

# Bayesian Matching with Relaxed Duplication Assumptions through Dirichlet Record Linkage

Brian Kunding

September 13, 2024

## Abstract

Probabilistic record linkage is the use of statistical methods to identify unique entities within and across databases. Bayesian methods can adopt different prior distributions for various assumptions on the nature of record duplications in the linkage task, but become computationally expensive as the size of the linkage task grows. In this paper, we propose Dirichlet Record Linkage (DRL), a method for linking a duplicate-free reference file to a target file that may have internal duplications. We use a Dirichlet process prior for the parameter governing the number of matches a record in the reference file has in the target file. We use the sequence of parameters in stick breaking representation of the Dirichlet process to produce a computationally efficient Gibbs sampler to conduct the linkage task. We demonstrate the speed and accuracy of our approach relative to other recent Bayesian methods through simulations and case studies. In particular, we show that DRL exhibits strong performance using default, non-informative hyperparameters.

(Beka commented that I should mention the loss function in the abstract and introduction. To me, the loss function is not innovative enough to really call attention to it; it is used implicitly in many classification tasks. Thoughts?)

# 1 INTRODUCTION

Probabilistic record linkage uses statistical methods to identify matching records within and across databases, generally without the use of unique identifiers. This is an increasingly important task in “data cleaning,” either for its own sake, or as a preliminary step before conducting subsequent analysis. This problem is increasingly important in health care (Gutman et al. 2013, Hof et al. 2017), governmental statistics (Jaro 1995, Fortini et al. 2001), human rights statistics (Lum et al. 2013, Price et al. 2015, Sadosky et al. 2015, Sadinle 2014, 2018), author name disambiguation (Lai et al. 2011, Zhang et al. 2018), and forensic science (Tai 2018). In this paper, we propose a method for conducting record linkage between a duplicate free reference file and a target file that may have duplications within itself.

Many probabilistic record linkage methods are based on the foundational work of Fellegi & Sunter (1969). In these Fellegi-Sunter (FS) methods, the data analyst first creates comparison vectors for each pair of records in the data files. These vectors indicate how similar the records are on a set of variables measured in both files, known as the linkage variables. Using these comparison vectors, the analyst classifies each pair as a match or non-match. This approach has been extended for a wide variety of applications (Fair 2004, Winkler 1999, Wagner et al. 2014, Gill & Goldacre 2003). An alternative paradigm is to model the linkage variables directly (e.g., Tancredi & Liseo 2011, Steorts et al. 2016, Marchant et al. 2021, Betancourt et al. 2022). In this article, we build on the contributions to the comparison vector approach.

In the original FS formulation, the matching status of each record pair is considered independent from the matching status of all other record pairs. This can be suboptimal when we know that each file is internally free of duplications or when we desire a set of declared matches that respects transitivity. These concerns can be addressed through a Bayesian framework through different prior distributions on the linkage structure to formalize different assumptions on the type of matchings between files. For example, Sadinle (2017) proposed a prior distribution on the one-to-one matchings between two files, and Aleshin-Guendel & Sadinle (2023) proposed a prior on the partitions of records corresponding to unique entities found in arbitrarily many files. These more tailored record linkage methods have

the advantage of producing coherent sets of matches that respect transitivity without the need for post-processing, and are generally more accurate than the standard FS approach. However, they also face limited scalability due to the computation required to model more complex dependencies between matching statuses.

Given these challenges, much work has been done to increase the speed and scalability of FS methods. One approach is to use probabilistic or deterministic methods to reduce the number of record pairs considered as possible matches. With “blocking,” the analyst uses some feature in the data to allocate records into smaller partitions (or blocks), and runs a linkage algorithm independently on each block (Christen 2012). Blocking on an unreliable field, however, can lead to missed matches, making this form of blocking often undesirable (Steorts et al. 2014). “Filtering,” is a similar technique which uses some deterministic or probabilistic criteria to set the match probability for certain comparison vectors to zero after they have been created (Murray 2016, McVeigh et al. 2019). Without reducing the number of comparison vectors in the model, Enamorado et al. (2019) used a lower dimensional representation of the set of comparison vectors that allowed for faster parameter estimation of the standard FS model. Kunderinger, Reiter & Steorts (2024) then adopted this approach for a faster implementation of the bipartite model of Sadinle (2017). Kunderinger, Aleshin-Guendel & Steorts (2024) then used variational inference for further computational gains.

In this work, we continue to build on these advances and provide a scalable record linkage method without the stringent assumption that each file is free of duplicates. We propose Dirichlet Record Linkage (DRL), an extension of the comparison vector record linkage framework for the scenario of linking one duplicate free reference file to a target file that may have duplications within itself (referred to as “internal duplications”). To do so, we model the parameter governing the amount of matches between files as a Dirichlet process. Since the resulting set of possible matches for each record is extremely high dimensional, we propose a Gibbs sampler utilizing the stick-breaking representation of the Dirichlet process for tractable posterior inference. We adopt the dimension reduction techniques of Kunderinger, Reiter & Steorts (2024) for additional gains in speed and scalability. Through simulation

studies, we show that this method is able to link a reference file to a target file with various amounts of internal duplications without making any distributional assumptions on the nature of these duplications. Through a case study of voter registration records, we show that this method is scalable to a linkage task involving 400,000 records without the use of blocking.

In what follows, Section 2 reviews the work of Fellegi & Sunter (1969) and a selection of Bayesian extensions to contextualize our contribution. Section 3 introduces **DRL**, and describes our method for efficient posterior sampling. Sections 4 and 5 demonstrate the speed and accuracy of our method relative to several other FS techniques in a series of simulations and a case study of voter registration records in North Carolina. Finally, Section 6 discusses areas for further research.

## 2 THE FELLEGI-SUNTER FRAMEWORK FOR RECORD LINKAGE

Consider two data files  $A$  and  $B$  comprising  $n_A$  and  $n_B$  records, respectively, and including  $F$  linkage variables measured in both files. Let  $[x] := \{1, \dots, x\}$  denote the sequence of integers from 1 to  $x$ . For  $i \in [n_A]$ , let record  $i$  be given by  $A_i = (A_{i1}, \dots, A_{iF})$ , so that  $A = (A_i : i \in [n_A])$ . Similarly, for  $j = 1, \dots, n_B$ , let record  $j$  be given by  $B_j = (B_{j1}, \dots, B_{jF})$ , so that  $B = (B_j : j \in [n_B])$ . By convention, we label the two files such that  $n_A \geq n_B$ .

In record linkage tasks, records that refer to the same entity should be similar, and records that refer to different entities should be dissimilar. To represent this, Fellegi & Sunter (1969) proposed the comparison vector  $\gamma_{ij} = (\gamma_{ij}^1, \dots, \gamma_{ij}^F)$ , where  $\gamma_{ij}^f$  is a comparison for field  $f$  between records  $A_i$  and  $B_j$ . Binary comparisons are commonly used due to their simplicity; partial agreement or more complicated comparisons can be used through similarity metrics or distance functions (Winkler 1990, Bilenko et al. 2003, Elmagarmid et al. 2007). We assume that each field’s comparison is discretized, and we let  $L_f$  denote the number of levels for field  $f$ . We collect all of the comparison vectors as  $\mathbf{\Gamma} = \{\gamma_{ij}\}_{i=1, j=1}^{n_A, n_B}$ .

Consider records from  $A$  and  $B$  as a disjoint set of nodes. We have an edge between two records if they are coreferent (or matching). Our parameter of interest is the collection

of edges representing coreferent records. This parameter can be expressed in various ways depending on the model assumptions for a particular approach to record linkage. For example, Fellegi & Sunter (1969) used a coreferent matrix  $\Delta \in \{0, 1\}^{n_A \times n_B}$ , where

$$\Delta_{ij} = \begin{cases} 1, & \text{if records } A_i \text{ and } B_j \text{ refer to the same entity;} \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

This sparse matrix representation can become cumbersome for large linkage tasks. In the setting where each record in  $B$  can match to at most one record in  $A$ , Sadinle (2017) proposed using the more compact vector  $\mathbf{Z} = (Z_1, \dots, Z_{n_B})$  for the records in  $B$  such that

$$Z_j = \begin{cases} i, & \text{if records } A_i \text{ and } B_j \text{ refer to the same entity;} \\ n_A + j, & \text{if record } B_j \text{ does not have a match in } A. \end{cases}$$

When presenting the theory in this paper, we use either representation for convenience depending on the context. We can go back and forth between the two using  $\Delta_{ij} = I(Z_j = i)$ , where  $I(\cdot) = 1$  when the expression inside the parentheses is true, and  $I(\cdot) = 0$  otherwise.

## 2.1 Models for Comparison Vector Based Record Linkage

Record linkage models within the FS framework consist of a model for the comparison vectors  $\mathbf{\Gamma}$ , and a model for the linkage structure that classifies the record pairs as matches and non-matches. In all such methods, the model for the comparison vectors takes the form

$$\begin{aligned} \gamma_{ij} \mid \Delta_{ij} = 1 &\stackrel{iid}{\sim} \mathcal{M}(\mathbf{m}), \\ \gamma_{ij} \mid \Delta_{ij} = 0 &\stackrel{iid}{\sim} \mathcal{U}(\mathbf{u}). \end{aligned}$$

Here,  $\mathcal{M}$  and  $\mathcal{U}$  are the distributions for matching and non-matching record pairs, and  $\mathbf{m}$  and  $\mathbf{u}$  are their respective parameter values. When using comparison vectors with discrete agreement levels,  $\mathcal{M}$  and  $\mathcal{U}$  are collections of independent multinomial distributions for each linkage feature. Accordingly,  $\mathbf{m} = (\mathbf{m}_1, \dots, \mathbf{m}_F)$ , where  $\mathbf{m}_f = (m_{f1}, \dots, m_{fL_f})$  and  $m_{fl} = \mathbb{P}(\gamma_{ij}^f = l \mid \Delta_{ij} = 1)$  for all fields  $f$  and agreement levels  $l$ . The  $\mathbf{u}$  parameters are defined similarly, with  $u_{fl} = \mathbb{P}(\gamma_{ij}^f = l \mid \Delta_{ij} = 0)$ .

The original approach of Fellegi & Sunter (1969) models the matching status of each record pair independently from all other record pairs. Accordingly, the authors model  $\Delta_{ij}$  through independent, identically distributed Bernoulli random variables. We generally assume that data in the records is missing completely at random, and is thus ignorable (Little & Rubin 2002). Letting  $I_{obs}(\cdot)$  denote an indicator of whether a record field is observed, the likelihood function after marginalizing out the missing data is

$$\mathcal{L}(\Delta, m, u \mid \Gamma) = \prod_{i=1}^{n_A} \prod_{j=1}^{n_B} \prod_{f=1}^F \prod_{l=1}^{L_f} \left[ m_{fl}^{I(\Delta_{ij}=1)} u_{fl}^{I(\Delta_{ij}=0)} \right]^{I(\gamma_{ij}^f=l)I_{obs}(\gamma_{ij}^f)}. \quad (2)$$

The independence assumption at the level of the individual record pairs makes this model straightforward to fit through the EM algorithm (Sariyar & Borg 2010). Recently, much work has been done to use this model for increasingly large linkage tasks, such as through the **fastLink** package in R by Enamorado et al. (2019), and the **splink** library in Python by Linacre et al. (2022). However, the full independence assumption of the record pair matching tends to produce a large number of false matches, and post-processing is required to get a more accurate estimate of the linkage structure Jaro (1989).

Results tend to improve when prior knowledge about the linkage structure is incorporated into the model Sadinle (2017). For example, there are many situations where practitioners know that there are no duplications within each file, and this can be incorporated into the model in a Bayesian framework through a prior distribution on the linkage structure. For this setting, Sadinle (2017) proposed the “beta distribution for bipartite matching,” given by

$$\mathbb{P}(\mathbf{Z} \mid \pi) = \frac{[n_A - n_{12}(\mathbf{Z})]!}{n_A!} \pi^{n_{12}(\mathbf{Z})} (1 - \pi)^{n_B - n_{AB}(\mathbf{Z})}, \quad (3)$$

$$\pi \sim \text{Beta}(\alpha_\pi, \beta_\pi),$$

where  $n_{AB} = \sum_{j=1}^{n_B} I(Z_j < n_A + j)$  is the number of links between files  $A$  and  $B$ , and  $\alpha_\pi$  and  $\beta_\pi$  are fixed and known hyperparameters. This prior induces a Gibbs sampler that strictly enforces one-to-one matching, and has been shown to outperform the base FS method when model assumptions are satisfied. This model is implemented in R through the BRL package, but has computational complexity that grows quadratically with the size of the linkage task, and is therefore limited to small to moderate data files (Sadinle 2020).

More recently, Aleshin-Guendel & Sadinle (2023) extended the FS framework for the setting of arbitrary amount of duplications within and across an arbitrary number of files. To do so, they presented a generative process for multifile partitions, and parameterized each step of this generative process. In particular, they use a truncated Poisson distribution for the number of records associated with each unique entity exhibited in the linkage task, where the user designates an upper limit for the cluster size. Similarly the beta prior for bipartite matching, this Gibbs sampler as implemented in `multilink` becomes infeasible for larger linkage tasks.

