Bayesian Analysis (0000) **00**, Number 0, pp. 1

Efficient and Scalable Bipartite Matching with Fast Beta Linkage (fabl) DO NOT SHARE

Anonymous Authors *

Abstract.

Keywords: bipartite record linkage, Bayesian methods, hashing techniques, parallel/distributed computing, Markov chain Monte Carlo.

1 Introduction

2 The Bipartite Record Linkage Problem

Assume two databases (data sources or data files) $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ contain n_1 and n_2 records, respectively. Without loss of generality, assume $n_1 \geq n_2$. In addition, assume duplicate records occur across databases but not within a single database, which is known as *bipartite record linkage*. The goal of this problem is to ascertain which records refer to the same entity across the two databases.

2.1 The Bipartite Record Linkage Matching

This problem can be formulated by a *bipartite matching* between two sets of records from the two databases. Within the bipartite record linkage framework, let us consider records from $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ as disjoint set of nodes (containing all the feature or attribute information). An edge between two records represents them being a *match* or *coreferent*.

One way of representing a bipartite matching is through a *coreference matrix* (or linkage structure) (Fortini et al., 2001; Larsen, 2002, 2005; Tancredi and Liseo, 2011; Sadinle, 2014; Steorts et al., 2016). The coreference matrix $\boldsymbol{\Delta}$ is a $n_1 \times n_2$ matrix whose (i, j)th entry is

$$\Delta_{ij} = \begin{cases} 1 & \text{if records } i \in \boldsymbol{X}_1 \text{ and } j \in \boldsymbol{X}_2 \text{ refer to the same entity;} \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

While this construction has been used in the literature, it is not computationally convenient. The *matching label* $\boldsymbol{Z} = (Z_1, \ldots, Z_{n_2})$ for the records in database n_2 was introduced by Sadinle (2017) such that

*Anonymous Address ,

 DOI: 0000

$$Z_j = \begin{cases} i & \text{if records } i \in \boldsymbol{X}_1 \text{ and } j \in \boldsymbol{X}_2 \text{ refer to the same entity;} \\ n_1 + j & \text{if record } j \in \boldsymbol{X}_2 \text{ does not have a match in database } \boldsymbol{X}_1. \end{cases} \tag{2}$$

The coreference matrix and the matching label are interchangeable because $\Delta_{ij} = I(Z_j = i)$, where $I(\cdot)$ is the indicator function.

2.2 Approaches to Bipartite Record Linkage

There are two main approaches to Bayesian bipartite record linkage — comparison and direct based frameworks. Direct-based approaches generatively model the field (feature or attribute) information in the databases (Gutman et al., 2013; Steorts et al., 2014; Steorts, 2015; Zanella et al., 2016; Steorts et al., 2017; Kaplan et al., 2018). On the other hand, comparison-based approaches model the field comparisons between record pairs (Sadinle, 2014, 2017; McVeigh et al., 2019).

In this paper, we propose a scalable extension of Sadinle (2017), which relaxes the bipartite extension implicitly in the prior and "corrects" for this in a post-processing step of Jaro (1989). As such, we provide the first independent updates to the matching label $\boldsymbol{Z}$, contrasting prior work of sequential updates. Furthermore, we propose additional computational speeds up (see section **??**).

3 The Fellegi-Sunter Approach

Based upon the original paper of Fellegi and Sunter (1969), we can think of the set of ordered record pairs $\boldsymbol{X}_1 \times \boldsymbol{X}_2$ as the union of the set of matches and non-matches

$$\boldsymbol{M} = \{(i,j);\ i \in \boldsymbol{X}_1,\ j \in \boldsymbol{X}_2,\ \Delta_{ij} = 1\}; \tag{3}$$

$$\boldsymbol{U} = \{(i,j);\ i \in \boldsymbol{X}_1,\ j \in \boldsymbol{X}_2,\ \Delta_{ij} = 0\}. \tag{4}$$

Record pairs that are estimated as matches are called *links*; pairs estimated as non-matches are called *non-links*. The FS approach uses pairwise comparisons of the records to estimate their match status.

3.1 Comparison Data

In record linkage applications, records that refer to the same entity should be similar. Records that refer to different entities should be dissimilar. This motivates a *comparison vector*

$$\boldsymbol{\gamma}_{ij} = (\gamma_{ij}^1, \ldots, \gamma_{ij}^f, \ldots, \gamma_{ij}^F), \tag{5}$$

where F denotes the number of criteria used to compare the records. Often the F criteria correspond to one comparison per field that the databases have in common. For instance, binary agreement is the commonly used due to its simplicity; others have used partial agreement patterns using similarity metrics or distance functions (Sadinle, 2017).

