $\mathcal{M}_{ij}$ is the set containing (i, j) with the highest posterior probability of being a maximal matching set, i.e., $\mathcal{M}_{ij} := \arg\max_{\mathcal{A}:(i,j)\in\mathcal{A}} P(\mathbf{\Omega}_{\mathcal{A},\mathbf{\Lambda}} = 1)$. Next, a *shared most probable maximal matching set* is a set that is the most probable maximal matching set of all records it contains, i.e., a set $\mathcal{A}^\star$ such that $\mathcal{M}_{ij} = \mathcal{A}^\star$ for all $(i, j) \in \mathcal{A}^\star$. We then estimate the overall linkage structure by linking records if and only if they are in the same shared most probable matching set. The resulting estimated linkage structure is guaranteed to have the transitivity property since (by construction) each record is an element of at most one shared most probable maximal matching set. By Theorem 4.1 of Tancredi and Liseo (2011), this is the optimal Bayesian decision rule under squared error loss and absolute number of errors. Then trivially under our original definition of the shared MPMMSs, it is still optimal.

3.4 Functions of Linkage Structure

The output of the Gibbs sampler also allows us to estimate the value of any function of the variables, parameters, and linkage structure by computing the average value of the function over the posterior samples. For example, estimated summary statistics about the population of latent individuals are straightforward to calculate. Indeed, the ease with which such estimates can be obtained is yet another benefit of the Bayesian paradigm, and of MCMC in particular.

4 Assessing Accuracy of Matching and Application to NLTCS

We test our model using data from the National Long Term Care Survey (NLTCS), a longitudinal study of the health of elderly (65+) individuals (http://www.nltcs.aas.duke.edu/). The NLTCS was conducted approximately every six years, with each wave containing roughly 20,000 individuals. Two aspects of the NLTCS make it suitable for our purposes: individuals were tracked from wave to wave with unique identifiers, but at each wave, many patients had died (or otherwise left the study) and were replaced by newly-eligible individuals. We can test the ability of our model to link records across files by seeing how well it is able to track individuals across waves, and compare its estimates to the ground truth provided by the unique identifiers.

To show how little information our method needs to find links across files, we gave it access to only four variables, all known to be noisy: date of birth, sex, state of residence, and the regional office at which the subject was interviewed. We linked individuals across the 1982, 1989, and 1994 survey waves. Our model had little information on which to link, and not *all* of its assumptions strictly hold (e.g., individuals can move between states across waves). We demonstrate our method's validity using error rates, confusion matrices, posterior matching sets and linkage probabilities, and estimation of the unknown number of observed individuals from the population.

Appendix C provides a simulation study of the NLTCS with varying levels of distortion at the field level. We conclude from this that SMERE is able to handle low to moderate levels of distortion (Figure 9). Furthermore, as distortion increases, so do the false negative rate (FNR) and false positive rate (FPR) (Figure 8).

4.1 Error Rates and Confusion Matrix

Since we have unique identifiers for the NLTCS, we can see how accurately our model matches records. A *true link* is a match between records which really do refer to the same latent individual; a *false link* is a match between records which refer to different latent individuals; and a *missing link* is a match which is not found by the model. Table 6 gives posterior means for the number of true, false, and missing links. For the NLTCS, the FNR is 0.11, while the FPR is 0.046, when we block by date of birth year (DOB) and sex.

More refined information about linkage errors comes from a confusion matrix, which compares records' estimated and actual linkage patterns (Figure 1 and Table 7 in Appendix D). Every row in the confusion matrix is diagonally dominated, indicating that correct classifications are overwhelmingly probable. The largest off-diagonal entry, indicating a mis-classification, is 0.07. For instance, if a record is estimated to be in both the 1982 and 1989 waves, it is 90% probable that this estimate is correct. If the true pattern for 1982 and 1989 is wrong, the estimate is most probably 1982 (0.043) and 1989 (0.033) followed by all years (0.018), and and then the other waves with small estimated probabilities.

4.2 Example of Posterior Matching Probabilities

We wish to search for sets of records that match record 10084 in 1982. In the posterior samples of Λ, this record is part of three maximal matching sets that occur with nonzero estimated posterior probability, one with high and two with low posterior matching probabilities (Table 4). This record has a posterior probability of 0.995 of simultaneously matching both record 6131 in 1989 and record 5583 in 1994. All three records denote a male, born 07/01/1910, visiting office 25 and residing in state 14. The unique identifiers show that these three records are in fact the same individual. Practically speaking, only matching sets with reasonably large posterior probability, such as the set in the last column of Table 4, are of interest.

4.3 Example of Most Probable MMSs

For each record in the NLTCS, we wish to produce its most probable maximal matching set (MP-MMS). We then wish to identify those MPMMSs that are shared. Finally, we wish to show the linked records for the shared most probable MMSs visually for a subset of the NLTCS (it is too large to show for the entire dataset).

On each Gibbs iteration, we record all the records linked to a given latent individual; this is an MMS. We aggregate MMSs across Gibbs iterations, and their relative frequencies are approximate posterior probabilities for each MMS. Each record is labeled with the most probable MMS to which it belongs. Finally, we link two records when they share the same most probable MMS, giving us
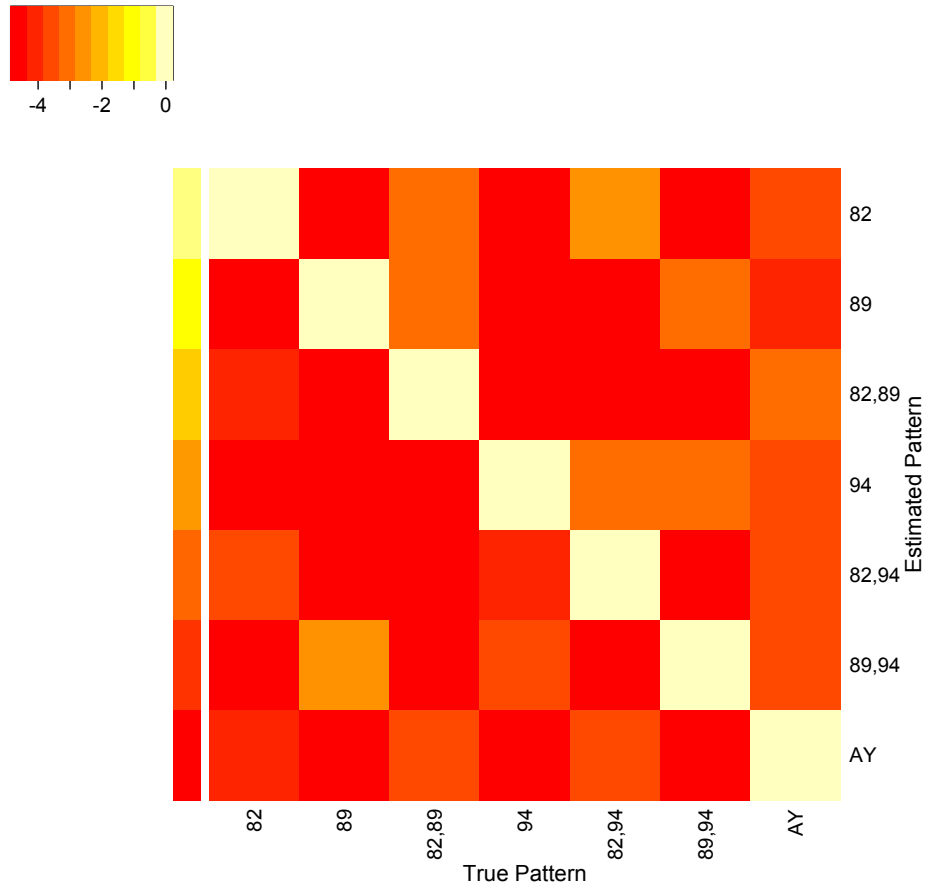
Figure 1: Heatmap of the natural logarithm of the relative probabilities from the confusion matrix, running from yellow (most probable) to dark red (least probable). The largest probabilities are on the diagonal, showing that the linkage patterns estimated for records are correct with high probability. Mis-classification rates are low and show a tendency to under-link rather than over-link, as indicated by higher probabilities for cells above the diagonal than for cells below the diagonal.

the shared most probable MMS. From this we are able to compute a FNR and a FPR (which we discuss in Section 5.1). We give an example of the most probable MMSs for the first ten rows in Table 2.

record	MPMMS	posterior probability
1.0	$\{1.0, 1.16651\}$	0.51
1.16651	$\{1.0, 1.16651\}$	0.51
1.1	$\{1.1\}$	1.00
1.10	$\{1.10\}$	0.75
1.2024	$\{1.2024, 1.21667, 1.39026\}$	0.82
1.3043	$\{1.3043\}$	0.58
1.21667	$\{1.2024, 1.21667, 1.39026\}$	0.82
1.39026	$\{1.2024, 1.21667, 1.39026\}$	0.82
1.2105	$\{1.2105, 1.21719, 1.39079\}$	0.81
1.21719	$\{1.2105, 1.21719, 1.39079\}$	0.81

Table 2: First 10 rows of most probable maximal matching sets. We represent the file and record by $A.B$, where the file comes first and the record follows the period sign. Hence, 1.10, refers to the tenth record in file one. We use this encoding as it's consistent with our coding practice and easy to refer back to the data in the NLTCS.