### 2.1.1 Fast Beta Linkage

Kundinger, Reiter & Steorts (2024) proposed to model each element of  $\mathbf{Z}$  independently, relaxing the assumption that there are no duplicates within each database. Under this relaxation, they introduced the fast beta linkage (`fabl`) prior:

$$\mathbb{P}(Z_j = i \mid \pi) = \begin{cases} \pi/n_A, & \text{if } i \leq n_A, \\ 1 - \pi, & \text{if } i = n_A + j, \end{cases} \quad (4)$$

$$\pi \sim \text{Beta}(\alpha_\pi, \beta_\pi). \quad (5)$$

This prior says that a record  $B$  has some match in  $A$  with probability  $\pi$ , and that each record in  $A$  is equally likely to be that match. The `fabl` prior closely mimics the beta prior for bipartite matchings in (3) without enforcing the bipartite matching restriction within the Gibbs sampler. This prior was originally proposed for efficient MCMC sampling, but we recently showed that this prior was conducive to variational inference. This implementation of the model, dubbed “Variational Beta Linkage” (or `vabl`) was shown to conduct linkage 50-250 times fast than the original MCMC implementation in `fabl` (Kundinger, Aleshin-Guendel & Steorts 2024). If desired, a bipartite matching can be acquired through a simple post-processing step after computing the Bayes estimate.

### 3 DIRICHLET RECORD LINKAGE

Although the `fabl` framework is effective and efficient when  $A$  has no internal duplications, significant problems arise when a record in  $B$  has multiple matching records in  $A$ . In particular, posterior probability is normalized among all records in  $A$ , such that none of the matches achieves high enough posterior probability to be identified in the Bayes estimate. This amounts to a paradox: the more matches that record  $B_j$  has in  $A$ , the less likely the algorithm is to find a match. Here, we attempt to resolve this paradox by extending the `fabl` framework to handle these internal duplications in  $A$ . We emphasize that we are not interested in deduplication within  $A$  for its own sake, but rather aim to conduct record linkage in light the problems posed by these duplications. Therefore, duplications within files that have no duplications across files will go undetected; we believe this acceptable for many practical applications.

We provide updated notation to describe matching in this setting. We drop the convention of labelling datasets such that  $n_A \geq n_B$ , and instead let  $B$  denote the file that we assume is free of internal duplicates. In this setting, we refer to the duplicate free file  $B$  as the reference file, and the other file  $A$  as the target file. Let  $Z_j = (Z_{j,1}, \dots)$  be a set containing the indices for all of the records in  $A$  that are a match with record  $B_j$ , with  $Z_j = \emptyset$  denoting when  $B_j$  has no match in  $A$ . Let  $\mathbf{Z} = \{Z_j | j \in [n_B]\}$  denote the collection of all such sets for all records in  $B$ . Lastly, let  $|Z_j|$  denote the number of records in  $A$  that are linked to  $B_j$ .

Define a vector of probabilities  $\boldsymbol{\pi} = (\pi_0, \dots)$  where  $\pi_k$  is the probability that some record  $B_j$  has exactly  $k$  matches in  $A$ . To avoid setting a maximum number of matches for any  $B_j$ , we assign  $\boldsymbol{\pi}$  a Dirichlet process prior. That is,

$$f(\boldsymbol{\pi}) = \sum_{k=0}^{\infty} \pi_k \delta_{\pi_k}(\boldsymbol{\pi}),$$

where  $\delta_{\pi_k}$  is the indicator function which evaluates to zero everywhere, except for  $\delta_{\pi_k}(\pi_k) = 1$ . We model each  $\pi_k$  as a product of conditional probabilities. Let  $\boldsymbol{\eta} = (\eta_1, \dots)$ , and let  $\eta_k$  be the probability that some record in  $B$  has at least  $k$  matches, given that it has at least



$k - 1$  matches. This gives us the stick breaking representation

$$\pi_k = (1 - \eta_{k+1}) \prod_{c=1}^k \eta_c, \quad (6)$$

where  $\eta_k$  are independent random variables from a  $\text{Beta}(\alpha_\eta, \beta_\eta)$  distribution, for fixed and known hyperparameters  $\alpha_\eta$  and  $\beta_\eta$ . We use notation for an arbitrary  $\alpha_\eta$  as it makes future derivations more readable, but we fix  $\alpha_\eta = 1$  as is done in a typical Dirichlet process prior (Teh et al. 2006).

Conditional on  $B_j$  having  $k$  matches, we construct a prior specification on  $\mathbf{Z}$  such that each matching  $q$  of length  $|q|$  is equally likely. Since there are  $\binom{n_A}{|q|} = n_A! / [(n_A - |q|)!|q|!]$  ways to select  $|q|$  matching records out of all  $n_A$  possible records, we let

$$\mathbb{P}(Z_j = q | \boldsymbol{\pi}) = \frac{(n_A - |q|)!|q|!}{n_A!} \pi_{|q|}. \quad (7)$$

In Appendix 7.1.1, we show that the full conditional distribution for  $Z_j$  is given by

$$p(Z_j = q | \boldsymbol{\Gamma}, \mathbf{m}, \mathbf{u}, \boldsymbol{\pi}) \propto \frac{(n_A - |q|)!|q|!}{n_A!} \pi_{|q|} \prod_{i \in q} w_{ij}, \quad (8)$$

where for all  $i \in [n_A]$  and  $j \in [n_B]$ ,

$$w_{ij} = \prod_{f=1}^F \prod_{l=1}^{L_f} \left( \frac{m_{fl}}{u_{fl}} \right)^{I(\gamma_{ij}^f = l) I_{obs}(\gamma_{ij}^f)}.$$

Unfortunately, while the full conditional in (8) has a closed form, sampling from it directly is infeasible for almost any record linkage task. In particular, there are  $2^{n_A}$  possible matchings for each  $B_j$ , such that even enumerating these possible matchings would be computationally demanding. In Section 3.1, we show how we can sequentially sample individual components of  $Z_j$  at a much lower computational cost. In Section 3.2, we outline our approach to further reduce computational complexity

### 3.1 Sampling $Z_j$

We use the stick breaking representation in (6) to generalize the fast beta prior in (4), producing a sequence of priors that allows for each record in  $B$  to match to multiple records

in  $A$ . In each iteration of the Gibbs sampler, we sample an initial set of links using  $\eta_1$ . For each record in  $B$  that was found to have a link, we remove the linked record in  $A$  from consideration, and then sample another potential link with  $\eta_2$ . We continue, using  $\eta_k$  in the  $k^{th}$  matching step, until no new links are found, at which we point the matching phase terminates. Crucially, there is no need to specify a maximum number of links per record, as this is estimated through the model.

We use the superscript  $(s)$  to denote the iteration in the Gibbs sampler. Let  $Z_{j,-k}^{(s)} = (Z_{j,1}^{(s)}, \dots, Z_{j,k-1}^{(s)})$  denote the indices for records in  $A$  that have been matched to  $B_j$  before matching step  $k$  in Gibbs iteration  $s$  (with  $Z_{j,-k}^{(s)} = \emptyset$ ). Importantly, the size and ordering of  $Z_j^{(s)}$  is random across Gibbs iterations. Let  $D_k^{(s)} = \{j \in [n_B] | Z_{j,k}^{(s)} \neq \emptyset\}$  to denote the set of records in  $B$  linked to at least  $k$  records in  $A$  in Gibbs iteration  $s$  (with  $D_0^{(s)} = [n_B]$ ).

When  $B_j$  has been linked to  $k-1$  records, we use a prior that says that the probability that  $B_j$  has a  $k^{th}$  match is  $\eta_k$ , and that all  $n_A - (k-1)$  remaining records in  $A$  are equally likely to be linked. For each  $j \in D_{k-1}^{(s)}$ , we use

$$\mathbb{P}(Z_{j,k}^{(s+1)} = q_k | \eta_k, Z_{j,-k}^{(s+1)}) = \begin{cases} \frac{\eta_k}{n_A - (k-1)}, & q_k \notin [n_A] \setminus Z_{j,-k}^{(s+1)}, \\ 1 - \eta_k, & q_k = \emptyset. \end{cases} \quad (9)$$

This sequence of priors leads to sequence of posteriors that can be used to sample arbitrarily many links for record  $B_j$ . These full conditional distributions are given by

$$\mathbb{P}(Z_{j,k}^{(s+1)} = q_k | Z_{j,-k}^{(s+1)}, \eta_k, \mathbf{m}, \mathbf{u}, \Gamma) \propto \begin{cases} \frac{\eta_k}{n_A - (k-1)} w_{q_k, j}, & q_k \in [n_A] \setminus Z_{j,-k}^{(s+1)}, \\ 1 - \eta_k, & q_k = \emptyset. \end{cases} \quad (10)$$

We provide a derivation in Appendix 7.1.2.

This sequential sampler produces an output  $Z_j^{(s)} = q = (q_1, \dots, q_k)$  when  $Z_{j,c}^{(s)} = q_c$  for steps  $c \in [k]$ , and the  $k+1$  step produces  $Z_{j,k+1} = \emptyset$ . While this vector is necessarily ordered, all reorderings of the same elements are equivalent for the purposes of record linkage. That is,  $Z_j^{(s)} = (i, i')$  and  $Z_j^{(s)} = (i', i)$  both communicate that record  $B_j$  is matched to records  $A_i$  and  $A_{i'}$ . Let  $\sigma(q)$  denote all  $|q|!$  possible orderings of the elements of  $q$ . Marginalizing

over such orderings, we have

$$\begin{aligned}
\mathbb{P}(Z_{j,k+1}^{(s)} = \emptyset | \Gamma, \mathbf{m}, \mathbf{u}, \boldsymbol{\eta}) & \sum_{q' \in \sigma(q)} \prod_{c=1}^k \mathbb{P}(Z_{j,c}^{(s)} = q'_c | \Gamma, \mathbf{m}, \mathbf{u}, \boldsymbol{\eta}) \\
& \propto (1 - \eta_{k+1}) \sum_{q' \in \sigma(q)} \prod_{c=1}^k \frac{\eta_c}{n_A - (c-1)} \prod_{c=1}^k w_{q'_c, j} \\
& = (1 - \eta_{k+1}) k! \prod_{c=1}^k \frac{\eta_c}{n_A - (c-1)} \prod_{c=1}^k w_{q_c, j} \\
& = \frac{(n_A - k)! k!}{n_A!} (1 - \eta_{k+1}) \prod_{c=1}^k \eta_c \prod_{c=1}^k w_{q_c, j} \\
& = \frac{(n_A - k)! k!}{n_A!} \pi_k \prod_{c=1}^k w_{q_c, j} \\
& = \mathbb{P}(Z_j^{(s)} = q | \Gamma, \mathbf{m}, \mathbf{u}, \boldsymbol{\pi}).
\end{aligned}$$

Thus, the probability that the sequential process samples each of the components of  $q$  (in any order) is equal to the joint probability of  $q$  as expressed in the joint distribution in (8).

We summarize the full Gibbs sampler in Algorithm 1. We initialize  $\mathbf{Z}$  to reflect no matches across data files; that is,  $Z_j = \emptyset$  for all  $j \in [n_B]$ . We then iteratively sample  $\mathbf{m}$ ,  $\mathbf{u}$ ,  $\boldsymbol{\eta}$ , and  $\mathbf{Z}$  for a prespecified number of Gibbs iterations. The full conditional for  $Z_{j,k}^{(s)}$  is provided in Appendix 7.1.2, the full conditional for  $\boldsymbol{\eta}^{(s)}$  is provided in Appendix 7.1.3, and full conditionals for  $\mathbf{m}^{(s)}$  and  $\mathbf{u}^{(s)}$  are provided in Appendix 7.1.4.

### 3.2 Efficient Posterior Inference

As described in Algorithm 1, the Gibbs sampler for DRL has computational complexity that depends on both  $n_A$  and  $n_B$ , making it infeasible for large linkage tasks. Adapting the methodology of Kundinger, Reiter & Steorts (2024), however, we use hashing to substantially reduce this computational complexity.

Since each component  $\gamma_{ij}^f$  of the comparison vector is discrete, there are only finitely many possible realizations of the comparison vector  $\gamma_{ij}$ . Let  $P$  be the number of unique agreement patterns observed in  $\boldsymbol{\gamma}$ . This number is bounded above by  $P^* = \prod_{f=1}^F (L_f + 1)$ , where the addition of 1 to  $L_f$  for each field accounts for the possibility of missing values.