3.2 Fellegi-Sunter Model

In this section, we review the Fellegi and Sunter (FS) model of Fellegi and Sunter (1969). Note that $\boldsymbol{\gamma}_{ij}$ is a realization of a random variable $\boldsymbol{\Gamma}_{ij}$ whose distribution depends on the coreference matrix Δ_{ij} of a record pair (i, j). Let $\mathcal{M}$ and $\mathcal{U}$ denote a product of individual models for each of the comparison components under a conditional independence assumption (Winkler, 1988; Jaro, 1989) or a complex log-linear model (Larsen and Rubin, 2001). Let the comparison vector $\boldsymbol{\gamma}_{ij}$ be a realization of the random vector $\boldsymbol{\Gamma}_{ij}$ that can take distribution $\mathcal{M}(\boldsymbol{m})$ or $\mathcal{U}(\boldsymbol{u})$, depending on whether the observed record pair is a match or non-match with $\boldsymbol{m}$ and $\boldsymbol{u}$ representing vectors of the parameters. Let p denote the proportion of matches. Following Winkler (1988); Jaro (1989); Larsen and Rubin (2001) consider

$$\begin{aligned} \boldsymbol{\Gamma}_{ij} \mid \Delta_{ij} = 1 &\overset{iid}{\sim} \mathcal{M}(\boldsymbol{m}), \\ \boldsymbol{\Gamma}_{ij} \mid \Delta_{ij} = 0 &\overset{iid}{\sim} \mathcal{U}(\boldsymbol{u}), \quad \Delta_{ij} \overset{iid}{\sim} \text{Bernoulli}(p), \end{aligned} \tag{6}$$

where p denotes the proportion of matches.

The mixture model of equation 6 relies on two assumptions: the comparison vectors are independent given the bipartite matching and the matching status of the record pairs are independent of another. Both assumptions imply the comparison vectors are marginally independent. Estimation has been typically accomplished via the expectation-maximization (EM) algorithm (Dempster et al., 1977).

3.3 Beta Record Linkage Model

As such, this motivated the mixture model of Sadinle (2017), who considered the matching status of the record pairs via a bipartite matching

$$\begin{aligned} \boldsymbol{\Gamma}_{ij} \mid Z_i = i &\overset{iid}{\sim} \mathcal{M}(\boldsymbol{m}), \\ \boldsymbol{\Gamma}_{ij} \mid Z_i \neq i &\overset{iid}{\sim} \mathcal{U}(\boldsymbol{u}), \quad \boldsymbol{Z} \sim \mathcal{B}, \end{aligned} \tag{7}$$

where $\mathcal{M}(\boldsymbol{m})$ and $\mathcal{U}(\boldsymbol{u})$ are models for the comparison vectors for matches and non-matches and $\mathcal{B}$ is a prior of the space of bipartite matches such as the beta prior of Sadinle (2017). For complete details, we refer to the aforementioned paper, which inherently provides the one-to-one constraint of the bipartite matching problem. While theoretically desirable, unfortunately, the Gibbs sampling updates to the matching label $\boldsymbol{Z}$ are sequential.

4 The Fast Beta Linkage (fabl) Model

In this section, we propose the fast beta linkage method (fabl), which relaxes the bipartite assumption of Sadinle (2017) to induce a Gibbs sampler that is independent in its updates for the matching label $\boldsymbol{Z}$. This allows for much scalable computation.

4.1 Likelihood for Comparison Data

Following Sadinle (2017), the comparison fields are conditionally independent given the matching status of the record pairs, which provides the likelihood of the comparison data

$$\mathcal{L}(\boldsymbol{Z}, \Phi \mid \boldsymbol{\gamma}) = \prod_{i=1}^{n_1} \prod_{j=1}^{n_2} \prod_{f=1}^{F} \prod_{\ell=1}^{L_f} \left[m_{f\ell}^{I(Z_j=i)} u_{f\ell}^{I(Z_j \neq i)} \right]^{I(\gamma_{ij}^f = \ell)}, \tag{8}$$

where $m_{f\ell} = \mathbb{P}(\Gamma_{ij}^f = \ell \mid Z_j = i)$ denotes the probability of a match having a level ℓ of disagreement in field f and $u_{f\ell} = \mathbb{P}(\Gamma_{ij}^f = \ell \mid Z_j \neq i)$ denotes the probability of a non-match having a level ℓ of disagreement in field f. Denote $\boldsymbol{m}_f = (m_{f1}, \ldots, m_{fL_f})$, $\boldsymbol{u}_f = (u_{f1}, \ldots, u_{fL_f})$, $\boldsymbol{m} = (\boldsymbol{m}_1, \ldots, \boldsymbol{m}_F)$, $\boldsymbol{u} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_F)$, and $\Phi = (\boldsymbol{m}, \boldsymbol{u})$.