In Figure 2, we provide the shared most probable MMSs for the first 204 records of the NLTCS. Color indicates whether or not the probability of a shared most probable MMS was above (green) or below (red) the value 0.8. Transitivity is clear from the fact that each connected component is a clique, i.e., each record is connected to every other record in the same set. In Figure 3, we replot the same records, however, we add a feature to each set. Each edge is either straight or wavy to indicate whether the link was correct (straight) or incorrect (wavy) according to the ground truth from the NLTCS. Overall, we see that the the wavy sets tend to be red, meaning that these matches were assigned a lower posterior probability by the model, while the straight sets tend to be green, meaning that these matches were assigned a higher posterior probability by the model.

We can view the red wavy links as individuals we would push to clerical review since the algorithms has trouble matching them. For the NLTCS, manual review would possibly not do much better than our algorithm since there many individuals match on everything except unique ID. Since there is no other information to match on, we cannot hope to improve in this application.

Useful as these figures are, they would become visually unreadable for the *entire* NLTCS or

any record linkage problem of this scale or larger. This is a first step at showing that shared most probable MMSs can be visualized. Visualization on the entire graph structure would be an important advancement moving forward.
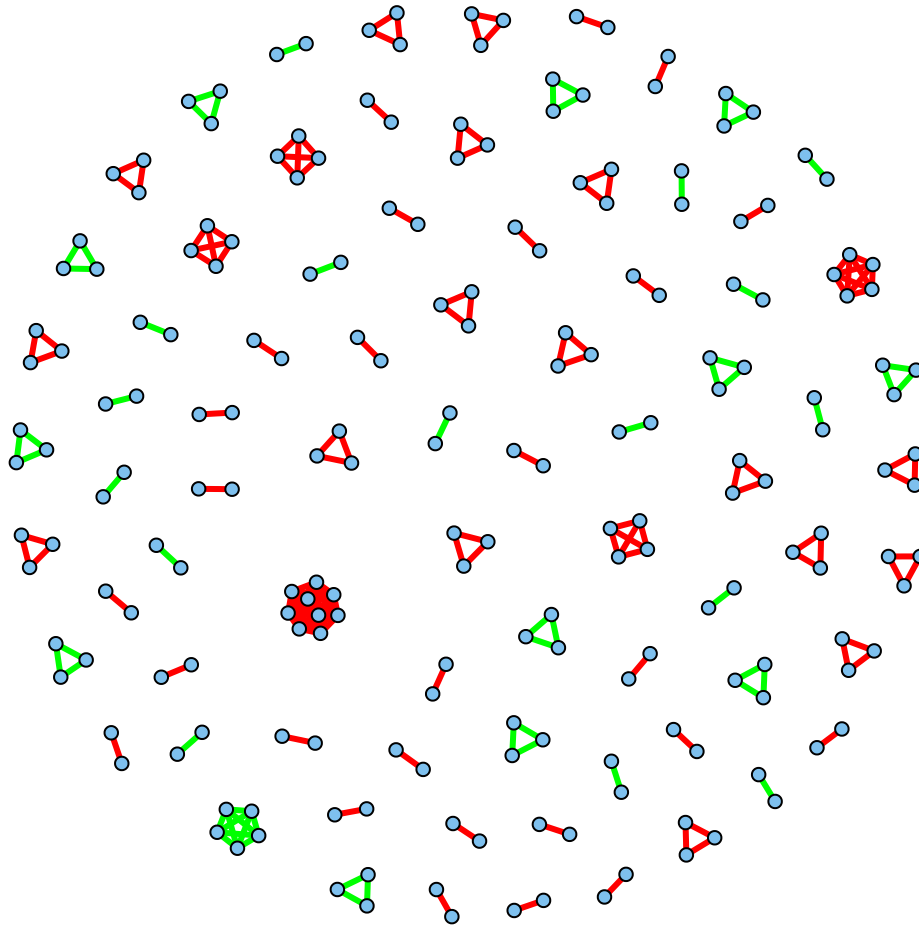

Figure 2: This illustrates the first 204 records, where each node is a record and edges are drawn between records with shared most probable MMSs. Transitivity is clear from the fact that each connected component is a clique, or rather, each record is connected to every other record in the same set. Color indicates whether or not the probability of the shared most probable MMS was above (green) or below (red) a threshold of 0.8.
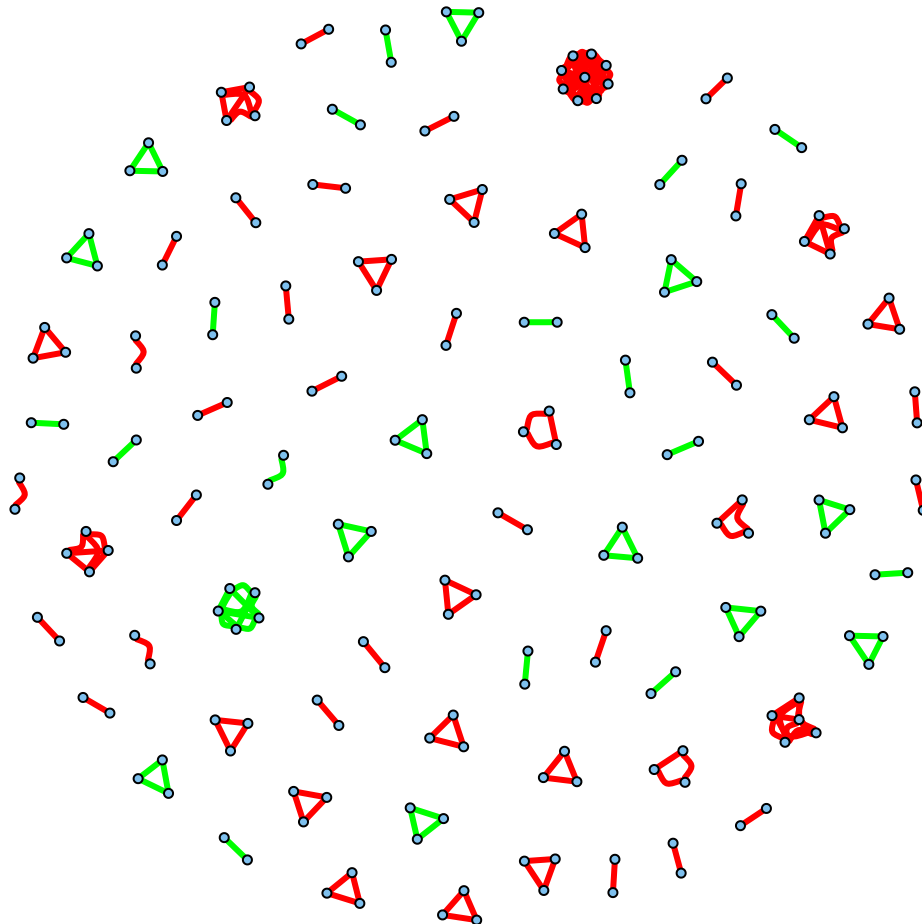
Figure 3: The same as Figure 2 with two added features: there are wavy and straight edges. The wavy edges indicate that SMERED and the NLTCS do not agree on the linkage, hence a false link. The straight edges indicate linkage agreement. We can see that there are a fair number of red, wavy edges indicating low probability and incorrect links. There are also many straight edged green high probability correct links. This illustrates one level of accuracy of the algorithm in the sense that it finds correct links and identifies incorrect ones. We can view the red links as individuals we would push to clerical review since the algorithms has trouble matching them (if this is warranted).