---

**Algorithm 1** Gibbs Sampler for Dirichlet Record Linkage

---

**Input** : Data  $\mathbf{\Gamma}$  and hyperparameters  $\boldsymbol{\alpha}, \boldsymbol{\beta}$ , and  $\beta_\eta$

**Initialize**:  $Z_j = \emptyset, \forall j \in [n_B]$

**for**  $s \in [S]$  **do**

**for**  $f \in [F]$  **do**

        Sample  $\mathbf{m}_f^{(s+1)} \sim \text{Dirichlet}(\alpha_{f1}(\mathbf{Z}^{(s)}), \dots, \alpha_{fL_f}(\mathbf{Z}^{(s)}))$ , where

$$\alpha_{fl}(\mathbf{Z}^{(s)}) = \alpha_{fl} + \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} I_{\text{obs}}(\gamma_{ij}^f) I(\gamma_{ij}^f = l) I(i \in Z_j^{(s)})$$

        Sample  $\mathbf{u}_f^{(s+1)} \sim \text{Dirichlet}(\beta_{f1}(\mathbf{Z}^{(s)}), \dots, \beta_{fL_f}(\mathbf{Z}^{(s)}))$

$$\beta_{fl}(\mathbf{Z}^{(s)}) = \beta_{fl} + \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} I_{\text{obs}}(\gamma_{ij}^f) I(\gamma_{ij}^f = l) I(i \notin Z_j^{(s)})$$

**end**

    Set  $k = 1$

**while**  $D_{k-1}^{(s+1)} \neq \emptyset$  **do**

        Sample  $\eta_k^{(s+1)} \sim \text{Beta}(n_k^{(s)} + \alpha_\eta, n_{k-1}^{(s)} - n_k^{(s)} + \beta_\eta)$ , where

$$n_k^{(s)} = \sum_{j=1}^{n_B} I(|Z_j^{(s)}| \geq k)$$

**for**  $j \in D_{k-1}^{(s+1)}$  **do**

$$\text{Sample } Z_{j,k}^{(s+1)} \propto \begin{cases} \frac{\eta_k^{(s+1)}}{n_A - (k-1)} w_{q_k,j}^{(s+1)}, & q_k \in [n_A] \setminus Z_{j,-k}^{(s+1)}, \\ 1 - \eta_k^{(s+1)}, & q_k = \emptyset. \end{cases}$$

**end**

        Set  $k = k + 1$

**end**

**end**

**return**  $\{\mathbf{Z}^{(s)}, \mathbf{m}^{(s)}, \mathbf{u}^{(s)}, \boldsymbol{\eta}^{(s)}\}_{s=1}^S$

---

This upper bound does not scale with  $n_A$  or  $n_B$ , but rather is determined by  $F$  and  $L_f$ . By viewing each comparison vector as the realization of one agreement pattern rather than  $F$  components, we greatly reduce the dimensionality of the linkage task and calculate summary statistics for efficient posterior inference. We provide a more thorough explanation of this hashing procedure in Appendix 7.2.

Most importantly for DRL, we use hashing to break the sampler in (10) into two steps. First, we sample the agreement pattern for  $Z_{j,k}$  among the  $P$  agreement patterns exhibited in the data. Second, we sample the record  $Z_{j,k}$  uniformly among all records in  $A$  with comparison vectors exhibiting that pattern. We provide details in Appendix 7.3.1. We also provide an efficient posterior update for  $\mathbf{m}$  and  $\mathbf{u}$  in Appendix 7.1.4. These changes can greatly improve the speed of the sampler, as described in Lemma 1 and the subsequent discussion.

**Lemma 1.** *Let  $n_A$  and  $n_B$  be the number of records in files  $A$  and  $B$ , respectively. Let  $F$  be the number of fields used for comparisons across records, and  $P$  be the number of patterns that comparison vectors exhibit in  $A \times B$ . We assume  $C$  cores available for parallelization and a Gibbs sampler with  $T$  iterations. Then, the overall computational complexity of DRL with hashing is  $O(Fn_An_B/C) + O(TPn_An_B/C)$ .*

*Proof.* We showed in Kundering, Reiter & Steorts (2024) that the complexity for constructing the comparison vectors was  $O(Fn_An_B/C)$ . We also showed that the complexity for sampling  $\mathbf{m}$  and  $\mathbf{u}$  is  $O(P)$ . We therefore focus on here on the complexity of sampling  $\mathbf{Z}$ .

At each matching step, we sample  $Z_{j,k}$  for each  $j \in D_{k-1}$ . In each instance, we first sample the agreement pattern of the match with complexity  $O(P)$ , and then we sample the record exhibiting that pattern with complexity  $O(1)$ . For any matching step  $k$ ,  $|D_{k-1}| \leq n_B$ , so that the complexity is at worst  $O(Pn_B)$ . For any linkage task, each record  $j \in [n_B]$  matches to at most  $n_A$  records in  $A$ , so that the complexity is at worst  $O(Pn_An_B)$ . Therefore, with parallelization over  $C$  cores for  $T$  iterations, sampling  $\mathbf{Z}$  has complexity  $O(TPn_An_B/C)$ . Thus, the overall computational complexity of DRL with hashing is  $O(Fn_An_B/C) + O(TPn_An_B/C)$ .  $\square$

The complexity proven in Lemma 1 is a loose upper bound that is only seen in the

unrealistic linkage scenario where every record in  $B$  matches to every record in  $A$  (so that all records in the both data files refer to the same entity). In practice, sampling  $\mathbf{Z}^{(s)}$  at iteration  $s$  has complexity  $O\left(P \sum_{k=1}^{n_A} n_k^{(s)}\right)$  where  $n_k^{(s)} = \sum_{j=1}^{n_B} I(|Z_j^{(s)}| \geq k)$  is the number of records in  $B$  with at least  $k$  links in  $A$ . This quantity grows with the number of records in  $B$  and with the rate of internal duplication in  $A$ , but does not grow with  $n_A$ . In particular, since records in  $B$  generally have a small number duplicates in  $A$ , most terms in the summation  $\sum_{k=1}^{n_A} n_k^{(s)}$  are zero. We show in Section 4 that DRL exhibits similar computation time to `fabl` when the rate of duplication is low, and modest increases to computation time as the rate of duplication grows.

### 3.3 Bayes Estimate

After generating MCMC samples through Algorithm 1, we require a loss function to produce a point estimate of the linkage structure. Unfortunately, many of the loss functions that have been proposed for different record linkage models are not amenable to DRL. For example, Sadinle (2017) proposed a loss function that ensured a Bayes estimate that respected one-to-one matching and allowed certain components of the estimate to remain undeclared and left for clerical review. However, this loss function relies on making exactly one linkage decision per record in  $B$ , and is not amenable to our scenario where we do not set a maximum number of links per record. More recently, Aleshin-Guendel & Sadinle (2023) proposed a loss function that estimated a partition, and thus respected transitivity, and also allowed for leaving parts of the Bayes estimate undeclared. The implementation of this loss function however relies on each MCMC sample itself being a partition, which does not hold true in our relaxed setting.

For this reason, we instead use a generalization of the absolute number of errors loss function from Tancredi & Liseo (2011). It is more convenient here to use the matrix representation of the linkage structure  $\Delta$  as in (1), rather than the  $\mathbf{Z}$  notation used throughout Section 3. Specifically, we seek an estimate  $\hat{\Delta}$  of the linkage structure  $\Delta$  that

minimizes the loss function

$$L(\Delta, \hat{\Delta}) = \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} \tau \hat{\Delta}_{ij} (1 - \Delta_{ij}) - \Delta_{ij} (1 - \hat{\Delta}_{ij}),$$

where  $\tau > 0$  is a penalization term specified by the user to weigh the importance of false matches relative to false non-matches. This loss function is equivalent to

$$L(\Delta, \hat{\Delta}) = \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} \tau \hat{\Delta}_{ij} - (1 + \tau) \hat{\Delta}_{ij} \Delta_{ij} + \Delta_{ij}.$$

Minimizing  $L(\Delta, \hat{\Delta})$  is equivalent to maximizing the quantity

$$W(\hat{\Delta}) = \hat{\Delta}_{ij} [(1 + \tau) \Delta_{ij} - \tau].$$

The expectation of this quantity with respect to the linkage structure is given by

$$E[W(\hat{\Delta})] = \hat{\Delta}_{ij} [(1 + \tau) \mathbb{P}(\Delta_{ij} = 1) - \tau],$$

which is maximized by the decision rule

$$\hat{\Delta}_{ij} = \begin{cases} 1, & \text{if } \mathbb{P}(\Delta_{ij} = 1) > \frac{\tau}{1+\tau}, \\ 0 & \text{otherwise.} \end{cases} \quad (11)$$

When the user specifies  $\tau = 1$ , indicating that false matches and false non-matches are weighted equally, we recover the decision rule of Tancredi & Liseo (2011) with a threshold of 0.5 to declare matches. This decision rule also functions similarly to that of Sadinle (2017) under certain choices of the losses.

Since the DRL prior as specified in (7) does not strictly enforce the requirement that there are no duplicates in  $B$ , it is possible for this Bayes estimate to link multiple records in  $B$  to the same record in  $A$ . To obtain a Bayes estimate that corresponds to our model assumptions, we minimize the expected loss subject to the constraint that  $\hat{Z}_{j,k} \neq \hat{Z}_{j',k}$  for all  $j \neq j'$  and all  $k$ . This means that when multiple records in  $B$  are declared as matching to the same record in  $A$  in the initial Bayes estimate, we simply accept the one with highest posterior probability, and declare all other pairs as non-matching. We make the simplifying assumption that all records that match to the same  $B_j$  also match to each other, but this is not verified in the model. We believe this assumption is well suited for many practical applications, and is worth the computational advantages of the DRL approach.

## 4 SIMULATIONS

We demonstrate the accuracy of DRL through an adaptation of the simulation study from Aleshin-Guendel & Sadinle (2023). We compare the performance of DRL against models implemented in `fastLink`, `vab1`, and `multilink`. BRL and `fab1` provide similar accuracy to `vab1`, and are therefore omitted.

For each simulation, we construct two data files  $A$  and  $B$  of size  $n_A = n_B = 500$ . File  $B$  is constructed to have no internal duplicates, and we designate 10% of the records in  $B$  (or 50 records) to have some number of matches in  $A$ . The number of matches these records have in  $A$  is sampled from a Poisson distribution truncated to the interval  $[1, 5]$ . We use Poisson distributions with means 0.1, 1, and 2, and denote these settings as low, medium, and high duplication settings respectively. Each of these duplicate records has up to three altered fields relative to the original record in  $A$ . After accounting for the records in  $A$  that have a match in  $B$ , the rest of the records in  $A$  are comprised of unique entities. We follow the methodology of Christen & Vatsalan (2013) to generate original records for the unique entities by simulating first name, last name, age, postal code, and occupation from frequency tables based on demographic information in Australia. Fields are then chosen for distortion uniformly at random. Names are subject to string insertions, deletions and substitutions, as well as common keyboard, phonetic, and optical recognition errors. Age and occupation are distorted through keyboard errors and missingness.

We create comparison vectors for these linkage tasks according to the thresholds described in Table 1. We use uniform priors for  $\mathbf{m}$  and  $\mathbf{u}$ , with  $\alpha_{fl} = \beta_{fl} = 1$  for all  $f$  and  $l$ . We use uniform priors for  $\pi$  for `vab1`, and also uniform priors for each  $\eta_k$  in the sequential sampler for DRL. For `multilink`, we correctly designate that there are no duplicates in  $B$  and that  $A$  can have up to 5 records that are linked together (and to one record in  $B$ ). We use the default setting of a Poisson distribution with mean 1 for the number of records clustered together in  $A$ . In consultation with the authors of `multilink`, we also implement the method with filtering, requiring that both first and last name have an agreement level of 1 or 2 in order for the record pair to be considered for matching. For DRL and `multilink`, we use a Gibbs sampler with 1000 iterations and discard the first 100 as burn-in. For `fastLink`



Fields	Similarity	Level of Disagreement		
		1	2	3
First and Last Name	Levenstein	0	(0, .25]	(.5, .1]
Age, Postal Code, and Occupation	Binary	Agree	Disagree	

Table 1: Construction of comparison vectors for accuracy study with simulated data files of Section 4.

and **vab1**, we run the algorithm with a convergence tolerance of 1e-04, in line with default settings. **fastLink** results are provided without the Jaro correction that enforces one-to-one matching.

We compute point estimates of the linkage structure using comparable decision rules for each method. For each Bayesian method, we use losses reflecting equal importance for false matches and false non-matches. For **vab1** and **DRL**, this results in a probability threshold of 0.5 to declare record pairs as a match. We choose this threshold for **fastLink** as well. For each of these, we obtain an initial point estimate, then post-process so that no two records in  $B$  match to same record in  $A$ . For **multilink**, Aleshin-Guendel & Sadinle (2023) do not provide closed form expressions for the decision rule for their Bayes estimate, but we select analogous losses.