Following Sadinle (2017), we modify the model to take into account data that is missing completely at random. That is, using equation 8 and the assumption of ignorability, we marginalize over the missing comparisons to find

$$\mathcal{L}(\boldsymbol{Z}, \Phi \mid \boldsymbol{\gamma}^{\text{obs}}) = \prod_{f=1}^{F} \prod_{\ell=1}^{L_f} \left[m_{f\ell}^{\sum_{i,j} I_{\text{obs}}(\gamma_{ij}^f) I(\gamma_{ij}^f = \ell) I(Z_j = i)} \times u_{f\ell}^{\sum_{i,j} I_{\text{obs}}(\gamma_{ij}^f) I(\gamma_{ij}^f = \ell) I(Z_j \neq i)} \right] \tag{9}$$

$$= \prod_{f=1}^{F} \prod_{\ell=1}^{L_f} \left[m_{f\ell}^{a_{f\ell}(\boldsymbol{Z})} u_{f\ell}^{b_{f\ell}(\boldsymbol{Z})} \right], \tag{10}$$

where

$$a_{f\ell}(\boldsymbol{Z}) = \sum_{i,j} I_{\text{obs}}(\gamma_{ij}^f) I(\gamma_{ij}^f = \ell) I(Z_j = i) \quad \text{and} \tag{11}$$

$$b_{f\ell}(\boldsymbol{Z}) = \sum_{i,j} I_{\text{obs}}(\gamma_{ij}^f) I(\gamma_{ij}^f = \ell) I(Z_j \neq i),$$

which for a given matching label $\boldsymbol{Z}$ represents the number of matches and non-matches with observed disagreement level ℓ in comparison f and $I_{\text{obs}}(\cdot)$ is the indicator of whether its argument is observed.

Under equation 9, it is convenient to use independent, conditionally conjugate priors

$$\boldsymbol{m}_f \sim \text{Dirichlet}(\alpha_{f0}, \ldots, \alpha_{fL_f}) \tag{12}$$

$$\boldsymbol{u}_f \sim \text{Dirichlet}(\beta_{f0}, \ldots, \beta_{fL_f}), \tag{13}$$

for $f = 1, \ldots, F$.

4.2 Fast Beta Linkage Prior

While the bipartite prior of Sadinle (2017) may be theoretically desirable, we propose a *non-bipartite prior* for purely computational reasons. That is, for each Z_j, $j \in \{1, \ldots, n_1, n_1 + j\}$ assume independent draws from

$$\mathbb{P}(Z_j = i \mid \pi) = \begin{cases} \dfrac{\pi}{n_1} & \text{if } z_j \leq n_1 \\ 1 - \pi & z_j = n_1 + j. \end{cases} \tag{14}$$

Also, assume

$$\pi \overset{ind}{\sim} \text{Beta}(\alpha_\pi, \beta_\pi), \tag{15}$$

where α_π and β_π are fixed and known. As already mentioned, the independence assumption above leads to a Gibbs sampler that is independent and does not require sequential updates of the matching label $\boldsymbol{Z}$.

4.3 Gibbs Sampler

In this section, we provide the Gibbs sampler of fabl for completeness that explores the joint posterior distribution of $\boldsymbol{Z}$, Φ, and π given the observed comparison data $\boldsymbol{\gamma}_{\text{obs}}$, for the likelihood and priors presented earlier.