4.4   Estimation of Attributes of Observed Individuals from the Population

The number of observed unique individuals N is easily inferred from the posterior of $\mathbf{\Lambda}|\mathbf{X}$, since N is simply the number of unique values in $\mathbf{\Lambda}$. Defining $N|\mathbf{X}$ to be the posterior distribution of N, we can find this by applying a function to the posterior distribution on $\mathbf{\Lambda}$, as discussed in Section 3.4. (Specifically, $N = |\#\mathbf{\Lambda}|$, where $\#\mathbf{\Lambda}$ maps $\mathbf{\Lambda}$ to its set of unique entries, and $|A|$ is the cardinality of the set A.) Doing so, the posterior distribution of $N|\mathbf{X}$ is given in Figure 4. Also, $\hat{N} := E(N|\mathbf{X}) = 35,992$ with a posterior standard error of 19.08. The posterior median and mode are 35,993 and 35,982 respectively. Since the true number of observed unique individuals is 34,945, we are slightly undermatching, which leads to an overestimate of N. This phenomenon most likely occurs due to individuals migrating between states across the three different waves. It is difficult to improve this estimate since we do not have additional information as described above.
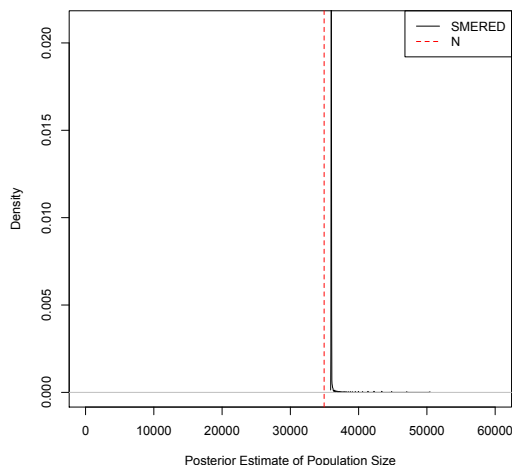

Figure 4: Posterior density of the number of observed unique individuals N (black) compared to the posterior mean (red).

We can also estimate attributes of sub-groups. For example, we can estimate the number of individuals within each wave or combination of waves—that is, the number of individuals with any given linkage pattern. (We summarize these estimates here with posterior expectations alone, but the full posterior distributions are easily computed.) Recall for each $j' = 1, \ldots, N$, $R_{ij'} = \{j : \lambda_{ij} = j'\}$. For example, the posterior expectation for the number of individuals appearing in lists

i_i and i_2 but not i_3 is approximately

$$\frac{1}{S_G} \sum_{h=1}^{S_G} \sum_{j'} I\left(\left|R_{i_1 j'}^{(h)}\right| = 1\right) I\left(\left|R_{i_2 j'}^{(h)}\right| = 1\right) I\left(\left|R_{i_3 j'}^{(h)}\right| = 0\right).$$

(The inner sum is a function of $\mathbf{\Lambda}^{(h)}$, but a complicated one to express without the R_{ij}.)

Table 5 reports the posterior means for the overlapping waves and each single wave of the NLTCS and compares this to the ground truth. In the first wave (1982), our estimates perform exceedingly well with relative error of 0.11%, however, as waves cross and we try to match people based on limited information, the relative errors range from 8% to 15%. This is not surprising, since as patients age, we expect their proxies to respond, making patient data more prone to errors. Also, older patients may move across states, creating further matching dilemmas. We are unaware of any alternative algorithm that does better using this data with only these fields available.

5 De-duplication

Our application of SMERE to the NLTCS assumes that each list had no duplicates, however, many other applications will contain duplicates within lists. We showed in Section 3.1 that we can theoretically handle de-duplication across and within lists. We apply SMERE with de-duplication (SMERED) to the NLTCS by (i) running SMERED on the three waves to show that the algorithm does not falsely detect duplicates when there really are none, and (ii) combining all the lists into one file, hence creating many duplicates, to show that SMERED can find them.

5.1 Application to NLTCS

We combine the three files of the NLTCS mentioned in Section 4 which contain 22,132 duplicate records out of 57,077 total records. We run SMERED on settings (i) and (ii), evaluating accuracy with the unique IDs. We compare our results to "ground truth" (Table 5). In the case of the NLTCS, compiling all three files together and running the three waves separately under SMERED yields similar results, since we match on similar covariate information. There is no covariate information to add to from thorough investigation to improve our results, except under simulation study.

22

When running SMERE for three files, the FNR is 0.11 and the FPR is 0.37. When running SMERED and estimating a single linkage structure by linking records in shared most probable maximal matching sets, the FNR is 0.11 and the FPR is 0.37. We contrast this with the results obtained when running SMERED under the shared most probable MMSs (MPMMS) for a single compiled file (Table 6), which yields an FNR of 0.10 and an FPR of 0.17. Clearly, SMERE produces the best results in terms of both FNR and FPR; however, if we want to consider the record linkage and de-duplication problem simultaneously, the SMERED algorithm with linkages applied through the shared MPMMS lowers the FPR nearly in half.

The dramatic increase in the FPR and number of false links shown in Table 5 is explained by how few field variables we match on. Their small number means that there are many records for different individuals that have identical or near-identical values. On examination, there are $2,558$ possible matches among "twins," records which agree exactly on all attributes but have different unique IDs. Moreover, there are 353,536 "near-twins," pairs of records that have different unique IDs but match on all but one attribute. This illustrates why the matching problem is so hard for the NLTCS and other data sources like it, where survey-responder information such as name and address are lacking. However, if it is known that each file contains no duplicates, there is no need to consider most of these twins and near-twins as possible matches.

We would like to put SMERED's error rates in perspective by comparing to another method, but we know of no other that simultaneously links and de-duplicates three or more files. Thus, we compare to the simple baseline of linking records when, and only when, they agree on all fields (cf. Fleming, King and Juda, 2007). See Table 5 for the relevant error rates. Recall that SMERED produces a FNR and FPR of 0.10 and 0.37. The baseline has an FPR of 0.09, much lower than ours, and an FNR of 0.09, which is the same. We attribute the comparatively good performance of the baseline to there being only five categorical fields per record. With more information per record, exact matches would be much rarer, and the baseline's FPR would shrink while its FNR would tend to 1. Furthermore, we extend the baseline to the idea of "near-twins." Under this, we find that the FPR and FNR are 12.61 and 0.05, where the FPR is orders of magnitude larger than ours, while the FNR is slightly lower. While the FPR is high under our model, it is much worse

for the baseline of "near-twins." Our methods tend to "lump" distinct latent individuals, but it is much less prone to lump than the baseline, at a minor cost in splitting.

We also consider a slightly modified version of our proposed methodology wherein we estimate the overall linkage structure by linking those records that are in any shared MMS with a posterior probability of at least 0.8. The resulting linkage structure is still guaranteed to be transitive, but it includes fewer links. This procedure attains a similar FNR (0.10) to the original methodology, but it has a substantially lower FPR (0.17).

Furthermore, we compare to the baseline using a range of thresholded FNRs and FPRs using the shared most probable MMSs. We apply thresholded values ranging from [0.2,1] to the shared most probable MMSs (and after thresholding calculating each FNR and FPR). This allows us to plot the tradeoff for FNR and FPR under under SMERE and SMERED as seen in Figure 5. SMERE and SMERED perform similarly compared to the baseline of exact matching, however, we note that either FPR or FNR can be changed for better performance to be gained. When the baseline changes to "near-twins," however, both our algorithms in terms of performance radically beat the "near-twins" baseline, showing the value of a model-based approach when there are distortions or noise in the data. We are not able to catch *all* of the effects of distortion; but, under a very simplistic and easily implemented model, where we match on very little information, we do well.