To compare the accuracy of each method, we use recall, precision, and F-measure as defined by Christen (2012). Recall is the proportion of true matches found by the estimate; given by

$$\hat{\theta}_{\text{Recall}} = \frac{\sum_{i=1}^{n_A} \sum_{j=1}^{n_B} \Delta_{ij} \hat{\Delta}_{ij}}{\sum_{i=1}^{n_A} \sum_{j=1}^{n_B} \Delta_{ij}}.$$

Precision is the proportion of links found by the estimate that are true matches, given by

$$\hat{\theta}_{\text{Precision}} = \frac{\sum_{i=1}^{n_A} \sum_{j=1}^{n_B} \Delta_{ij} \hat{\Delta}_{ij}}{\sum_{i=1}^{n_A} \sum_{j=1}^{n_B} \hat{\Delta}_{ij}}.$$

Lastly, we compute F-measure, a pseudo-average of recall and precision, given by

$$\hat{\theta}_{\text{F-measure}} = 2 \left( \frac{\hat{\theta}_{\text{Recall}} \times \hat{\theta}_{\text{Precision}}}{\hat{\theta}_{\text{Recall}} + \hat{\theta}_{\text{Precision}}} \right).$$

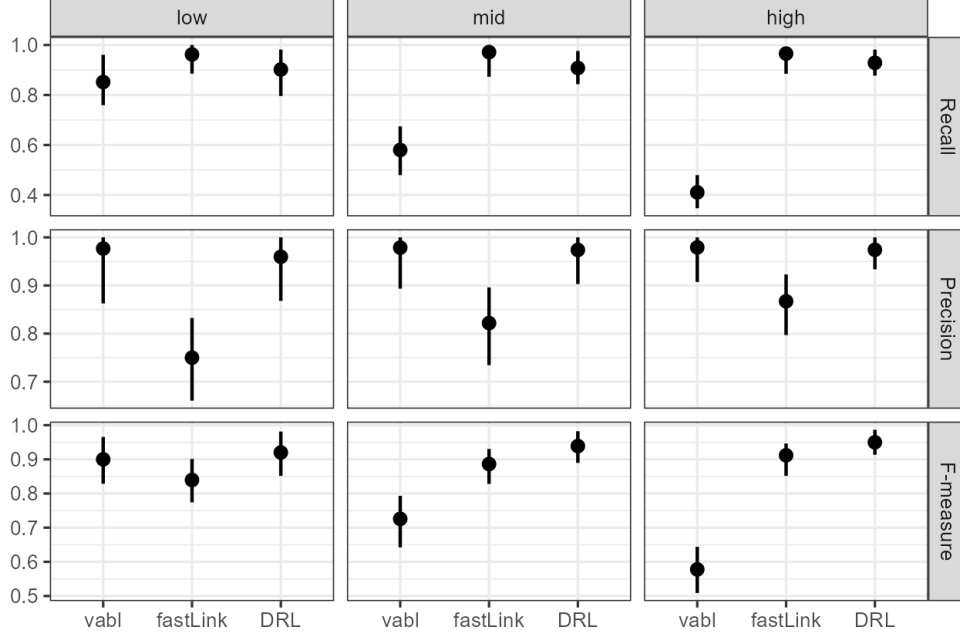


Figure 1: Results from simulation study of described in Section 4 for **vabl**, **fastLink**, and **DRL**. For each method and simulation setting, dots show the median, and bars show the 2.5<sup>th</sup> and 97.5<sup>th</sup> percentiles of recall and precision of Bayes estimates across 100 simulated datasets.

In Figure 1, we see that in the low duplication setting, **vabl** performs comparatively with **DRL** in terms of recall and precision. However, as the number of internal duplications in  $A$  grows, recall suffers. This is because **vabl** is constrained to identify at most one match for each record in  $B$ , and in many cases, posterior probability is split among several truly matching record pairs such that none of them reach the threshold to be declared a match. In contrast, **fastLink** maintains higher recall at all duplication settings, and precision improves as the rate of duplication grows. Upon closer analysis, the actual number of false positive matches remains similar across the duplication settings, but the larger number of true positives results in higher precision.

The results from **multilink** require a more nuanced interpretation. While **DRL** only evaluates the record pairs between each record in  $A$  and one particular  $B_j$ , **multilink** must also consider the comparison vectors between record pairs in  $A$ . Notably, each record in  $A$  can have up to three errors relative to the original record in  $B$ , which means that records in

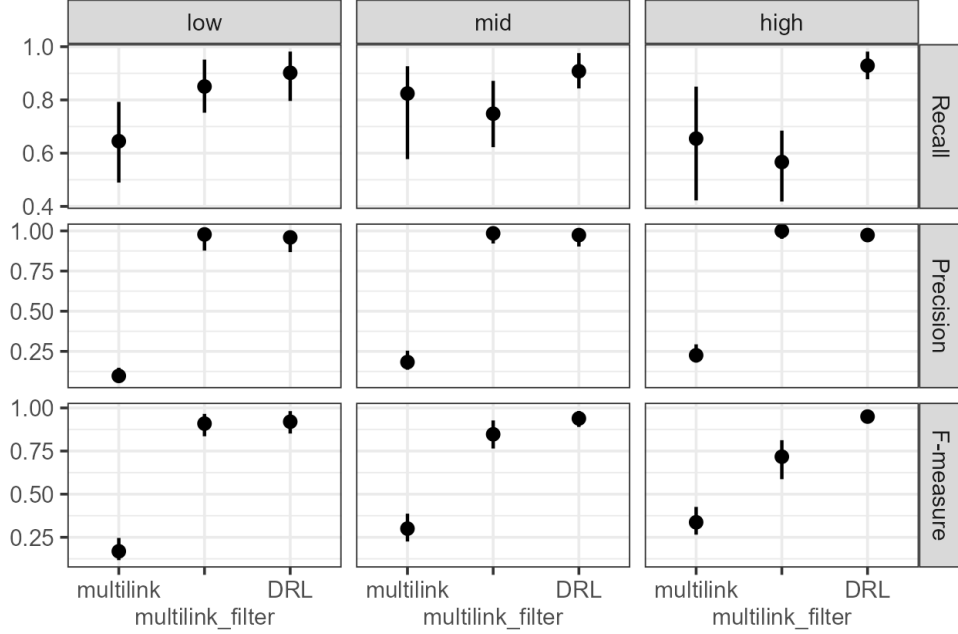


Figure 2: Results from simulation study of described in Section 4 for **multilink** (with and without filtering) and **DRL**.

$A$  can have up to six errors relative to each other. This means that **multilink** is engaging in a more difficult matching process, and in this sense, is not directly comparable to **DRL**. With that caveat mentioned, we see in Figure 2 that **multilink** without filter has poor precision under these simulation settings. With filtering however, the precision greatly improves. Nonetheless, as the rate of duplication increases, recalls suffers for both implementations of **multilink**. This because 5% to 10% of truly matching record pairs get removed from consideration by our chosen filtering scheme, and because of the difficulty of matching large clusters as described above.

In contrast, **DRL** maintains strong performance at all duplication rates considered. Notably, **DRL** has only slightly poorer precision than **vab1** even in the low duplication setting where the model assumptions of **vab1** are closest to being satisfied. Additionally, **DRL** is consistently able to match records in  $B$  to several records in  $A$  without needing to set a maximum number of links through the prior distribution, as is needed in **multilink**.

We note that **vab1** is incredibly fast, reaching convergence in around 0.01 seconds on average. Due to the design of the **fastLink** package in R, we cannot directly measure the

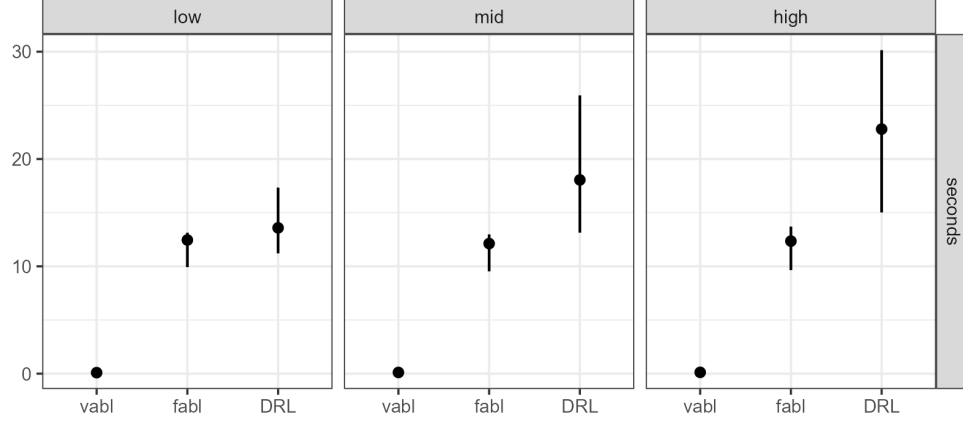


Figure 3: Computation time from simulation study of described in Section 4 for `fabl` and `DRL`.

speed of the EM algorithm separately from the construction of the comparison vectors, but we assume that it is even faster than `vabl`. When we use filtering, `multilink` completes 1000 iterations of the Gibbs sampler in about 0.5 seconds. Without filtering however, it takes around 1000 seconds. Lastly, `DRL` completes the 1000 Gibbs iterations in about 15 seconds.

In Figure 3 we see that `fabl` completes in an average of 10 seconds at all duplications levels in the simulation. In contrast, computation time for `DRL` increases as the duplication level grows from an average of 10 seconds at the low level to 15 seconds at the highest level. We note that `vabl` is incredibly fast, reaching convergence in around 0.01 seconds on average. For this reason, we recommend `vabl` in cases where it is known that records in  $B$  match to at most record in  $A$ , and recommend `DRL` when the duplications within  $A$  makes `vabl` unreliable.

Due to the design of the `fastLink` package in `R`, we cannot directly measure the speed of the EM algorithm separately from the construction of the comparison vectors, but we assume that it is even faster than `vabl`.

We note that this is only slightly slower than `fabl`, and clearly not approaching the upper bound on the complexity as shown in Lemma 1.

We conducted another simulation with 30% of records in  $B$  having matches in  $A$ , with

all other settings held constant. Under these settings, **fastLink** and **multilink** perform better, but **DRL** still outperforms. We include these results in Appendix 7.4.

## 5 CASE STUDY

We demonstrate our method on two snapshots of the North Carolina Voter Registration (NCVR) database taken two months apart (Christen 2014). The snapshots are filtered to include only those voters whose details changed over the two-month period, so there are no matching records with full agreement on all fields. We use first name, middle name, last name, as string fields, and street address and age as categorical fields. Unique voter registration numbers are provided, however they are known to contain some errors. The NCVR dataset is not publicly available due to sensitive information. However, we have permission to utilize it for publication by its owner.

Using the voter registration numbers, we can see that each file has internal duplication rates of about 1%. In this analysis, we deduplicate file  $B$ , and leave  $A$  with duplicates. In practice, such low amount of internal duplication may not warrant the use of **DRL** since **vab1** is considerably faster. However, we demonstrate here that even with such few internal duplications, **DRL** is effective at identifying these multiple matches and does not declare too many false matches.

We compare four approaches: **vab1**, **DRL**, **fastLink**, and **fastLink** with the Jaro correction. This task is too large for running **multilink** in its current implementation in **R**, and is thus omitted. We use a threshold of 0.5 to declare matches, and post-process so no two records in  $B$  match to the same record in  $A$ , as done in Section 4. Results are in Table 2.

We see that the initial **fastLink** estimate without one-to-one post-processing results in the highest recall, but leads to an undesirable amount of false positives. Since the posterior match probability of each record pair is computed independently, we expect such behavior on large linkage tasks such as this. When we use the methodology of Jaro (1989) to achieve a one-to-one matching, we get considerably better results. However, we lose the possibility of identifying cases where one record in  $B$  matches to two records in  $A$ , and we still have

Method	Recall	Precision	F-measure
DRL	0.9891623	0.9735309	0.9812843
vabl	0.9743739	0.9798799	0.9771191
fastLink	0.9988472	0.8494865	0.9181321
fastLink (with Jaro)	0.9514490	0.9664229	0.9588775

Table 2: Accuracy results for NCVR, showing the DRL has the highest F-measure.

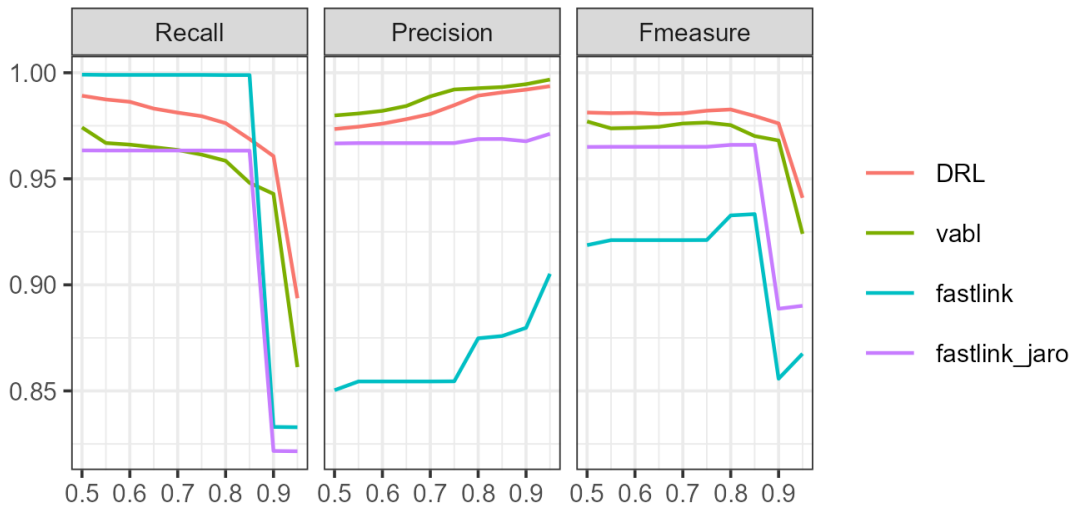


Figure 4: Accuracy metrics for NCVR data at various match probability thresholds. We see that DRL maintains the strongest F-measure for all thresholds considered.

worse recall and precision than attained under **vabl**.