We start the sampler with an empty matching, such that, $Z_j^{[0]} = n_1 + j$ for all $j \in \{1, \ldots, n_2\}$. For a current value of the matching labeling $\boldsymbol{Z}^{[t]}$, we obtain the next values of the Gibbs sampler iteration $\boldsymbol{m}_f^{[t+1]} = (m_{f0}^{[t+1]}, \ldots, m_{fL_f}^{[t+1]})$, $\boldsymbol{u}_f^{[t+1]} = (u_{f0}^{[t+1]}, \ldots, u_{fL_f}^{[t+1]})$, for $f = 1, \ldots, F$, $\pi^{[t+1]}$, and $\boldsymbol{Z}^{(t+1)} = (Z_1^{(t+1)}, \ldots, Z_{n2}^{(t+1)})$ as follows:

1. For $f = 1, \ldots, F$ sample

$$\boldsymbol{m}_f^{[t+1]} \mid \boldsymbol{\gamma}_{\text{obs}}, \boldsymbol{Z}^{[t]} \sim \text{Dirichlet}(a_{f0}(\boldsymbol{Z}^{(t)}) + \alpha_{f0}, \ldots, a_{fL_f}(\boldsymbol{Z}^{(t)}) + \alpha_{fL_f})),$$

and

$$\boldsymbol{u}_f^{[t+1]} \mid \boldsymbol{\gamma}_{\text{obs}}, \boldsymbol{Z}^{[t]} \sim \text{Dirichlet}(b_{f0}(\boldsymbol{Z}^{(t)}) + \beta_{f0}, \ldots, b_{fL_f}(\boldsymbol{Z}^{(t)}) + \beta_{fL_f})).$$

Collect these updated parameters into $\Phi^{[t+1]}$, where the functions $a_{f\ell}(\cdot)$ and $b_{f\ell}(\cdot)$ were defined in equation 27.

2. Sample

$$\pi^{[t+1]} \mid \boldsymbol{\gamma}_{\text{obs}}, \boldsymbol{Z}^{(t)} \sim \text{Beta}(n_{12}(\boldsymbol{Z}^{(t)}) + \alpha_\pi, n_2 - n_{12}(\boldsymbol{Z}^{(t)}) + \beta_\pi - 1))$$

3. Sample the entries of $\boldsymbol{Z}^{[t+1]}$ as follows:

$$P(Z_j^{[t+1]} \mid \Phi, \boldsymbol{\gamma}_{obs}, \pi) \propto \begin{cases} \frac{\pi}{n_1} \exp(w_{qj}(\Phi)) & \text{if } q \leq n_1 \\ 1 - \pi & \text{if } q = n_1 + j. \end{cases} \tag{16}$$

$$P(\boldsymbol{Z}^{[t+1]} \mid \Phi, \boldsymbol{\gamma}_{obs}, \pi) \propto \prod_{j} \prod_{q} P(Z_j^{[t+1]} = q \mid \Phi, \boldsymbol{\gamma}_{obs}, \pi), \tag{17}$$

which does not depend on the update of $Z_j^{[t+1]}$. Therefore, we can update the partition independently and in parallel.

See Appendix A for derivations of the full conditional distributions.

References

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). "Maximum likelihood from incomplete data via the EM algorithm." *Journal of the royal statistical society: series B (methodological)*, 39(1): 1–22. 3

Fellegi, I. P. and Sunter, A. B. (1969). "A Theory for Record Linkage." *Journal of the American Statistical Association*, 64(328): 1183–1210. 2, 3

Fortini, M., Liseo, B., Nuccitelli, A., and Scanu, M. (2001). "On Bayesian Record Linkage." *Research in Official Statistics*, 4(1): 185–198. 1

Gutman, R., Afendulis, C. C., and Zaslavsky, A. M. (2013). "A Bayesian Procedure for File Linking to Analyze End-of-Life Medical Costs." *Journal of the American Statistical Association*, 108(501): 34–47. 2

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. 2, 3

Kaplan, A., Betancourt, B., and Steorts, R. C. (2018). "Posterior Prototyping: Bridging the Gap between Bayesian Record Linkage and Regression." *arXiv e-prints*. 2

Kundinger, B., Reiter, J., and Steorts, R. C. (2023). "Efficient and Scalable Bipartite Matching with Fast Beta Linkage (fabl)."

Larsen, M. D. (2002). "Comments on Hierarchical Bayesian Record Linkage." In *Proceedings of the Joint Statistical Meetings, Section on Survey Research Methods*, 1995–2000. The American Statistical Association. 1

— (2005). "Advances in Record Linkage Theory: Hierarchical Bayesian Record Linkage Theory." In *Proceedings of the Joint Statistical Meetings, Section on Survey Research Methods*, 3277–3284. The American Statistical Association. 1

Larsen, M. D. and Rubin, D. B. (2001). "Iterative Automated Record Linkage Using Mixture Models." *Journal of the American Statistical Association*, 96(453): 32–41. 3

McVeigh, B. S., Spahn, B. T., and 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*. 2