5.2 Application to Italian Household Survey

We apply our method to the Italian Survey on Household and Wealth (SHIW), a sample survey conducted by the Bank of Italy every two years. The 2010 survey covers 7,951 households composed of 19,836 individuals. The 2008 survey covers 19,907 individuals and 13,266 individuals. We test our methods on all twenty regions, merging the 2008 and 2010 survey. We consider the following categorical variables: year of birth, working status, employment status, branch of activity, town size, geographical area of birth, sex, whether or not Italian national, and highest educational level obtained. We compare our method to that of Tancredi and Liseo (2011) as it is quite natural to our approach. The approach of Tancredi and Liseo (2011) can be framed as a special case of our linkage structure as we describe below.
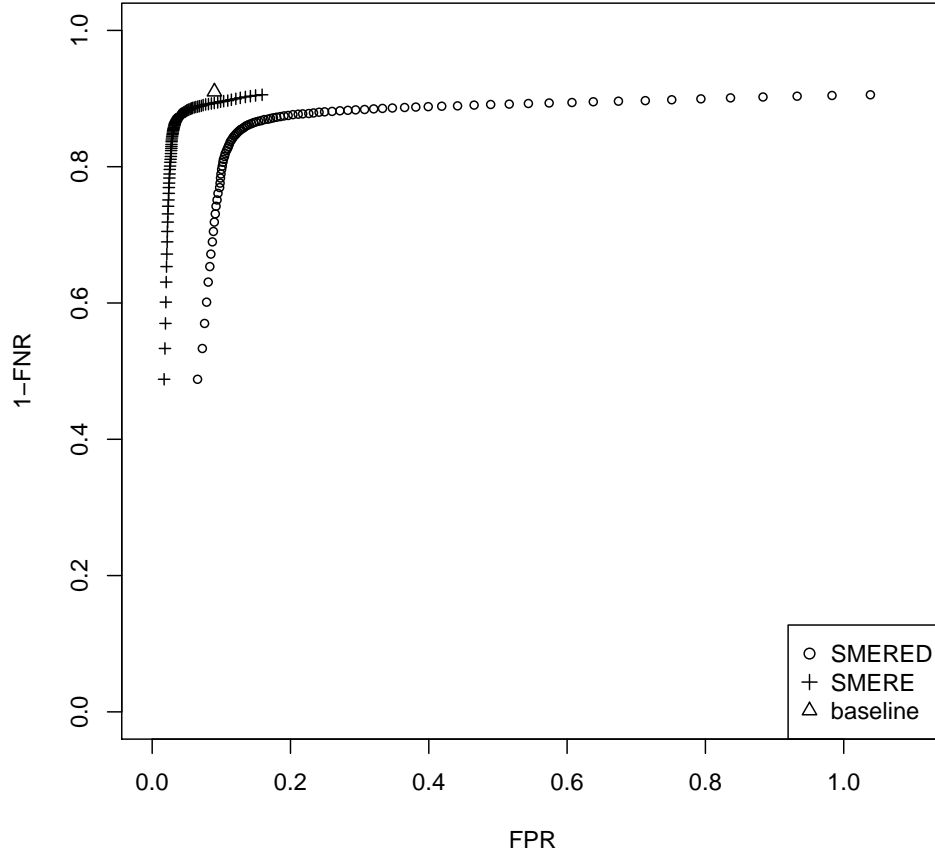
24

Figure 5: We plot both receiver operating characteristics (ROC) curves under SMERE and SMERED for the most probable MMSs (this is to avoid all to all comparisons) and compare them to the simple baseline (triangle). For the same FNR ($\approx$ the number of missing links), the FPR ($\approx$ the number of false links) is higher under SMERED than under SMERE. This is again due to problems in linkage using SMERED when the categorical information is very limited. When the FNR is small, performance is very similar under SMERE and SMERED. However, we note that the baseline can *never change*, whereas, under our algorithm we can relax the FPR or FNR for performance. The "near-twins" baseline does not appear on this plot since its FPR is 12.61.

Representing a partition as a matrix is not efficient as the number of records $N \to \infty$. An alternative is the coreference matrix (Matsakis, 2010; Sadinle, 2014; Tancredi and Liseo, 2011):

$$\boldsymbol{\Delta}_{qr} = \begin{cases} 1 & \text{if records } q \text{ and } r \text{ refer to same individual} \\ 0 & \text{otherwise.} \end{cases}$$

We make the relationship clear between that of the coreference matrix of Matsakis (2010); Sadinle (2014); Tancredi and Liseo (2011) and our proposed linkage structure clear.

Lemma: The conference matrix $\boldsymbol{\Delta}$ can be written as a function of $\boldsymbol{\Lambda}$. Hence, the linkage structure can be written or represented as a partition of records.

Proof: Let q and r be two records. Recall that the coreference matrix is defined as

$$\boldsymbol{\Delta}_{qr} = \begin{cases} 1 & \text{if records } q \text{ and } r \text{ refer to same individual} \\ 0 & \text{otherwise.} \end{cases}$$

In our notation, j refers to the individual in list i. Following this notation, for some record q, this corresponds to a list i_1 and records j_1 in our notation denoted by (i_1, j_1). Similarly, for record r, there is a corresponding list and record denoted by (i_2, j_2). Note: $\boldsymbol{\Lambda}_{i_1, j_1} = \boldsymbol{\Lambda}_{i_2, j_2}$ iff (i_1, j_1) and (i_2, j_2) refer to latent individual j' (which has an arbitrary indexing). This implies that $\boldsymbol{\Delta}_{qr} = I(\boldsymbol{\Lambda}_{i_1, j_1} = \boldsymbol{\Lambda}_{i_2, j_2})$. Hence, the coreference matrix can be written as a function of $\boldsymbol{\Lambda}$, and thus, $\boldsymbol{\Lambda}$ can be reordered such that it is a partition matrix.

Thus, we illustrate that for nearly every region in Italy for the SHIW, SMERED is superior in terms of FNR and FPR to that of Tancredi and Liseo (2011), something we attribute to the linkage structure, which is easily imbedded within the algorithm of Jain and Neal (2004) (See Table 3). Our method for every region takes 17 minutes, whereas the competitor approach takes 90 minutes.

| | Tancredi & Liseo | | SMERED | |
Region	FNR	FPR	FNR	FPR
1	1.51	0.80	0.37	0.14
2	0.13	0.18	0.34	0.02
3	0.44	0.50	0.38	0.15
4	0.45	0.47	0.62	0.19
5	1.26	0.75	0.50	0.15
6	0.54	0.58	0.36	0.08
7	0.70	0.58	0.42	0.10
8	1.52	0.81	0.43	0.17
9	1.30	0.81	0.56	0.22
10	0.73	0.55	0.39	0.15
11	0.97	0.63	0.54	0.18
12	0.86	0.79	0.40	0.12
13	0.27	0.41	0.40	0.12
14	0.33	0.31	0.46	0.07
15	1.11	0.89	0.56	0.25
16	0.86	0.53	0.51	0.17
17	0.29	0.34	0.42	0.08
18	0.34	0.30	0.36	0.08
19	1.18	0.71	0.50	0.16
20	0.97	0.54	0.45	0.11

Table 3: Method of SMERED versus Tancredi and Liseo (2011).

6 A User's Guide to Record Linkage

In this paper, we have proposed a new method and algorithm for simultaneous record linkage and de-duplication. We have focused on two applications, where anonymization occurs, and hence record linkage and de-duplication is useful for such methods.

We have assumed that (1) the lists are conditionally independent, (2) the data are categorical, and (3) the records have a minimal set overlap of fields that can be matched. We then matched based on our proposed method of MPMMS. We also assumed that blocks are independent and that the distortion probabilities *do not* depend on the list. Moreover, we have used a "non-infomative" prior on the linkage structure, and we show below that this leads to interesting discoveries and perhaps new research directions in terms of finding subjective priors. As a user and researcher, one might ask, how valid are these assumptions and what directions are left for exploration?

For applications such as those illustrated here (based on data from the NLTCS and the SHIW),

where the data have been stripped of "personal identifying information" for confidentiality reasons, the categorical assumption is in fact valid and very reasonable. For other applications, this assumption is at best an approximation and thus introduces potential errors. Exploring dependence among lists goes beyond the scope of the present paper and clearly will be context dependent, varying from one problem to the next. The work of Steorts (2015) explores the dependence between lists in an empirically motivated prior setting, showing improvements in the error rates when such dependencies are present. Moreover, the applications we have encountered have a common minimal set of overlapping fields (such as date of birth, location, etc). Situations where some subsets of lists have additional overlap require further study. These cases and those where there are *no overlapping fields* can be viewed as missing data problems, e.g., see Zhang, Shih, Müller et al. (2007), which proposes a spatial extension of Bayesian Adaptive Regression Trees in a record linkage context.