In Figure 4, we show the precision, recall, and F-measure under various choices of match probability thresholds, corresponding to different choices for  $\tau$  in the decision rule in (11). We see that **vabl** and **DRL** produce more refined match probability estimates, allowing the accuracy metrics to vary smoothly across the different thresholds. In contrast, under **fastLink**, all record pairs of the same agreement pattern have the same posterior probability, leading to much more coarse posterior distribution and more erratic curves. We see that **vabl** maintains the highest precision at all thresholds, but that **DRL** is able to identify enough more additional true matches that **DRL** maintains the highest overall F-measure.

Lastly, we note a limitation in using preexisting software for record linkage tasks. In its currently implementation, **fastLink** allows users to define normalized Levenshtein distance thresholds for coding field comparisons into three discrete levels. In this data however, it is common for a record to contain a middle initial (like “E”), rather than a full middle name (like “Elizabeth”). Using Levenshtein distance thresholds, the comparison field for middle name for such a pair of records could be coded as a full disagreement, when it should be more reasonably given its own agreement level. Likewise, when two records have a middle initial that happens to match, this is coded as a full agreement, even though it is possible these initials represent different names. In this analysis, 27.6% of the false matches under DRL at the 0.5 probability threshold included at least one record containing a middle initial rather than middle name. A more careful construction of the comparison vectors is likely to improve results for all methods considered.

## 6 CONCLUSION

In this paper, we have introduced Dirichlet Record Linkage, a new Bayesian method for matching one duplicate free reference file to another file that may have duplicates. We have discussed the trade offs between model complexity and speed across several comparison vector based methods. If a practitioner is sure that each file has no (or reasonably few) internal duplicates, then **vabl** allows for high accuracy and low computation time. If a practitioner is interested in deduplicating both datasets simultaneously, the additional computation of **multilink** may be worth the ability to needs to more carefully model the partition structure of the data. When a practitioner is confident that one dataset is relatively duplicate free, DRL offers an intermediate method that is robust to internal duplications in the other file while maintaining reasonable speed.

In future work, we seek to adapt the methodology presented here to more closely approximate the model implemented by **multilink**. At current, DRL either post-processes a Bayes estimate such that there are no implicit matchings between two records in the “duplicate-free” data file, or refrains from post-processing altogether, leaving a Bayes estimate that may not respect transitive closure. While we have argued this acceptable for many

practical purposes, we acknowledge that it is not acceptable for others. Additionally, while the MCMC sampler derived in this paper was shown to scale to sizeable linkage tasks, we also acknowledge the inherent scalability limitations of MCMC methods. We hope to explore the utility of variational inference in addressing these limitations(as in Kunding, Aleshin-Guendel & Steorts (2024)), and develop new methods for the challenges of modeling the joint distribution of clusters of matching records of varying sizes.



## REFERENCES

- Aleshin-Guendel, S. & Sadinle, M. (2023), ‘Multifile partitioning for record linkage and duplicate detection’, *Journal of the American Statistical Association* **118**(543), 1786–1795.
- Betancourt, B., Sosa, J. & Rodríguez, A. (2022), ‘A prior for record linkage based on allelic partitions’, *Computational Statistics & Data Analysis* **172**, 107 – 474.
- Bilenko, M., Mooney, R. J., Cohen, W. W., Ravikumar, P. & Fienberg, S. E. (2003), ‘Adaptive Name Matching in Information Integration’, *IEEE Intelligent Systems* **18**(5), 16–23.
- Christen, P. (2012), *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*, Data-Centric Systems and Applications, Springer-Verlag, Berlin Heidelberg.
- Christen, P. (2014), Preparation of a real temporal voter data set for record linkage and duplicate detection research, Technical report, Australian National University.  
**URL:** <http://users.cecs.anu.edu.au/~christen/publications/ncvoter-report-29june2014.pdf>
- Christen, P. & Vatsalan, D. (2013), Flexible and extensible generation and corruption of personal data, in ‘Proceedings of the 22nd ACM International Conference on Information and Knowledge Management’, CIKM ’13, Association for Computing Machinery, New York, NY, USA, p. 1165–1168.
- Elmagarmid, A. K., Ipeirotis, P. G. & Verykios, V. S. (2007), ‘Duplicate Record Detection: A Survey’, *IEEE Transactions on Knowledge and Data Engineering* **19**(1), 1–16.
- Enamorado, T., Fifield, B. & Imai, K. (2019), ‘Using a Probabilistic Model to Assist Merging of Large-Scale Administrative Records’, *American Political Science Review* **113**(2), 353–371.
- Fair, M. (2004), ‘Generalized record linkage system—statistics canada’s record linkage software’, *Austrian Journal of Statistics* **33**(1&2), 37–53.

- Fellegi, I. P. & Sunter, A. B. (1969), ‘A Theory for Record Linkage’, *Journal of the American Statistical Association* **64**(328), 1183–1210.
- Fortini, M., Liseo, B., Nuccitelli, A. & Scanu, M. (2001), ‘On Bayesian Record Linkage’, *Research in Official Statistics* **4**(1), 185–198.
- Gill, L. & Goldacre, M. (2003), ‘English national record linkage of hospital episode statistics and death registration records’, *Report to the Department of Health* .
- Gutman, R., Afendulis, C. C. & Zaslavsky, A. M. (2013), ‘A Bayesian Procedure for File Linking to Analyze End-of-Life Medical Costs’, *Journal of the American Statistical Association* **108**(501), 34–47.
- Hof, M. H., Ravelli, A. C. & Zwinderman, A. H. (2017), ‘A probabilistic record linkage model for survival data’, *Journal of the American Statistical Association* **112**(520), 1504–1515.
- Jaro, M. A. (1989), ‘Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida’, *Journal of the American Statistical Association* **84**(406), 414–420.
- Jaro, M. A. (1995), ‘Probabilistic linkage of large public health data files’, *Statistics in Medicine* **14**(5-7), 491–498.
- Kundinger, B., Aleshin-Guendel, S. & Steorts, R. C. (2024), ‘Variational Beta Linkage’, *TBD* .
- Kundinger, B., Reiter, J. & Steorts, R. C. (2024), ‘Efficient and Scalable Bipartite Matching with Fast Beta Linkage (fabl)’, *Bayesian Analysis (Accepted)* .
- Lai, R., D’Amour, A., Yu, A., Sun, Y., Torvik, V. & Fleming, L. (2011), ‘Disambiguation and co-authorship networks of the us patent inventor database’, *Harvard Institute for Quantitative Social Science, Cambridge, MA* **2138**.
- Linacre, R., Lindsay, S., Manassis, T., Slade, Z., Hepworth, T., Kennedy, R. & Bond, A. (2022), ‘Splink: Free software for probabilistic record linkage at scale.’, *International*

*Journal of Population Data Science* **7**(3).

**URL:** <https://ijpds.org/article/view/1794>

Little, R. J. A. & Rubin, D. B. (2002), *Statistical Analysis with Missing Data*, Wiley.

Lum, K., Price, M. E. & Banks, D. (2013), ‘Applications of multiple systems estimation in human rights research’, *The American Statistician* **67**(4), 191–200.

Marchant, N. G., Kaplan, A., Elazar, D. N., Rubinstein, B. I. & Steorts, R. C. (2021), ‘d-blink: Distributed end-to-end bayesian entity resolution’, *Journal of Computational and Graphical Statistics* **30**(2), 406–421.

McVeigh, B. S., Spahn, B. T. & Murray, J. S. (2019), ‘Scaling bayesian probabilistic record linkage with post-hoc blocking: An application to the california great registers’, *arXiv preprint arXiv:1905.05337*.

Murray, J. S. (2016), ‘Probabilistic record linkage and deduplication after indexing, blocking, and filtering’, *Journal of Privacy and Confidentiality* **7**(1), 3–24.

Price, M., Gohdes, A. & Ball, P. (2015), ‘Documents of war: Understanding the Syrian conflict’, *Significance* **12**(2), 14–19.

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

Sadinle, M. (2017), ‘Bayesian Estimation of Bipartite Matchings for Record Linkage’, *Journal of the American Statistical Association* **112**(518), 600–612.

Sadinle, M. (2018), ‘Bayesian propagation of record linkage uncertainty into population size estimation of human rights violations’, *The Annals of Applied Statistics* **12**(2), 1013–1038.

Sadinle, M. (2020), *BRL: Beta Record Linkage*. R package version 0.1.0.

**URL:** <https://CRAN.R-project.org/package=BRL>

Sadosky, P., Shrivastava, A., Price, M. & Steorts, R. C. (2015), ‘Blocking methods applied to casualty records from the Syrian conflict’, *arXiv e-prints* pp. 1–25. arxiv:1510.07714.

- Sariyar, M. & Borg, A. (2010), ‘The RecordLinkage Package: Detecting Errors in Data’, *The R Journal* **2**(2), 61–67.  
**URL:** <https://doi.org/10.32614/RJ-2010-017>
- Steorts, R. C., Hall, R. & Fienberg, S. E. (2016), ‘A Bayesian Approach to Graphical Record Linkage and Deduplication’, *Journal of the American Statistical Association* **111**(516), 1660–1672.
- Steorts, R. C., Ventura, S. L., Sadinle, M. & Fienberg, S. E. (2014), A Comparison of Blocking Methods for Record Linkage, *in* ‘Privacy in Statistical Databases’, Lecture Notes in Computer Science, Springer, Cham, pp. 253–268.
- Tai, X. H. (2018), ‘Record linkage and matching problems in forensics’, *IEEE International Conference on Data Mining Workshops* pp. 510–517.
- Tancredi, A. & Liseo, B. (2011), ‘A Hierarchical Bayesian Approach to Record Linkage and Size Population Problems’, *The Annals of Applied Statistics* **5**(2B), 1553–1585.
- Teh, Y. W., Jordan, M. I., Beal, M. J. & Blei, D. M. (2006), ‘Hierarchical dirichlet processes’, *Journal of the American Statistical Association* **101**(476), 1566–1581.  
**URL:** <http://www.jstor.org/stable/27639773>
- Wagner, D., Lane, M. et al. (2014), The person identification validation system (pvs): Applying the center for administrative records research and applications’ (carra) record linkage software, Technical report, Center for Economic Studies, U.S. Census Bureau.
- Winkler, W. E. (1990), String Comparator Metrics and Enhanced Decision Rules in the Fellegi–Sunter Model of Record Linkage, *in* ‘Proceedings of the Section on Survey Research Methods’, American Statistical Association, pp. 354–359.
- Winkler, W. E. (1999), The State of Record Linkage and Current Research Problems, Technical report, Statistical Research Division, U.S. Bureau of the Census.
- Wortman, J. P. H. (2019), Record linkage methods with applications to causal inference and election voting data, PhD thesis, Duke University.

Zhang, Y., Zhang, F., Yao, P. & Tang, J. (2018), ‘Name disambiguation in aminer: Clustering, maintenance, and human in the loop.’, *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* pp. 1002–1011.

## 7 APPENDIX

### 7.1 Full Conditional Distributions

In this section, we provide derivations for all full conditional distributions referenced in the paper. We provide the derivation for the full set  $Z_j$  in Appendix 7.1.1, for the component  $Z_{j,k}$  in Appendix 7.1.2, for  $\boldsymbol{\eta}$  in Appendix 7.1.3, and for  $bmm$  and  $\mathbf{u}$  in Appendix 7.1.4.