Sadinle, M. (2014). "Detecting duplicates in a homicide registry using a Bayesian partitioning approach." *Ann. Appl. Stat.*, 8(4): 2404–2434. 1, 2

— (2017). "Bayesian Estimation of Bipartite Matchings for Record Linkage." *Journal of the American Statistical Association*, 112(518): 600–612. 1, 2, 3, 4

Steorts, R. C. (2015). "Entity Resolution with Empirically Motivated Priors." *Bayesian Analysis*, 10(4): 849–875. 2

Steorts, R. C., Barnes, M., and Neiswanger, W. (2017). "Performance Bounds for Graphical Record Linkage." In Singh, A. and Zhu, J. (eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Pro-*

ceedings of Machine Learning Research, 298–306. Fort Lauderdale, FL, USA: PMLR. 2

Steorts, R. C., Hall, R., and Fienberg, S. E. (2014). "SMERED: A Bayesian Approach to Graphical Record Linkage and De-duplication." 33: 922–930. 2

— (2016). "A Bayesian Approach to Graphical Record Linkage and Deduplication." *Journal of the American Statistical Association*, 111(516): 1660–1672. 1

Tancredi, A. and Liseo, B. (2011). "A Hierarchical Bayesian Approach to Record Linkage and Size Population Problems." *Annals of Applied Statistics*, 5(2B): 1553–1585. 1

Winkler, W. E. (1988). "Using the EM Algorithm for Weight Computation in the Fellegi-Sunter Model of Record Linkage." *Proceedings of the Section on Survey Research Methods*, 667–671. 3

Zanella, G., Betancourt, B., Wallach, H., Miller, J., Zaidi, A., and Steorts, R. C. (2016). "Flexible Models for Microclustering with Application to Entity Resolution." In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, 1425–1433. NY, USA: Curran Associates Inc. 2

Appendix A: Full Conditionals

In this section, we provide the full conditional distributions for our Gibbs sampler in section 4.3.

First, we consider the full conditional distributions for $\boldsymbol{m}_{f\ell}$ and $\boldsymbol{u}_{f\ell}$. It follows from equations 9 and 12 that this is a standard Multinomial-Dirichlet:

$$\begin{aligned} \boldsymbol{m}_{f\ell} \mid \boldsymbol{Z}, \boldsymbol{\gamma}_{\text{obs}} &\sim \text{Dirichlet}(a_{f\ell}(\boldsymbol{Z}) + \alpha_{f\ell}) && (18) \\ &= \text{Dirichlet}(\sum_{i,j} I_{\text{obs}}(\gamma_{ij}^f) I(\gamma_{ij}^f = \ell) I(Z_j = i) + \alpha_{f\ell}) && (19) \end{aligned}$$

and

$$\begin{aligned} \boldsymbol{u}_{f\ell} \mid \boldsymbol{Z}, \boldsymbol{\gamma}_{\text{obs}} &\sim \text{Dirichlet}(b_{f\ell}(\boldsymbol{Z}) + \beta_{f\ell}) && (20) \\ &= \text{Dirichlet}(\sum_{i,j} I_{\text{obs}}(\gamma_{ij}^f) I(\gamma_{ij}^f \neq \ell) I(Z_j = i) + \beta_{f\ell}). && (21) \end{aligned}$$

Second, we consider the full conditional distribution for π. Recall that π provides the proportion of matches across the two databases subject to the restriction that the total number of links across both databases is $n_{12}(\boldsymbol{Z})$. Recall that n_2 is the larger of the two databases. This implies that the conditional distribution is

$$\begin{aligned} p(\pi \mid \boldsymbol{Z}) &\propto p(\boldsymbol{Z} \mid \pi) p(\pi) && (22) \\ &\propto \pi^{n_{12}(\boldsymbol{Z})} (1-\pi)^{n_2(\boldsymbol{Z}) - n_{12}(\boldsymbol{Z})} \times \pi^{\alpha_\pi - 1} (1-\pi)^{\beta_\pi - 1} && (23) \\ &\propto \pi^{n_{12}(\boldsymbol{Z}) + \alpha_\pi - 1} (1-\pi)^{n_2(\boldsymbol{Z}) - n_{12}(\boldsymbol{Z}) + \beta_\pi - 1} && (24) \end{aligned}$$

This implies

$$\pi \mid \boldsymbol{Z} \sim \text{Beta}(n_{12}(\boldsymbol{Z}) + \alpha_\pi, n_2(\boldsymbol{Z}) - n_{12}(\boldsymbol{Z}) + \beta_\pi).$$