Matching to the MPMMS is a good procedure in the sense that it's optimal for squared error loss and the MPMMS preserves transitivity. Thresholding based on this matching criteria is useful in practice since the MPMSS provides a principled way of accounting for the exact uncertainty of transitive matches. It then allows for an automated process to created an updated list of records that match above a posterior probability and push the rest to clerical review.

Blocking is necessary for dimension reduction and scalability of record linkage models. How to properly block at dataset remains an important issue in record linkage and should be evaluated by the false negative rate. Comparisons of blocking methods have been extensively explored in Steorts et al. (2014), which includes approaches such as breaking a string into characters, e.g., locality sensitive hashing and random projections.

The area that needs the most guidance moving forward is in choosing subjective priors from domain knowledge that are also robust to model-misspecification. We speak to the current limitations of our model and a similar model of Sadinle (2014, 2015), which we made connections with earlier.

Assigning Prior Probabilities to Partitions

Prior probabilities are assigned using $\mathbf{\Lambda}$ such that each record is assumed to be equally likely a priori to correspond to any of the N latent individuals. Thus, it treats the records as if they are a random sample drawn *with* replacement. (It is actually perfectly natural to take each of the N^N possible values of $\mathbf{\Lambda}$ to be equally likely a priori). We now translate this concept to priors on partitions, which requires answering how many possible values of $\mathbf{\Lambda}$ correspond to each partition ξ. (Under this translation, we can view both our prior and that of Sadinle (2014) as a prior on partitions).

A partition ξ splits the records up over $|\xi|$ latent individuals, where each of these $|\xi|$ latent individuals could have been assigned to one of N index numbers, with the catch that different latent individuals must have different index numbers. Thus, there are $N(N-1)\cdots(N-|\xi|+1) = N!/(N-|\xi|)!$ different $\mathbf{\Lambda}$ values that yield the partition ξ. Since each $\mathbf{\Lambda}$ value was assigned prior probability $1/N^N$, the prior probability of the partition ξ is

$$\pi(\xi) = \frac{N!}{(N-|\xi|)!\, N^N}.$$

Remark: any two partitions with the same number of latent entities are equally probably a priori.

A Generalization of the Probabilistic Structure

Our prior is constructed by assuming that the N records are randomly sampled with replacement from a population of N latent individuals. However, why should we assume that the "population" of latent individuals has the same size as the sample when considering such a mechanism? Since no more than N latent individuals can be represented in the sample, but there there *is* a difference between having a latent "population" of size N versus something larger than N—the larger the latent "population" is, the less likely it is for records to be linked. Instead, let the latent "population" have size M. There are now M^N possible values of $\mathbf{\Lambda}$, all of them equally likely. There are $M(M-1)\cdots(M-|\xi|+1) = M!/(M-|\xi|)!$ different $\mathbf{\Lambda}$ values that yield the partition ξ. Then the

prior probability of the partition ξ is

$$
\pi(\xi) = \begin{cases} \dfrac{M!}{(M - |\xi|)! \, M^N} & \text{if } |\xi| \leq M, \\[2ex] 0 & \text{if } |\xi| > M, \end{cases}
$$

noting the impossibility of any partition with more sets ($|\xi|$) then there are individuals in the latent "population" (M).

The Role of M

The latent "population" size M is a tuning parameter that can be altered to control the overall linking tendency. Suppose $M \to \infty$. Let $\xi^\star = \{\{1\}, \{2\}, \ldots, \{N\}\}$ denote the partition in which no records at all are linked, and note that $|\xi^\star| = N$. Then

$$
\pi(\xi^\star) = \frac{M(M-1)\cdots(M-N+1)}{M^N} = \left(\frac{M}{M}\right)\left(\frac{M-1}{M}\right)\cdots\left(\frac{M-N+1}{M}\right) \to 1
$$

as $M \to \infty$. If we sample from an essentially infinite population, then our sample is almost guaranteed to consist of N distinct individuals. However, this is clearly not what we want to do for record linkage. On the other end of the spectrum, setting $M < N$ immediately excludes certain linkage structures from consideration. Specifically, taking $M = N - r$ for some positive integer r implies that all linkage structures with less than r links have prior probability zero.

Choosing M via the Prior Mean

A guideline for choosing M could possibly be determined by looking at the prior expectation of $|\xi|$. Note that $|\xi|$ represents the number of unique individuals in the sample, so it's a quantity about which we might have some intuition in a particular problem. Let $\mu(M, N)$ denote this prior mean, i.e., let $\mu(M, N)$ denote the expected cardinality of a random partition ξ of N elements, with the

probability of each partition equal to

$$
\pi(\xi) = \begin{cases} \dfrac{M!}{(M - |\xi|)!\, M^N} & \text{if } |\xi| \le M, \\[2mm] 0 & \text{if } |\xi| > M. \end{cases}
$$

Note that

$$
\mu(M, N) = \sum_{j=0}^{N-1} \left(\frac{M-1}{M} \right)^j = M \left[1 - \left(\frac{M-1}{M} \right)^N \right] = M \left\{ 1 - \left[\left(1 - \frac{1}{M} \right)^M \right]^{N/M} \right\}
$$
$$
\approx M \left[1 - \left(e^{-1} \right)^{N/M} \right]
$$
$$
= M \left[1 - \exp\left(-\frac{N}{M} \right) \right]
$$

as long as M is fairly large. It is not possible to solve for M in terms of $\mu(M, N)$. However, $\mu(M, N)$ is a strictly increasing function of M, so we can pick the value of M numerically to yield a desired prior mean $\mu(M, N)$. This may be a direction of future research in terms of looking at subjective prior's and partitions. We now illustrate that our prior along with that of Sadinle (2014) is highly informative.

6.1 Highly Informative Priors

Any particular value of M leads to a prior that is still *highly informative* about the overall amount of linkage. Note that one way to measure the overall amount of linkage is to count the number of individuals represented in the sample. Both our prior as described above and that of Sadinle (2014) prior are highly informative about this quantity. For a small sample of size $N = 25$, The plot below shows our prior with $M = 50$ (blue) and that of Sadinle (2014) prior (red).

Changing M for our prior would shift our distribution to the left or right, but the *location* is not the problem. The problem is that *the tails of these priors are much too light*. Moreover, the larger the sample size, the more dramatic this problem will be. In other words, both priors are highly informative about the overall amount of linkage. Figure 6 illustrates that our prior is preferred over that of Sadinle (2014) since the choice of M allows us some control over the situation. This gives
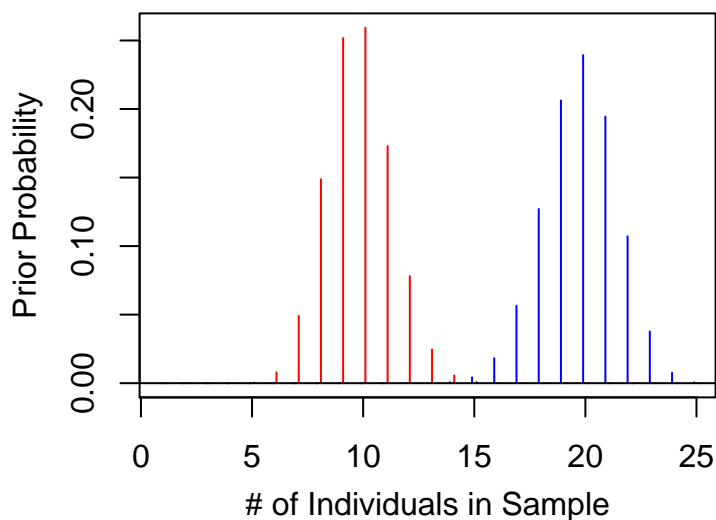
Figure 6: Illustrating that when $M = 50$ and $N = 25$ both the priors of our paper and Sadinle (2014) are highly informative.

insight at least in the situation of "non-informative" priors of this type and illustrating how very informative they are! More importantly, this very much drives home the point that we need well principled subjective priors since the light tails we see here mean that the data will not be able to overwhelm the prior if we do not choose M properly (which, in fact, defeats the purpose of being a Bayesian).