#### 7.1.1 Full Conditional for the Set $Z_j$

We express the likelihood in (2) for the model assumptions of DRL as

$$\mathcal{L}(\mathbf{Z}, \mathbf{m}, \mathbf{u} \mid \boldsymbol{\Gamma}) = \prod_{i=1}^{n_A} \prod_{j=1}^{n_B} \prod_{f=1}^F \prod_{l=1}^{L_f} \left[ m_{fl}^{I(i \in Z_j)} u_{fl}^{I(i \notin Z_j)} \right]^{I(\gamma_{ij}^f = l) I_{obs}(\gamma_{ij}^f)}.$$

Let  $\Gamma_{.j} = \{\gamma_{ij} \mid i \in [n_A]\}$  be the collection of  $n_A$  comparison vectors associated with record  $B_j$ . We follow the proof technique introduced by Wortman (2019) and elaborated by Kundinger, Reiter & Steorts (2024). When  $B_j$  does not link to any record in  $A$  (such that  $|Z_j| = 0$ ) the contribution to the likelihood is simply a product of  $u$  parameters, which we will call  $c_j$ .

$$\mathbb{P}(\Gamma_{.j} \mid \mathbf{m}, \mathbf{u}, \boldsymbol{\pi}, Z_j = \emptyset) = \prod_{i=1}^{n_A} \prod_{f=1}^F \prod_{l=1}^{L_f} u_{fl}^{I(\gamma_{ij}^f = l) I_{obs}(\gamma_{ij}^f)} = c_j.$$

When  $Z_j = q = (q_1, \dots, q_k)$  for some  $|q| > 0$ , we have

$$\mathbb{P}(\Gamma_{.j} \mid \mathbf{m}, \mathbf{u}, \boldsymbol{\pi}, Z_j = q) = \prod_{i \in q} \prod_{f=1}^F \prod_{l=1}^{L_f} m_{fl}^{I(\gamma_{ij}^f = l) I_{obs}(\gamma_{ij}^f)} \prod_{i \notin q} \prod_{f=1}^F \prod_{l=1}^{L_f} u_{fl}^{I(\gamma_{ij}^f = l) I_{obs}(\gamma_{ij}^f)}.$$

We multiply and divide by the  $u$  parameters for the matching record pairs to obtain

$$\begin{aligned} \mathbb{P}(\Gamma_{.j} \mid \mathbf{m}, \mathbf{u}, \boldsymbol{\pi}, Z_j = q) &= \prod_{i \in q} \prod_{f=1}^F \prod_{l=1}^{L_f} \left( \frac{m_{fl}}{u_{fl}} \right)^{I(\gamma_{ij}^f = l) I_{obs}(\gamma_{ij}^f)} \prod_{i=1}^{n_A} \prod_{f=1}^F \prod_{l=1}^{L_f} u_{fl}^{I(\gamma_{ij}^f = l) I_{obs}(\gamma_{ij}^f)} \\ &= c_j \prod_{i \in q} w_{ij}. \end{aligned}$$

Lastly, we multiply the likelihood by the prior in (7) to obtain the posterior distribution. We have

$$\begin{aligned}
p(Z_j = q | \Gamma, \mathbf{m}, \mathbf{u}, \boldsymbol{\pi}) &= \frac{\frac{(n_A - k)! |k|!}{n_A!} \pi_k c_j \prod_{i \in q} w_{ij}}{\sum_{h \in \mathcal{Z}} \frac{(n_A - |h|)! |h|!}{n_A!} \pi_{|h|} c_j \prod_{i \in h} w_{ij}} \\
&= \frac{\frac{(n_A - k)! |k|!}{n_A!} \pi_k \prod_{i \in q} w_{ij}}{\sum_{h \in \mathcal{Z}} \frac{(n_A - |h|)! |h|!}{n_A!} \pi_{|h|} \prod_{i \in h} w_{ij}} \\
&\propto \frac{(n_A - k)! |k|!}{n_A!} \pi_k \prod_{i \in q} w_{ij}.
\end{aligned}$$

Importantly, the constant  $c_j$  is not found in the final expression because the probability mass associated with every potential value for  $Z_j$  shares the same  $c_j$ . This does not occur due to proportionality.

### 7.1.2 Full Conditional for Component $Z_{j,k}$

We now provide the derivation of the sequential sampler, following the argument presented in Section 7.1.1. Suppose  $B_j$  has been linked to  $k$  records in  $A$ . Recall that  $Z_{j,-k}^{(s)} = (Z_{j,1}^{(s)}, \dots, Z_{j,k-1}^{(s)})$  denotes the vector of records already linked to  $B_j$  in Gibbs iteration  $s$ . When  $B_j$  has no additional link in  $A$ , the contribution to the likelihood is a product of the  $u$  parameters for all remaining records. That is,

$$\mathbb{P}(\Gamma_{\cdot j} | \mathbf{m}, \mathbf{u}, \boldsymbol{\pi}, Z_{j,k}^{(s)} = \emptyset, Z_{j,-k}^{(s)} = q_{-k}) = \prod_{i \notin q_{-k}} \prod_{f=1}^F \prod_{l=1}^{L_f} u_{fl}^{I(\gamma_{ij}^f = l) I_{obs}(\gamma_{ij}^f)} = c_{Z_{j,-k}^{(s)}}.$$

When  $Z_{j,k}^{(s)} = q_k$  for some  $q_k > 0$ , we have

$$\begin{aligned}
&\mathbb{P}(\Gamma_{\cdot j} | \mathbf{m}, \mathbf{u}, \boldsymbol{\pi}, Z_{j,k}^{(s)} = q_k, Z_{j,-k}^{(s)} = q_{-k}) \\
&= \prod_{f=1}^F \prod_{l=1}^{L_f} m_{fl}^{I(\gamma_{q_k,j}^f = l) I_{obs}(\gamma_{q_k,j}^f)} \prod_{i \notin (q_{-k}, q_k)} \prod_{f=1}^F \prod_{l=1}^{L_f} u_{fl}^{I(\gamma_{ij}^f = l) I_{obs}(\gamma_{ij}^f)} \\
&= \prod_{f=1}^F \prod_{l=1}^{L_f} \left( \frac{m_{fl}}{u_{fl}} \right)^{I(\gamma_{q_k,j}^f = l) I_{obs}(\gamma_{q_k,j}^f)} \prod_{i \notin q_{-k}} \prod_{f=1}^F \prod_{l=1}^{L_f} u_{fl}^{I(\gamma_{ij}^f = l) I_{obs}(\gamma_{ij}^f)} \\
&= c_{Z_{j,-k}^{(s)}} w_{q_k,j}
\end{aligned}$$

To obtain the posterior, we multiply by the prior in (9). The posterior distribution this is given by

$$\begin{aligned}
\mathbb{P}(Z_{j,k}^{(s)} = q_k | Z_{j,k-1}^{(s)}, \eta_k, \mathbf{m}, \mathbf{u}, \gamma) \\
&= \frac{\left( \frac{\eta_k}{n_A - (k-1)} c_{Z_{j,-k}^{(s)}} w_{q_k,j} \right)^{I(q_k \in [n_A] \setminus Z_{j,-k}^{(s)})} + \left( c_{Z_{j,-k}^{(s)}} (1 - \eta_k) \right)^{I(q_k = \emptyset)}}{\frac{\eta_k}{n_A - (k-1)} c_{Z_{j,-k}^{(s)}} \sum_{i \notin Z_{j,-k}} w_{ij} + c_{Z_{j,-k}} (1 - \eta_k)} \\
&= \frac{\left( \frac{\eta_k}{n_A - (k-1)} w_{q_k,j} \right)^{I(q_k \in [n_A] \setminus Z_{j,-k}^{(s)})} + (1 - \eta_k)^{I(q_k = \emptyset)}}{\frac{\eta_k}{n_A - (k-1)} \sum_{i \notin Z_{j,-k}} w_{ij} + (1 - \eta_k)} \\
&\propto \begin{cases} \frac{\eta_k}{n_A - (k-1)} w_{q_k,j}, & q_k \in [n_A] \setminus Z_{j,-k}^{(s)}, \\ 1 - \eta_k, & q_k = \emptyset. \end{cases}
\end{aligned}$$

Importantly, the constant  $c_{Z_{j,-k}^{(s)}}$  is not found in the final expression because the probability mass associated with every potential value for  $Z_{j,k}^{(s)}$  shares the same  $c_{Z_{j,-k}^{(s)}}$ . This does not occur due to proportionality.

### 7.1.3 Sampling Match Rate $\eta_k$

Define  $n_k = \sum_{j=1}^{n_B} I(|Z_j| \geq k)$  to be the number of records in  $B$  that are linked to at least  $k$  records in  $A$  (with  $n_0 = n_B$ ). Note that step  $k$  of the sequential sampler acts on  $n_{k-1}$  records in  $B$  and identifies  $n_k$  links. As such, the full conditional for  $\eta_k$  is a standard Beta-Binomial update. For each  $\eta_k$ , we have

$$\begin{aligned}
&\mathbb{P}(\eta_k | \Gamma, \mathbf{Z}, \mathbf{m}, \mathbf{u}) \\
&\propto \mathbb{P}(\mathbf{Z} | \eta_k) \mathbb{P}(\eta_k) \\
&= \prod_{j=1}^{n_B} \mathbb{P}(Z_j | \eta_k) \mathbb{P}(\eta_k) \\
&\propto \prod_{j=1}^{n_B} \left[ \eta_k^{I(|Z_j| \geq k)} (1 - \eta_k)^{1 - I(|Z_j| \geq k)} \right]^{I(|Z_j| \geq k-1)} \eta_k^{\alpha_\eta - 1} (1 - \eta_k)^{\beta_\eta - 1} \\
&\propto \eta_k^{n_k + \alpha_\eta - 1} (1 - \eta_k)^{n_{k-1} - n_k + \beta_\eta - 1}.
\end{aligned}$$

Thus we have

$$\eta_k | \mathbf{Z}, \mathbf{m}, \mathbf{u}, \Gamma \sim \text{Beta}(n_k + \alpha_\eta, n_{k-1} - n_k + \beta_\eta).$$



#### 7.1.4 Full Conditional for $\mathbf{m}$ and $\mathbf{u}$

The  $\mathbf{m}$  and  $\mathbf{u}$  parameters are updated through standard multinomial-Dirichlet distributions. For a particular  $m_{fl}$ , we have

$$\mathcal{L}(m_{fl}|\mathbf{\Gamma}, \mathbf{u}, \mathbf{Z}, \boldsymbol{\eta}) \propto \prod_{i=1}^{n_A} \prod_{j=1}^{n_B} m_{fl}^{I(i \in Z_j)I(\gamma_{ij}^f=l)I_{obs}(\gamma_{ij}^f)} m_{fl}^{\alpha_{fl}-1} = m_{fl}^{\alpha_{fl}(\mathbf{Z})-1},$$

where  $\alpha_{fl}(\mathbf{Z}) = \alpha_{fl} + \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} I_{obs}(\gamma_{ij}^f)I(\gamma_{ij}^f=l)I(i \in Z_j)$ . Analogous procedures lead to  $\mathcal{L}(u_{fl}|\mathbf{\Gamma}, \mathbf{m}, \mathbf{Z}, \boldsymbol{\eta}) \propto u_{fl}^{\beta_{fl}(\mathbf{Z})-1}$ , where  $\beta_{fl}(\mathbf{Z}) = \beta_{fl} + \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} I_{obs}(\gamma_{ij}^f)I(\gamma_{ij}^f=l)I(i \notin Z_j)$ . Thus, for the vectors of parameters  $\mathbf{m}_f$  and  $\mathbf{u}_f$ , we have

$$\mathbf{m}_f|\mathbf{\Gamma}, \mathbf{Z}, \mathbf{u}, \boldsymbol{\eta} \sim \text{Dirichlet}(\alpha_{f1}(\mathbf{Z}), \dots, \alpha_{fL_f}(\mathbf{Z})),$$

$$\mathbf{u}_f|\mathbf{\Gamma}, \mathbf{Z}, \mathbf{m}, \boldsymbol{\eta} \sim \text{Dirichlet}(\beta_{f1}(\mathbf{Z}), \dots, \beta_{fL_f}(\mathbf{Z})),$$

$$\text{where } \alpha_{fl}(\mathbf{Z}) = \alpha_{fl} + \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} I_{obs}(\gamma_{ij}^f)I(\gamma_{ij}^f=l)I(i \in Z_j), \quad (12)$$

$$\text{and } \beta_{fl}(\mathbf{Z}) = \beta_{fl} + \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} I_{obs}(\gamma_{ij}^f)I(\gamma_{ij}^f=l)I(i \notin Z_j). \quad (13)$$

## 7.2 Hashing for Efficient Posterior Inference

Following Kunderinger, Reiter & Steorts (2024) we use hashing to reduce the computational complexity of the Gibbs sampler. Since each component  $\gamma_{ij}^f$  of the comparison vector is discrete, there are only finitely many possible realizations of the comparison vector  $\gamma_{ij}$ . Let  $P$  be the number of unique agreement patterns observed in  $\boldsymbol{\gamma}$ . This number is bounded above by  $P^* = \prod_{f=1}^F (L_f + 1)$ , where the addition of 1 to  $L_f$  for each field accounts for the possibility of missing values. This upper bound does not scale with  $n_A$  or  $n_B$ , but rather is determined by  $F$  and  $L_f$ .