Recall the full likelihood from equation 9

$$\mathcal{L}(\boldsymbol{Z}, \Phi \mid \boldsymbol{\gamma}^{\text{obs}}) = \prod_{f=1}^{F} \prod_{\ell=1}^{L_f} \left[m_{f\ell}^{\sum_{i,j} I_{\text{obs}}(\gamma_{ij}^f) I(\gamma_{ij}^f = \ell) I(Z_j = i)} \times u_{f\ell}^{\sum_{i,j} I_{\text{obs}}(\gamma_{ij}^f) I(\gamma_{ij}^f = \ell) I(Z_j \neq i)} \right] \tag{25}$$

$$= \prod_{f=1}^{F} \prod_{\ell=1}^{L_f} \left[m_{f\ell}^{a_{f\ell}(\boldsymbol{Z})} u_{f\ell}^{b_{f\ell}(\boldsymbol{Z})} \right], \tag{26}$$

where

$$a_{f\ell}(\boldsymbol{Z}) = \sum_{i,j} I_{\text{obs}}(\gamma_{ij}^f) I(\gamma_{ij}^f = \ell) I(Z_j = i) \quad \text{and} \tag{27}$$

$$b_{f\ell}(\boldsymbol{Z}) = \sum_{i,j} I_{\text{obs}}(\gamma_{ij}^f) I(\gamma_{ij}^f = \ell) I(Z_j \neq i).$$

Now, let us consider when we have a match. This corresponds when record $i \leq n_1$ and leads to the following expression:

$$\mathcal{L}_{ij}(\boldsymbol{Z}, \Phi \mid \boldsymbol{\gamma}^{\text{obs}}) = \prod_{f=1}^{F} \left[\prod_{\ell=1}^{L_f} \left(\frac{m_{f\ell}}{u_{f\ell}} \right)^{I(\gamma_{ij}^f = \ell)} \right]^{I_{\text{obs}}(\gamma_{ij}^f)} \tag{28}$$

$$= \exp \left\{ \sum_{f=1}^{F} I_{\text{obs}}(\gamma_{ij}^f) \sum_{\ell=1}^{L_f} \log \left(\frac{m_{f\ell}}{u_{f\ell}} \right) I(\gamma_{ij}^f = \ell) \right\} \tag{29}$$

$$=: \exp\{w_{ij}(\Phi)\}. \tag{30}$$

In the case of a non-match, $i = n_1 + j$, and the likelihood is proportional to a constant. This result immediately follows from Sadinle (2017) and Wortman (2019), where the later provides more explicit derivational details for each of the two cases.

It follows that

$$\mathcal{L}_{ij}(\boldsymbol{Z}, \Phi \mid \boldsymbol{\gamma}^{\text{obs}}) \propto \begin{cases} \exp\{w_{ij}(\Phi)\} & \text{if} \quad i \leq n_1 \\ 1 & \text{if} \quad i = n_1 + j \end{cases} \tag{31}$$

It follows that we can update Z_j using the following expression:

$$p_{ij}(Z_j \mid \Phi, \boldsymbol{\gamma}^{\text{obs}}, \pi) \propto P(Z_j \mid \pi) \times \mathcal{L}_{ij}(\boldsymbol{Z}, \Phi \mid \boldsymbol{\gamma}^{\text{obs}}) \propto \begin{cases} \frac{\pi}{n_1} \exp\{w_{ij}(\Phi)\} & \text{if} \quad i \leq n_1 \\ 1 - \pi & \text{if} \quad i = n_1 + j \end{cases} \tag{32}$$

$$\tag{33}$$

We sample the conditional distribution for each Z_j by normalizing the probabilities $\frac{\pi}{n_1}\exp\{w_{ij}(\Phi)\}$ in equation 32 for each $i \leq n_1$ and equal to $1-\pi$ for $i = n_1 + j$, where j is any new label.

Using these probabilities, we randomly sample Z_j conditional on $\Phi, \boldsymbol{\gamma}^{\text{obs}}, \pi$. Observe that for any i, j $(i \neq j)$ the conditional update does not depend on the matching label providing a posteriori independence. It follows that we can update the conditional distribution of the entire matching label independently:

$$p(\boldsymbol{Z} \mid \Phi, \boldsymbol{\gamma}^{\text{obs}}, \pi) \propto \prod_j \left[\prod_i p_{ij}(Z_j \mid \Phi, \boldsymbol{\gamma}^{\text{obs}}, \pi) \right]. \tag{34}$$