Hence, it is up to the user to decide whether or not this model and evaluations is appropriate for the data at hand (and to do proper testing as we have suggested) Finally, as we have outlined there is much work left to be done in record linkage and this is a first step at advancing the methodology and understanding the most important steps for moving forward on what is a very important topic both in methods, theory, computer science, and applications.

7   Discussion

We have made three contributions in this paper. The first and more general one is to frame record linkage and de-duplication simultaneously, namely linking observed records to latent individuals and representing the linkage structure via $\mathbf{\Lambda}$. The second contribution is our specific parametric

Bayesian model, which, combined with the linkage structure, allows for efficient inference and exact error rate calculation (such as our most probable MMS and associated posterior probabilities). Moreover, this allows for easy integration with capture-recapture methods, where error propagation is *exact.* Third, we have suggested practical guidance to practitioners for doing record linkage using our proposed methods, outlining its strengths and its shortcomings. As with any parametric model, its assumptions only apply to certain problems; however, this work serves as a starting point for more elaborate models, e.g., incorporating missing fields, data fusion, complicated string fields, population heterogeneity, or dependence across fields, across time, or across individuals. Within the Bayesian paradigm, such model expansions will lead to larger parameter spaces, and therefore call for computational speed-ups, perhaps via online learning, variational inference, or approximate Bayesian computation.

Our work serves as a first basis for solving record linkage and de-duplication problems simultaneously using a noisy Bayesian model and a linkage structure that can handle large-scale databases. We hope that our approach will encourage the emergence of new record linkage approaches and applications along with more state-of-the-art algorithms for this kind of high-dimensional data.

sets of records	1.10084	3.5583; 1.10084	3.5583; 1.10084; 2.6131
posterior probability	0.001	0.004	0.995

Table 4: Example of posterior matching probabilities for record 10084 in 1982

	82	89	94	82, 89	89, 94	82, 94	82, 89, 94
NLTCS (ground truth)	7955	2959	7572	4464	3929	1511	6114
Bayes Estimates$_{\text{SMERE}}$	7964.0	3434.1	8937.8	4116.9	4502.1	1632.2	5413.0
Bayes Estimates$_{\text{SMERED}}$	7394.7	3009.9	6850.4	4247.5	3902.7	1478.7	5191.2
Relative Errors$_{\text{SMERE}}$ (%)	0.11	16.06	18.04	-7.78	14.59	8.02	-11.47
Relative Errors$_{\text{SMERED}}$ (%)	-7.04	1.72	-9.53	-4.85	-0.67	-2.14	-15.09

Table 5: Comparing NLTCS (ground truth) to the Bayes estimates of matches for SMERE and SMERED

	False links	True Links	Missing Links	FNR	FPR
NLTCS (ground truth)	0	28246	0	0	0
Bayes Estimates$_{\text{SMERE}}$	1299	25196	3050	0.11	0.05
Bayes Estimates$_{\text{SMERED}}$	10595	24900	3346	0.09	0.37
MPMMSs$_{\text{SMERED}}$	4819	25489	2757	0.10	0.17
Exact matching	2558	25666	2580	0.09	0.09
Near-twins matching	356094	26936	1310	0.05	12.61

Table 6: False, True, and Missing Links for NLTCS under blocking sex and DOB year where the Bayes estimates are calculated in the absence of duplicates per file and when duplicates are present (when combining all three waves). Also, reported FNR and FPR for NLTCS, Bayes estimates.

References

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.

BHATTACHARYA, I. and GETOOR, L. (2006). A latent dirichlet model for unsupervised entity resolution. In *SDM*, vol. 5. SIAM, 59.

CHRISTEN, P. (2012). A survey of indexing techniques for scalable record linkage and deduplication. *IEEE Transactions on Knowledge and Data Engineering*, **24**.

COPAS, J. and HILTON, F. (1990). Record linkage: Statistical models for matching computer records. *Journal of the Royal Statistical Society, Series A*, **153** 287–320.

DAI, A. M. and STORKEY, A. J. (2011). The grouped author-topic model for unsupervised entity resolution. In *Artificial Neural Networks and Machine Learning–ICANN 2011*. Springer, 241–249.

DOMINGOS, P. and DOMINGOS, P. (2004). Multi-relational record linkage. In *Proceedings of the KDD-2004 Workshop on Multi-Relational Data Mining*. ACM.

FELLEGI, I. and SUNTER, A. (1969). A theory for record linkage. *Journal of the American Statistical Association*, **64** 1183–1210.

FIENBERG, S., MAKOV, U. and SANIL, A. (1997). A Bayesian Approach to Data Disclosure: Optimal Intruder Behavior for Continuous Data . In *Privacy in Statistical Databases*, vol. 13. 75–89.

FLEMING, L., KING, C., III and JUDA, A. (2007). Small worlds and regional innovation. *Organization Science*, **18** 938–954.

GUTMAN, R., AFENDULIS, C. and ZASLAVSKY, A. (2013). A bayesian procedure for file linking to analyze end- of-life medical costs. *Journal of the American Statistical Association*, **108** 34–47.

HALL, R. and FIENBERG, S. (2012). Valid statistical inference on automatically matched files. In *Privacy in Statistical Databases 2012* (J. Domingo-Ferrer and I. Tinnirello, eds.), vol. 7556 of *Lecture Notes in Computer Science*. Springer, Berlin, 131–142.

HERZOG, T., SCHEUREN, F. and WINKLER, W. (2007). *Data Quality and Record Linkage Techniques.* Springer, New York.

JAIN, S. and NEAL, R. (2004). A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model. *Journal of Computational and Graphical Statistics*, **13** 158–182.

LAHIRI, P. and LARSEN, M. (2005). Regression analysis with linked data. *Journal of the American Statistical Association*, **100** 222–230.

LARSEN, M. D. and RUBIN, D. B. (2001). Iterative automated record linkage using mixture models. *Journal of the American Statistical Association*, **96** 32–41.

LISEO, B. and TANCREDI, A. (2013). Some advances on Bayesian record linkage and inference for linked data. URL http://www.ine.es/e/essnetdi_ws2011/ppts/Liseo_Tancredi.pdf.

MATSAKIS, N. E. (2010). *Active Duplicate Detection with Bayesian Nonparametric Models.* Ph.D. thesis, Massachusetts Institute of Technology.

REITER, J. P. and RAGHUNATHAN, T. E. (2007). The multiple adaptations of multiple imputation. *Journal of the American Statistical Association*, **102** 1462–1471.

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

SADINLE, M. (2015). *A Bayesian Partitioning Approach to Duplicate Detection and Record Linkage.* Ph.D. thesis, Carnegie Mellon University.

SADINLE, M. and FIENBERG, S. (2013). A generalized Fellegi-Sunter framework for multiple record linkage with application to homicide record-systems. *Journal of the American Statistical Association*, **108** 385–397.

STEORTS, R., VENTURA, S., SADINLE, M. and FIENBERG, S. (2014). A Comparison of Blocking Methods for Record Linkage . In *Privacy in Statistical Databases.* Springer, 253–268.

STEORTS, R. C. (2015). Entity resolution with empirically motivated priors. *Bayesian Analysis*, **10** 849–875.

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

Winkler, W. (1999). The state of record linkage and current research problems. Technical report, Statistical Research Division, U.S. Bureau of the Census.

Winkler, W. (2000). Machine learning, information retrieval, and record linkage. American Statistical Association, Proceedings of the Section on Survey Research Methods, 20–29. URL http://www.niss.org/affiliates/dqworkshop/papers/winkler.pdf.

Zhang, S., Shih, Y.-C. T., Müller, P. et al. (2007). A spatially-adjusted bayesian additive regression tree model to merge two datasets. *Bayesian Analysis*, **2** 611–633.