We adopt a hashing function to map the  $P$  unique agreement patterns present in the comparison vectors to the integers in  $[P]$ . Let

$$h_f^{(i,j)} = I_{obs}(\gamma_{ij}^f) 2^{\gamma_{ij}^f + I(f>1) \times \sum_{e=1}^{f-1} (L_e - 1)}$$

denote a hashed value for field  $f$  for record pair  $(i, j)$ . Summing over the fields for the agreement pattern of record pair  $(i, j)$  gives the hash value  $h^{(i,j)} = \sum_{f=1}^F h_f^{(i,j)}$ . We enumerate

the unique hashed agreement patterns from 1 to  $P$ . Denote each agreement pattern as  $h_p = (h_p^1, \dots, h_p^F)$ , so that when record pair  $(i, j)$  exhibits agreement pattern  $p$ , we write  $\gamma_{ij} = h_p$ . Collect the agreement patterns as  $\mathcal{P} = \{h_p \mid p \in [P]\}$ .

Let  $e(h_p)$  denote the  $\sum_{f=1}^F L_f$  length vector in which the  $l + \sum_{k=1}^{f-1} L_k$  component is 1 when  $h_p^f = l$ , and 0 otherwise. Observe that  $e(h_p)$  represents a long version of  $h_p$ , which is useful for computational purposes. Let  $N_{p_j} = \sum_{i=1}^{n_B} I(\gamma_{ij} = h_p)$  denote the number of records in  $A$  with which record  $j$  in  $B$  has agreement pattern  $p$ . Collect these counts in  $\mathcal{N} = \{N_{p_j} \mid j \in [n_B], p \in [P]\}$ . Let  $N_p = \sum_{j=1}^{n_B} N_{p_j}$  denote the total number of record pairs with agreement pattern  $p$ . Finally, let  $r_{p_j} = \{i \in [n_A] \mid \gamma_{ij} = h_p\}$  be the set of records in  $A$  with which record  $j$  in  $B$  has agreement pattern  $p$ , and collect these sets as  $\mathcal{R} = \{r_{p_j} \mid p \in [P], j \in [n_B]\}$ .

The set  $\tilde{\Gamma} = \{\mathcal{P}, \mathcal{R}, \mathcal{N}\}$  fully characterizes the information in  $\Gamma$  at a reduced storage cost. In particular, by letting

$$m_p = \prod_{f=1}^F \prod_{l=1}^{L_f} m_{fl}^{I(h_p^f=l)I_{obs}(h_p^f)}$$

and

$$u_p = \prod_{f=1}^F \prod_{l=1}^{L_f} u_{fl}^{I(h_p^f=l)I_{obs}(h_p^f)},$$

the likelihood function in (2) can be written as

$$\mathcal{L}(\mathbf{Z}, \mathbf{m}, \mathbf{u} \mid \tilde{\Gamma}) = \prod_{j=1}^{n_B} \prod_{p=1}^P \prod_{i \in r_{p_j}} m_p^{I(i \in Z_j)} u_p^{I(i \notin Z_j)}. \quad (14)$$

Thus this hashing procedure allows for exact inference on the original model, without any approximation.

### 7.2.1 Batched Computation of Comparison Vectors

When it becomes infeasible to store all  $n_A n_B$  comparison vectors at one time, we partition  $A$  and  $B$  into  $t_1$  and  $t_2$  smaller disjoint batches, as proposed by Kunderinger, Reiter & Steorts (2024). Let  $\{A^1, \dots, A^{t_1}\}$  be a partition of  $A$  such that

$$\cup_{a=1}^{t_1} A^a = A \quad \text{and} \quad A^a \cap A^{a'} = \emptyset \quad \text{for all} \quad a \neq a'.$$

Likewise, let  $\{B^1, \dots, B^{t_2}\}$  be a partition of  $B$  such that

$$\bigcup_{b=1}^{t_2} B^b = B \quad \text{and} \quad B^b \cap B^{b'} = \emptyset \quad \text{for all} \quad b \neq b'.$$

For each  $a$  and  $b$ , we compute comparison vectors for all records in  $A^a \times B^b$ , creating the comparison matrix  $\mathbf{\Gamma}^{ab}$ .

We then perform hashing on the batch of comparison data, obtain a compressed  $\tilde{\mathbf{\Gamma}}^{ab}$ , and delete the memory intensive matrix  $\mathbf{\Gamma}^{ab}$  before continuing with the next batch of data. That is, we calculate

$$\begin{aligned} r_{p_j}^{ab} &= \{i \in [n_A] \mid \gamma_{ij} = h_p, i \in A^a, j \in B^b\}, \\ N_{p_j}^{ab} &= |r_{p_j}^{ab}|. \end{aligned}$$

We compute these quantities for each batch in parallel. Finally, summary statistics from each pairwise batch comparison are combined to recover summary statistics for the full comparison matrix  $\mathbf{\Gamma}$  through

$$\begin{aligned} r_{p_j} &= \{r_{p_j}^{ab}\}_{a=1, b=1}^{t_1, t_2} \\ \text{and } N_{p_j} &= \sum_{a=1}^{t_1} \sum_{b=1}^{t_2} N_{p_j}^{ab}. \end{aligned} \tag{15}$$

We can further reduce memory costs through storage efficient indexing (SEI) (Kundinger, Reiter & Steorts 2024). All records  $i$  in  $A$  that share agreement pattern  $p$  with record  $B_j$  have the same weight  $w_p$  (as described in Appendix 7.2). These records have the same probability to be identified as the link for record  $B_j$ . Thus, records  $i \in r_{p_j}$  with large  $N_{p_j}$  are unlikely to be sampled consistently enough to be deemed a match in the Bayes estimate. Rather than store all of these record labels, we store only a small number  $S$ . For each  $r_{p_j}^{ab}$ , we sample  $S$  indices without replacement to form  $\text{SEI}(r_{p_j}^{ab})$ . We collect these memory-reduced lists to form  $\text{SEI}(r_{p_j})$  as in (15), and collect these to form  $\text{SEI}(\mathcal{R})$ .

Lastly, when using SEI along with batching in this manner, we implicitly restrict the sequential sampler described in Section 3.1. If  $|\text{SEI}(r_{p_j})| < N_{p_j}$  after conducting SEI, there is a non-zero probability we can sample an agreement pattern in (16) and not have a corresponding record label for it when sampling in (17). Therefore, when using SEI, we must limit the amount of sequential sampling steps to some number  $K < S$ .

### 7.3 Efficient Gibbs Sampler

A computationally efficient method to sample  $\mathbf{Z}$  using the hashing described in Appendix 7.2 is provided in Appendix 7.3.1. The full conditional for  $\boldsymbol{\eta}$  is provided in Appendix 7.1.3, and full conditionals for  $\mathbf{m}$  and  $\mathbf{u}$  are provided in Appendix 7.1.4. We summarize the complete efficient Gibbs sampler in Algorithm 2.

#### 7.3.1 Efficient Sequential Sampling for Components $Z_{j,k}$

Let  $\tilde{\Gamma} = \{\mathcal{P}, \mathcal{R}, \mathcal{N}\}$  denote the set of summary statistics described in Appendix 7.2. We provide a computationally efficient implementation of the sampler in (10) by using, and updating, quantities in  $\mathcal{R}$  and  $\mathcal{N}$  as the sequential sampler progresses. Let  $r_{p_j}(k) = r_{p_j}(Z_{j,1}, \dots, Z_{j,k-1})$  be the set of records of agreement pattern  $p$  available for matching at the  $k^{th}$  matching step, and let  $N_{p_j}(k) = |r_{p_j}(k)|$  be the number of such records. We first sample among  $P + 1$  options for the agreement pattern between  $B_j$  and its potential link. Define  $r$  as an arbitrary set of records. We have

$$p\left(Z_{j,k} \in r \mid \tilde{\Gamma}, \mathbf{m}, \mathbf{u}, \eta_k\right) \propto \begin{cases} \frac{\eta_k N_{p_j}(k)}{n_A - (k-1)} w_p, & r = r_{p_j}(k); \\ 1 - \eta_k, & r = \emptyset. \end{cases} \quad (16)$$

Since all remaining records in  $A$  sharing the same agreement pattern with  $B_j$  are equally likely, we then sample among candidate records uniformly using

$$p(Z_{j,k} = q \mid Z_{j,k} \in r, \mathbf{m}, \mathbf{u}, \eta_k) = \begin{cases} \frac{1}{N_{p_j}(k)}, & r = r_{p_j}(k) \text{ and } q \in r; \\ 1, & r = \emptyset \text{ and } q = \emptyset. \end{cases} \quad (17)$$

Sampling from (16) and (17) is mathematically equivalent to sampling from the full conditional for  $Z_{j,k}$  in (10).

#### 7.3.2 Efficient Sampling for $m$ and $u$

Using the hashing techniques provided in Appendix 7.2 and likelihood as expressed in (14), we can perform these updates references in Appendix 7.1.4 in a more efficient manner. Let  $\boldsymbol{\alpha}_0 = (\alpha_{11}, \dots, \alpha_{FL_F})$  and  $\boldsymbol{\beta}_0 = (\beta_{11}, \dots, \beta_{FL_F})$  be concatenated vectors of prior parameters

---

**Algorithm 2** Efficient Gibbs Sampler for Dirichlet Record Linkage

---

**Input** : Data  $\tilde{\Gamma}$  and hyperparameters  $\alpha, \beta$ , and  $\beta_\eta$

**Initialize**:  $Z_j = \emptyset, \forall j \in [n_B]$

**for**  $s \in [S]$  **do**

**for**  $f \in [F]$  **do**

        Sample  $\mathbf{m}_f^{(s+1)} \sim \text{Dirichlet}(\alpha_{f1}(\mathbf{Z}^{(s)}), \dots, \alpha_{fL_f}(\mathbf{Z}^{(s)}))$ , where

$\alpha_{fl}(\mathbf{Z}^{(s)})$  is component  $l + \sum_{k=1}^{f-1} L_k$  of the sum  $\alpha_{fl} + \sum_{p=1}^P n_p(\mathbf{Z}^{(s)}) \times e(h_p)$

        Sample  $\mathbf{u}_f^{(s+1)} \sim \text{Dirichlet}(\beta_{f1}(\mathbf{Z}^{(s)}), \dots, \beta_{fL_f}(\mathbf{Z}^{(s)}))$ , where

$\beta_{fl}(\mathbf{Z}^{(s)})$  is component  $l + \sum_{k=1}^{f-1} L_k$  of the sum  $\beta_{fl} + \sum_{p=1}^P (N_p - n_p(\mathbf{Z}^{(s)})) \times e(h_p)$

**end**

    Set  $k = 1$

**while**  $D_{k-1}^{(s+1)} \neq \emptyset$  **do**

        Sample  $\eta_k^{(s+1)} \sim \text{Beta}\left(n_k^{(s)} + \alpha_\eta, n_{k-1}^{(s)} - n_k^{(s)} + \beta_\eta\right)$ , where

$n_k^{(s)} = \sum_{j=1}^{n_B} I(|Z_j^{(s)}| \geq k)$

**for**  $j \in D_{k-1}^{(s+1)}$  **do**

            Sample  $r^{(s+1)} \propto \begin{cases} \frac{\eta_k^{(s+1)} N_{p_j}^{(s+1)}(k)}{n_A - (k-1)} w_p^{(s+1)}, & r^{(s+1)} = r_{p_j}^{(s+1)}(k); \\ 1 - \eta_k^{(s+1)}, & r^{(s+1)} = \emptyset. \end{cases}$

            Sample  $Z_{j,k}^{(s+1)} \propto \begin{cases} \frac{1}{N_{p_j}^{(s+1)}(k)}, & r^{(s+1)} = r_{p_j}^{(s+1)}(k) \text{ and } q \in r^{(s+1)}; \\ 1, & r^{(s+1)} = \emptyset \text{ and } q = \emptyset. \end{cases}$

**end**

        Set  $k = k + 1$

**end**

**end**

**return**  $\{\mathbf{Z}^{(s)}, \mathbf{m}^{(s)}, \mathbf{u}^{(s)}, \boldsymbol{\eta}^{(s)}\}_{s=1}^S$

---

for the  $\mathbf{m}$  and  $\mathbf{u}$  distributions respectively. The terms needed for the posterior updates for the  $\mathbf{m}$  and  $\mathbf{u}$  parameters are given by the appropriate components of the vectors

$$\boldsymbol{\alpha}(\mathbf{Z}) = \boldsymbol{\alpha}_0 + \sum_{p=1}^P n_p(\mathbf{Z}) \times e(h_p), \quad (18)$$

$$\boldsymbol{\beta}(\mathbf{Z}) = \boldsymbol{\beta}_0 + \sum_{p=1}^P (N_p - n_p(\mathbf{Z})) \times e(h_p), \quad (19)$$

where  $n_p(\mathbf{Z}) = \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} I(i \in Z_j) I(\gamma_{ij} = h_p)$  is the number of matching record pairs with agreement pattern  $h_p$ . Specifically, the  $l + \sum_{k=1}^{f-1} L_k$  components of (18) and (19) provide the posterior updates for level  $l$  and field  $f$  in (12) and (13). These vectorized summations can be computed more efficiently than summing over  $n_A n_B$  record pairs for each field and agreement level.