A Motivating Example of Linkage Structure and Distortion

We now present a simple example of the ideas of distortion and linkage, which illustrates the relationships between the observed data $\boldsymbol{X}$, the latent individuals $\boldsymbol{y}$, the linkage structure $\boldsymbol{\Lambda}$, and the distortion indicators $\boldsymbol{z}$. Suppose the "population" (individuals represented in at least one list) has four members, where name and address are stripped for anonymity and they are listed by state, age, and sex. For instance, the latent individual vector $\boldsymbol{y}$ might be

$$
\boldsymbol{y} = \begin{bmatrix} \text{NC, 72, F} \\ \text{SC, 73, F} \\ \text{PA, 91, M} \\ \text{VA, 94, M} \end{bmatrix}.
$$

The observed records $\boldsymbol{X}$ are given in three separate lists, which would combine into a three-dimensional array. We write this here as three two-dimensional arrays for notational simplicity:

$$
\text{List 1} = \begin{bmatrix} \text{NC, 72, F} \\ \text{SC, 70, F} \\ \text{PA, 91, M} \end{bmatrix}, \quad
\text{List 2} = \begin{bmatrix} \text{SC, 37 , F} \\ \text{VA, 93, M} \\ \text{PA, 92, M} \end{bmatrix}, \quad
\text{List 3} = \begin{bmatrix} \text{NC, 72 , F} \\ \text{NC, 72, F} \\ \text{SC, 72, F} \\ \text{VA, 94, M} \end{bmatrix}
$$

Here, for the sake of keeping the illustration simple, only age is distorted.

Comparing $\boldsymbol{X}$ to $\boldsymbol{y}$, the intended linkage and distortions are then

$$
\boldsymbol{\Lambda} = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 3 \\ 1 & 1 & 2 & 4 \end{bmatrix}, \quad
\boldsymbol{z_1} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad
\boldsymbol{z_2} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \quad
\boldsymbol{z_3} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}.
$$

In this linkage structure, every entry of $\boldsymbol{\Lambda}$ with a value of 2 means that some record from $\boldsymbol{X}$ refers to the latent individual with attributes "SC, 73, F." Here, the age of this individual is distorted in

all three lists, as can be seen from $\boldsymbol{z}$. (Note that $\boldsymbol{z}$, like $\boldsymbol{X}$, is also really a three-dimensional array.) Looking at $\boldsymbol{z}_1$ and $\boldsymbol{z}_3$, we see that there is only a single record in either list that is distorted, and it is only distorted in one field. In list 2, however, every record is distorted, though only in one field.

Figure 7 illustrates the interpretation of our linkage structure as a bipartite graph in which each edge links a record to a latent individual. For clarity, Figure 7 shows that X_{11} and X_{22} are the same individual and shows that X_{13}, X_{21}, and X_{34} correspond to the same individual. The rest are non-matches.
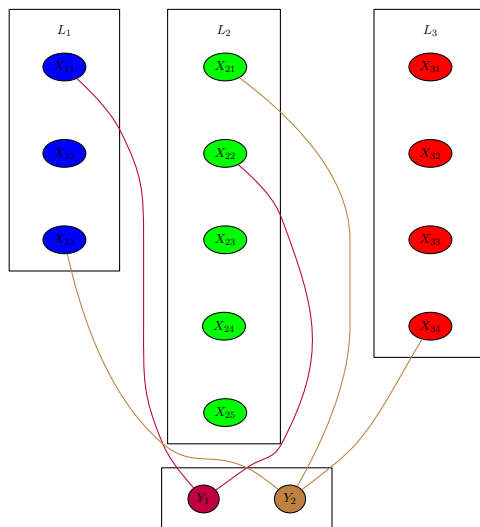

Figure 7: Illustration of records $\boldsymbol{X}$, latent random variables $\boldsymbol{Y}$, and linkage (by edges) $\boldsymbol{\Lambda}$.

B Hybrid MCMC Algorithm (SMERED)

We now describe in more detail the Metropolis-within-Gibbs algorithm with split-merge proposals and optional record linkage blocking (with pseudo-code given at the end). The entire loop below is repeated for a number of MCMC iterations $S_H = S_G \times S_M$. Additionally, the algorithm allows multiple split-merge operations to be performed in a single Metropolis-Hastings proposal step. Let T denote the allowed number of split-merge operations within each Metropolis-Hastings step. Let $\boldsymbol{\Lambda}^{(m)}, \boldsymbol{y}^{(m)}, \boldsymbol{z}^{(m)}, \boldsymbol{\theta}^{(m)}, \boldsymbol{\beta}^{(m)}$ denote the values of the MCMC chain at step m.

1. Repeat the following sequence of steps T times:

(a) As already described in Section 3.2, we sample pairs of records from different files uniformly at random within blocks.

(b) If the two records chosen above in (a) are currently assigned to the same latent individual, we propose to split them as follows:

    i. Let j' denote the latent individual to which both records are currently assigned.

    ii. Let C denote the set of all *other* records—not including the two chosen in step (a) above—who are also assigned to latent individual j'. Note that this set C may be empty.

    iii. Give the two records new assignments of latent individuals, calling these new assignments j_1 and j_2. One of the two individuals stays assigned to j', while the other is assigned to a latent individual currently not assigned to any records.

    iv. Randomly assign all the other records in C to either j_1 or j_2, which partitions C into sets C_1 and C_2. The inclusion of this step is important as the algorithm does not actually actually split or merge *records*—it actually splits or merges *latent individuals*. Note that the sets C_1 and C_2 are designed to *include* the two records we chose in step (a).

    v. The latent individuals j_1 and j_2 get their values $\boldsymbol{y}'_{j_1}$ and $\boldsymbol{y}'_{j_2}$ assigned by simply taking them to be *equal* (without distortion) to the exact record values for one of the individuals in the sets C_1 and C_2 (respectively), chosen at random.

    vi. For each record in C_1 and C_2, the corresponding distortion indicators $z'_{ij\ell}$ for each field are resampled from their respective conditional distributions. Note that some of these may be guaranteed to automatically be 1 (whenever a record differs from its corresponding latent individual on a particular field).

    vii. The above steps generate new proposals $\boldsymbol{\Lambda}'$, $\boldsymbol{y}'$, and $\boldsymbol{z}'$. We now decide whether to accept or reject the proposal according to the Metropolis acceptance probability. If we accept the proposal, then we take $\boldsymbol{\Lambda}^{(m+1)} = \boldsymbol{\Lambda}'$, $\boldsymbol{y}^{(m+1)} = \boldsymbol{y}'$, and $\boldsymbol{z}^{(m+1)} = \boldsymbol{z}'$. If we reject the proposal, then we take $\boldsymbol{\Lambda}^{(m+1)} = \boldsymbol{\Lambda}^{(m)}$, $\boldsymbol{y}^{(m+1)} = \boldsymbol{y}^{(m)}$, and $\boldsymbol{z}^{(m+1)} = \boldsymbol{z}^{(m)}$.

Algorithm 1: Split and MErge REcord linkage and Deduplication (SMERED)

Data: $\boldsymbol{X}$ and hyperparameters

Initialize the unknown parameters $\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{y}, \boldsymbol{z}$, and $\boldsymbol{\Lambda}$.

for $i \leftarrow 1$ **to** S_G **do**
   for $j \leftarrow 1$ **to** S_M **do**
      for $t \leftarrow 1$ **to** S_T **do**
         Draw records R_1 and R_2 uniformly and independently at random.
         if R_1 *and* R_2 *refer to the same individual* **then**
            propose splitting that individual, shifting $\boldsymbol{\Lambda}$ to $\boldsymbol{\Lambda}'$
            $r \leftarrow \min\{1, \pi(\boldsymbol{\Lambda}', \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\beta}|\boldsymbol{x})/\pi(\boldsymbol{\Lambda}, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\beta}|\boldsymbol{x})\}$
         endif
         else
            propose merging the individuals R_1 and R_2 refer to, shifting $\boldsymbol{\Lambda}$ to $\boldsymbol{\Lambda}'$
            $r \leftarrow \min\{1, \pi(\boldsymbol{\Lambda}', \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\beta}|\boldsymbol{x})/\pi(\boldsymbol{\Lambda}, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\beta}|\boldsymbol{x})\}$
         endif
         Resampling $\boldsymbol{\Lambda}$ by accepting proposal with Metropolis probability r or rejecting
         with probability $1 - r$.
      end
      Resample $\boldsymbol{y}$ and $\boldsymbol{z}$.
   end
   Resample $\boldsymbol{\theta}, \boldsymbol{\beta}$.
end

return $\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{\beta}\boldsymbol{X}, \boldsymbol{y}|\boldsymbol{X}, \boldsymbol{z}|\boldsymbol{X}$, *and* $\boldsymbol{\Lambda}|\boldsymbol{X}$.

(c) If instead the two records chosen above in (a) are currently assigned to different latent individuals, we propose to merge them. The same basic steps happen: a new state is created in which all records which belong to the same individual as either input record are all merged into a new individual. The fields for this individual are sampled uniformly from these records, distortion variables are all re-sampled, and then acceptance probability is tested.

2. Finally, new values $\boldsymbol{\theta}^{(m+1)}$ and $\boldsymbol{\beta}^{(m+1)}$ are drawn from their distributions conditional on the values of $\boldsymbol{\Lambda}^{(m+1)}$, $\boldsymbol{y}^{(m+1)}$, and $\boldsymbol{z}^{(m+1)}$ that we just selected:

$$\boldsymbol{\theta}^{(m+1)}, \beta^{(m+1)} \overset{\text{draw}}{\sim} \pi(\boldsymbol{\theta}, \boldsymbol{\beta} \mid \boldsymbol{y}^{(m)}, \boldsymbol{z}^{(m)}, \boldsymbol{\Lambda}^{(m)}, \boldsymbol{x}).$$

B.1 Time Complexity

Scalability is crucial to any record linkage algorithm. Current approaches typically run in polynomial (but super-linear) time in $N_{\max}$. (The method of Sadinle and Fienberg (2013) is $O(N_{\max}^k)$, while that of Domingos and Domingos (2004) finds the maximum flow in an $N_{\max}$-node graph, which is $O(N_{\max}^3)$, but independent of k.) In contrast, our algorithm is linear in both $N_{\max}$ and MCMC iterations.

Our running time is proportional to the number of Gibbs iterations S_G, so we focus on the time taken by one Gibbs step. Recall the notation from Section 3, and define $M = \frac{1}{p}\sum_{\ell=1}^p M_\ell$ as the average number of possible values per field ($M \geq 1$). The time taken by a Gibbs step is dominated by sampling from the conditional distributions. Specifically, sampling $\boldsymbol{\beta}$ and $\boldsymbol{y}$ are both $O(pN_{\max})$; sampling $\boldsymbol{\theta}$ is $O(pMN) + O(pN_{\max}) = O(pMN)$, as is sampling $\boldsymbol{z}$. Sampling $\boldsymbol{\Lambda}$ is $O(pN_{\max}M)$ if done carefully. Thus, all these samples can be drawn in time linear in $N_{\max}$.

Since there are S_M Metropolis steps within each Gibbs step and each Metropolis step updates $\boldsymbol{y}$, $\boldsymbol{z}$, and $\boldsymbol{\Lambda}$, the time needed for the Metropolis part of one Gibbs step is $O(S_M pN_{\max}) + O(S_M pMN) + O(S_M pN_{\max}M)$. Since $N \leq N_{\max}$, the run time becomes $O(pS_M N_{\max}) + O(MpS_M N_{\max}) = O(MpS_M N_{\max})$. On the other hand, the updates for $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ occur once each Gibbs step implying the run time is $O(pMN) + O(pN_{\max})$. Since $N \leq N_{\max}$, the run time becomes $O(pMN_{\max} + pN_{\max}) = O(pMN_{\max})$. The overall run time of a Gibbs step is $O(pMN_{\max}S_M) + O(pMN_{\max}) = O(pMN_{\max}S_M)$. Furthermore, for S_G iterations of the Gibbs sampler, the algorithm is order $O(pMN_{\max}S_GS_M)$. If p and M are all much less than $N_{\max}$, we find that the run time is $O(N_{\max}S_GS_M)$.

Another important consideration is the number of MCMC steps needed to produce Gibbs samples that form an adequate approximation of the true posterior. This issue depends on the convergence properties of the hybrid Markov chain used by the algorithm, which are beyond the scope of the present work.

C   Simulation Study

We provide a simulation study based on the model in Section 3.1, and we simulate data from the NLTCS based on our model, with varying levels of distortion. The varying levels of distortion (0, 0.25%, 0.5%, 1%, 2%, 5%) associated with the simulated data are then run using our MCMC algorithm to assess how well we can match under "noisy data." Figure 8 illustrates an approximate linear relationship with FPR (plusses) and the distortion level, while for FNR (triangles) exhibits a sudden large increase as the distortion level moves from 2% to 5%. Figure 9 demonstrates that for moderate distortion levels (per field), we can estimate the true number of observed individuals extremely well via estimated posterior densities. However, once the distortion is too *noisy*, our model has trouble recovering this value.

In summary, as records become more noisy or distorted, our matching algorithm typically matches less than 80% of the individuals. Furthermore, once the distortion is around 5%, we can only hope to recover approximately 65% of the individuals. Nevertheless, this degree of accuracy is in fact quite encouraging given the noise inherent in the data and given the relative lack of identifying variables on which to base the matching.
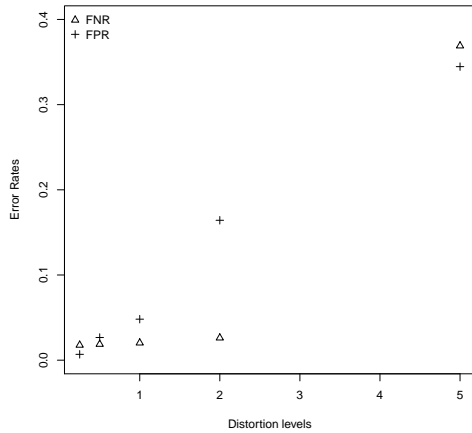
Figure 8: FNR and FPR versus distortion percentage. FPR shows an approximately linear relationship with distortion percentage, while FNR exhibits a sudden large increase as the distortion level moves from 2% to 5%.
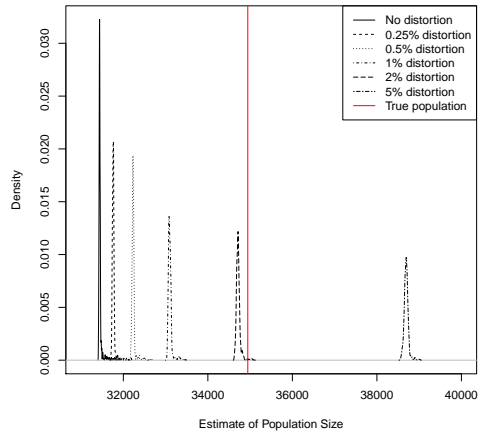

Figure 9: Posterior density estimates for 6 levels of distortion (none, 0.25%, 0.5%, 1%, 2%, and 5%) compared to ground truth (in red). As distortion increases (and approaches 2% per field), we overmatch N, however as distortion quickly increases to high levels (5% per field), the model undermatches. This behavior is expected to increase for higher levels of distortion. The simulated data illustrates that under our model, we are able to capture the idea of moderate distortion (per field) extremely well.

D   Confusion Matrix for NLTCS

Est vs Truth	82	89	82,89	94	82, 94	89, 94	AY	RS
82	8051.9	0.0	385.1	0.0	162.9	0.0	338.6	8938.5
89	0.0	2768.4	291.1	0.0	0.0	240.6	131.7	341.8
82, 89	118.4	2.2	8071.7	0.0	4.4	0.4	803.2	9000.3
94	0.0	0.0	0.0	7255.4	139.3	240.5	325.12	7960.32
82, 94	163.1	0.0	9.5	97.0	2662.2	0.09	331.5	3263.39
89, 94	0.0	186.8	6.1	190.6	1.5	7365.8	488.2	8239
AY	62.5	1.6	164.4	28.9	51.7	10.6	15923.7	18342.02
NLTCS	8396	2959	4464	7572	1511	3929	6114	

Table 7: Confusion Matrix for NLTCS

Est vs Truth	82	89	82,89	94	82, 94	89, 94	AY
82	0.9600	0.00000	0.04300	0.0000	0.0540	0.0000	0.0180
89	0.0000	0.94000	0.03300	0.0000	0.0000	3.1e-02	0.0072
82, 89	0.0140	0.00074	0.90000	0.0000	0.0015	5.1e-05	0.0440
94	0.0000	0.00000	0.00000	0.9600	0.0460	3.1e-02	0.0180
82,94	0.0190	0.00000	0.00110	0.0130	0.8800	1.1e-05	0.0180
89,94	0.0000	0.06300	0.00068	0.0250	0.0005	9.4e-01	0.0270
AY	0.0074	0.00054	0.01800	0.0038	0.0170	1.3e-03	0.8700

Table 8: Misclassification errors of confusion matrix for NLTCS