## 7.4 Additional Simulation

We repeat the simulation from Section 4 with a larger proportion of the records in  $B$  having matches in  $A$ . Specifically, we increase the number of such records to 150, rather than 50. As shown in Figures 5 and 6, the precision for `multilink` without filtering and for `fastLink` is much better in this setting, but still is not competitive with DRL.

(I don't really thinking this appendix is necessary)

## 7.5 Simulation Traceplots

We provide traceplots for a select a representative linkage task at the high duplication rate from the simulation described in Section 4. Traceplots for the total number of links,  $\boldsymbol{\eta}$ ,  $\mathbf{m}$ , and  $\mathbf{u}$  are provided in Figures 7 through 10.

We note that the variance for  $\eta_k$  reflected in the traceplots until  $k = 7$ . This is because as  $k$  increases, the sequential sampler acts upon fewer records in  $B$ , and there are lower counts involved in the posterior update for  $\eta_k$ . Starting at  $k = 8$ , there are visible stretches of iterations in which  $\eta_k$  remains constant. This because one record in  $B$  linking to 8 or more records in a rare event, and the sequential sampler generally terminates before then.

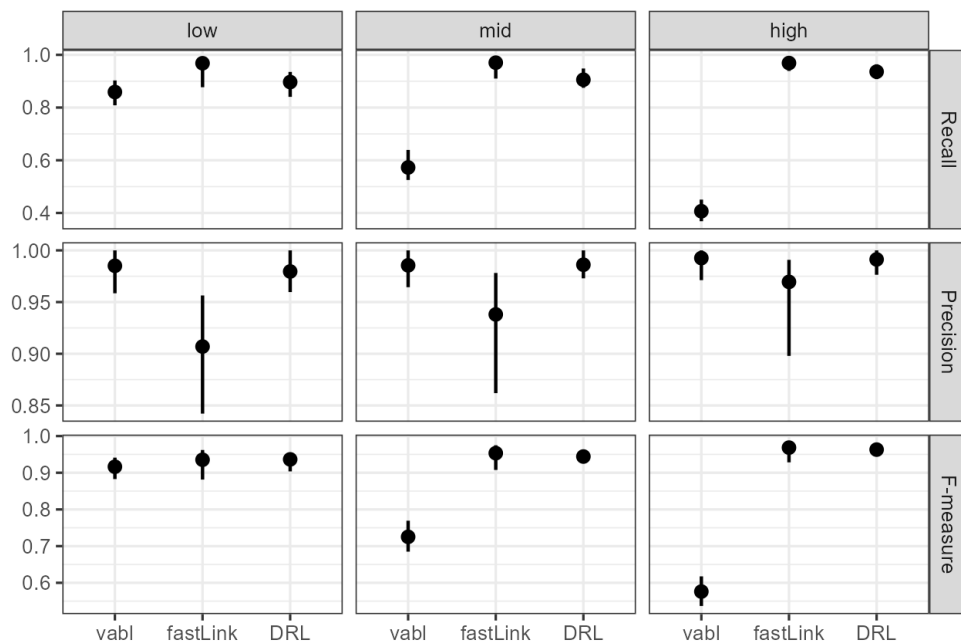


Figure 5: Results from simulation study of described in Appendix 7.4 for `vabl`, `fastLink`, and `DRL`.

(Serge’s paper only provides one traceplot, and Sadinle’s doesn’t provide any. I’d like to omit  $m$  and  $u$ , as I feel there is nothing interesting to say about them. Also, it turns out that the code I used to generate the data only altered age and occupation by adding missingness, not by adding errors (which is why the  $m$  parameters look as they do). This doesn’t invalidate the simulation, but it invites questions. I can figure out how to fix that, but it might take a while.)

## 7.6 Case Study Traceplots

(If I want to include these, I have to re-run the algorithm. I can do that, but I figured I’d send you the updated draft first.)

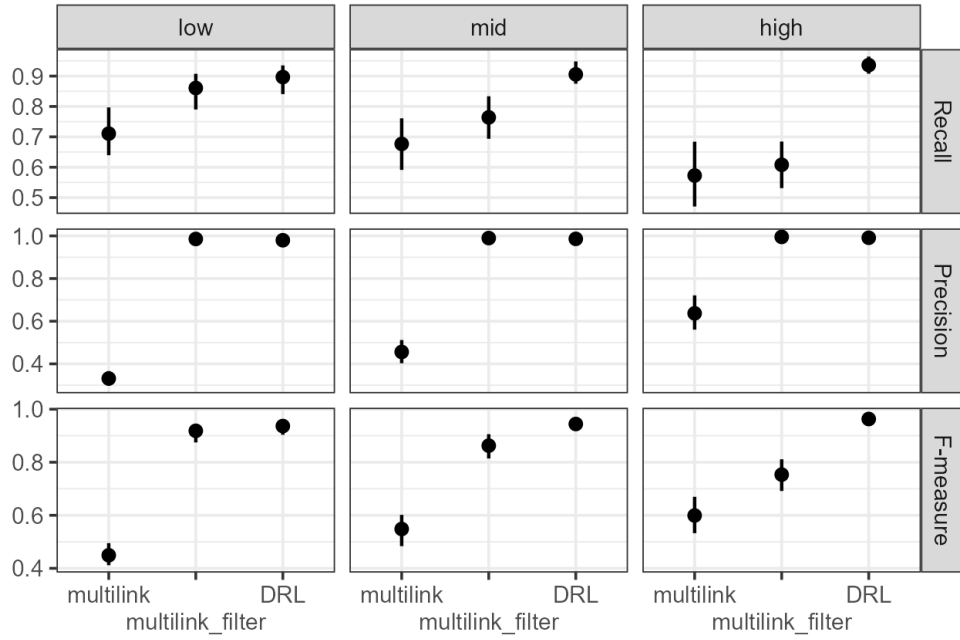


Figure 6: Results from simulation study of described in Appendix 7.4 for **multilink** (with and without filtering) and DRL.

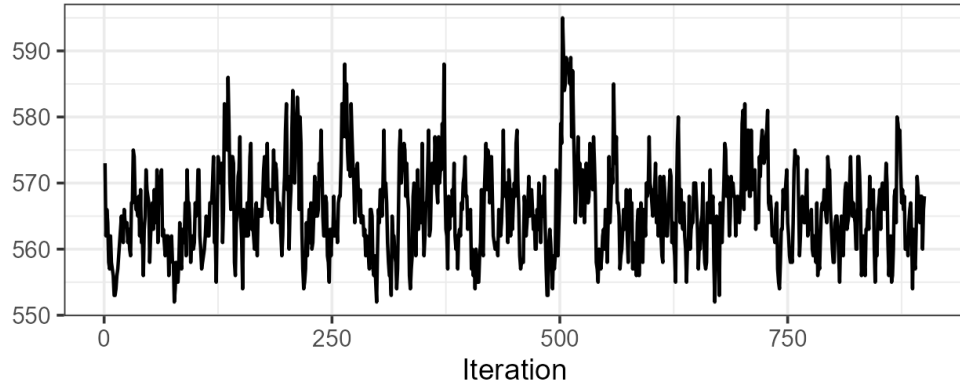


Figure 7: Traceplot for total number of matches identified under DRL for sample linkage task in simulation described in Section 4.



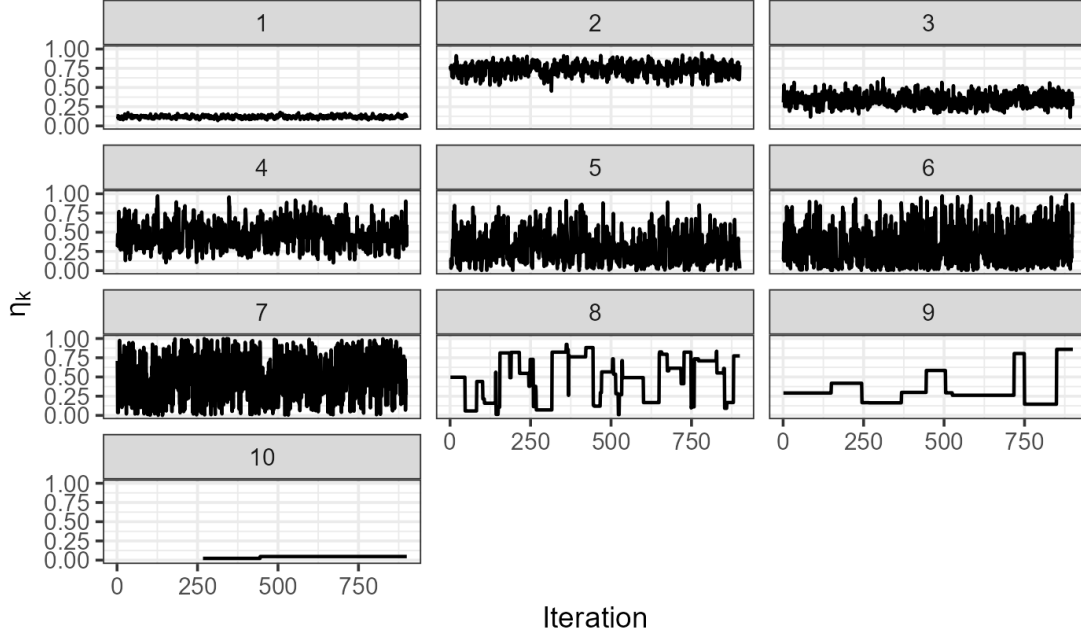


Figure 8: Traceplot for  $\eta$  under DRL for sample linkage task in simulation described in Section 4.

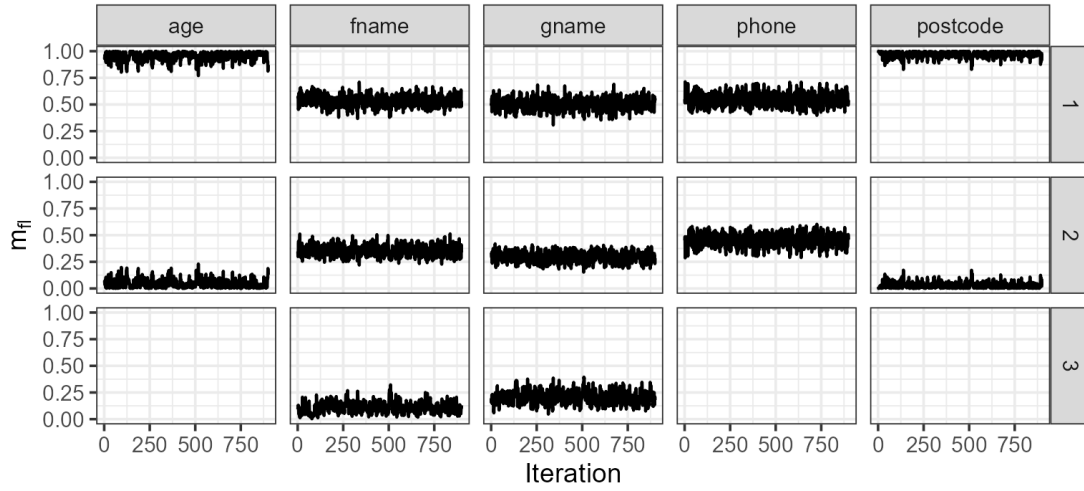


Figure 9: Traceplot for  $m$  under DRL for sample linkage task in simulation described in Section 4.

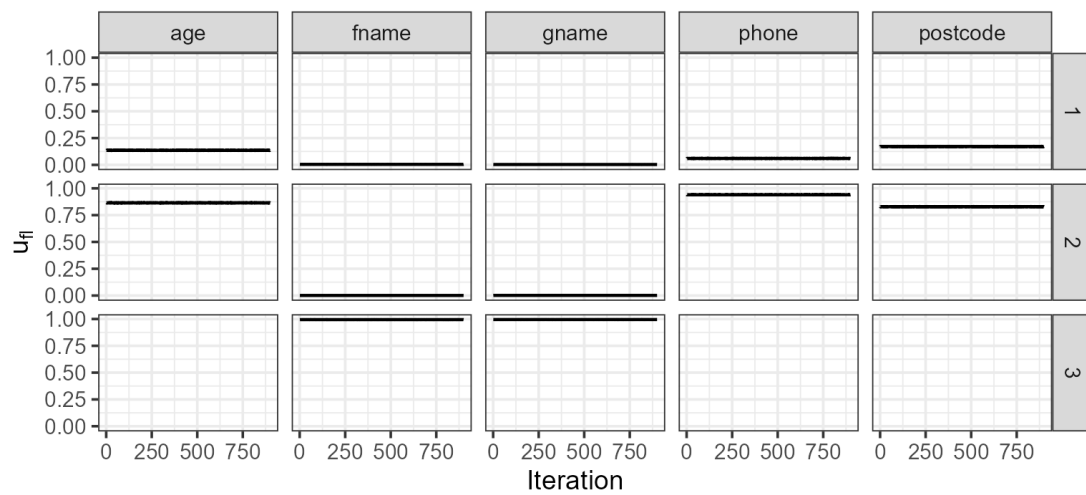


Figure 10: Traceplot for  $\mathbf{u}$  under DRL for sample linkage task in simulation described in Section 4.