Entity Resolution with an Application to the El Salvadoran Conflict

Bihan Zhuang and Rebecca C. Steorts[a]

[a]Department of Statistical Science, Duke University April 19, 2019

Entity resolution (record linkage or de-duplication) is the process of removing duplicate entities in large, noisy databases. Entity resolution is made even more difficult when unique identifiers are not present and many of the observed records are subject to missing values. Furthermore, entity resolution has tradeoffs regarding assumptions of the data generation process, error rates, and computational scalability that make it a difficult task for real applications. In this paper, we are motivated to study a real data set from El Salvador, where a Truth Commission formed by the United Nations in 1992 collected data on killings that occurred during the Salvadoran civil war (1980-1991). Due to the data collection process, victims can be duplicated, as they may have been reported by different relatives, friends, or grass roots teams working in the area. Our motivation is to be able (1) to build flexible and robust models that are computationally fast, (2) to better understand what types of models are well suited for conflict data, (3) and finally provide estimates and evaluations of the number of documented identifiable deaths for our motivating data set.
Keywords: record linkage, entity resolution, de-duplication, conflict data, Bayesian methods, El Salvador

1 INTRODUCTION

Very often information about social entities is scattered across multiple databases. Combining that information into one database can result in enormous benefits for analysis, resulting in richer and more reliable conclusions. In most practical applications, however, analysts cannot simply link records across databases based on unique identifiers, such as social security numbers, either because they are not a part of some databases or are not available due to privacy concerns. In such cases, analysts need to use methods from statistical and computational science known as *entity resolution* (also called *record linkage* or *de-duplication*) to proceed with analysis. Entity resolution (ER) is not only a crucial task for social science and industrial applications, but is a challenging statistical and computational problem itself, because many databases contain errors (noise, lies, omissions, duplications, etc.), and the number of parameters to be estimated grows with the number of records (Bhattacharya and Getoor, 2006; Dai and Storkey, 2011; Christen, 2012; Gutman et al., 2013; Christen, 2008; Bilenko and Mooney, 2003; Hsu et al., 2000; McCallum and Wellner, 2004; Murphy et al., 2007; Miller et al., 2000; Sariyar and Borg, 2010; Sariyar et al., 2012; Ball, 2000;

Lum et al., 2013; Larsen, 2002, 2005; Sadinle, 2014; Jewell et al., 2013; Cohen et al., 2003). To meet present and near-future needs, entity resolution methods must be flexible and scalable to large databases; furthermore, they must be able to handle uncertainty and be easily integrated with post-linkage statistical analyses, such as logistic regression or capture recapture. All this must be done while maintaining accuracy and low error rates.

Turning to the context of an armed conflict, creating such models is incredibly challenging as typically grass roots movements, families, friends collect multiple reports on the same victims. This naturally causes duplications to occur in the data. In this paper, we study a real example from El Salvador, where a Truth Commission was formed by the United Nations in 1992. This Truth Commission collected data on killings that occurred during the Salvadoran civil war (1980 – 1991). Given the data collection process, a victim can be reported by different family and friends, and thus, one important aspect is to remove any duplications in the data in order to make it more reliable. In addition, removing such duplications allows one to estimate the number of documented identifiable deaths.

1.1 The United Nations Truth Commission for El Salvador

Between 1980 and 1991, the Republic of El Salvador witnessed a civil war between the central government, the left-wing guerrilla Farabundo Mart National Liberation Front (FMLN), and the right-wing para-military death squads. After the peace agreement in 1992, the United Nations created a Commission on the Truth (UNTC) for El Salvador, which invited members of Salvadoran society to report war-related human rights violations, which mainly focused on killings and disappearances. In order to collect such information the UNTC invited individuals through newspaper, radio, and television advertisements to come forward and testify. The UNTC opened offices through El Salvador where witnesses could provide their testimonials, and this resulted in a list of potential victims with names, date of death, and reported location.

Due to the fact that testimonials were provided to the UNTC many years after the civil war, it is expected that witnesses could not recall some of the details of the killings. In addition, some details regarding testimonials of the same individual, may contain conflicting or differing information. This is a natural characteristic of this data and leads to more noise, distortions, and missingness in the data. Furthermore, a victim can be reported multiple times, which leads to the an issue with duplication in the data. Finally, there is not unique identifiers for this data set that are thought to be reliable, which motivates the use of fully unsupervised Bayesian methods.

Our work builds off the seminal work of Sadinle (2014), where we use the same data set for consistency. We refer to Sadinle (2014) for complete details regarding the UNTC data set. The entire data set contains 5395 records. The fields (features) used in this paper are full name, date of death (year, month and day), municipality, and department. Table 1 provides an illustrative example of how duplicates can appears in the UNTC data set. Records 1, 2, and 3 in Table 1 represent an example of three duplicated records that most likely refer to the same person. This example is one example, where we may have non-coreferent decisions made by models or by humans in making decisions between pairs of records due to the nature of the data at hand. One advantage to our proposed approach is that we never look at pairwise comparisons of records. Turing to the second example in our table, records 3 and 4 agree on all the same information except on given name and family name. This illustrates potential issues that one faces regarding Hispanic names. For example, record 5 may refer to

the same person in record 4. Record 5 could have typographical and missing information as it's quite common for a given name of JULIAN ANDRES to drop a given name to JULIAN. Turning too the family name, It's also quite common for one of the family names to be dropped. Thus, RAMOS ROJAS could be shortened to RAMOS. The typographical errors are quite common as the original data was scanned using OCR, and this is the most likely reason that such errors would appear. In short, declaring records 4 and 5 as co-referent depends highly we believe on the agreement or disagreement on the fields.

Record	Given name	Family name	Year	Month	Day	Municipality
1.	JOSE	FLORES	1981	1	29	A
2.	JOSE	FLORES	1981	2	NA	A
3.	JOSE	FLORES	1981	3	20	A
4.	JULIAN ANDRES	RAMOS ROJAS	1986	8	5	B
5.	JILIAM	RMAOS	1986	8	5	B

Table 1: Illustrative example of duplicated records in the UNTC data set. Records 1 – 3 should refer to the same entity. Records 4 — 5 should refer to the same entity.

2 CURRENT APPROACHES TO ENTITY RESOLUTION

Many modern record linkage techniques can be viewed as an extension of the Fellegi-Sunter approach (FS), which computes pairwise probabilities of matching for all pairs of records using a likelihood ratio test (Fellegi and Sunter, 1969; Newcombe et al., 1959). While modern FS methods are used today, such implementations assume that only two databases can be linked and that there are no duplicates within each database (Gutman et al., 2013; Tancredi and Liseo, 2011; Larsen and Rubin, 2001; Belin and Rubin, 1995; Murray, 2016). Furthermore, such approaches are known to be quite sensitive to the choice of the threshold that the likelihood ratio test is based upon. In short, these assumptions are inadequate for many record linkage tasks. Bayesian methods have been recently utilized in record linkage due to their flexibility and exact error propagation; however, they have been limited primarily to two-database matching, issues with scalability to large databases, and model misspecification (Copas and Hilton, 1990; Gutman et al., 2013; Tancredi and Liseo, 2011; Sadinle, 2014, 2016). These contributions, while valuable, do not easily generalize to multiple databases and to duplication within databases.

The most relevant work to our proposed methodology is that of Sadinle (2014), which deals with a special case of record linkage known as duplicate detection. Duplicate detection refers to removing duplicate entities within a data file (but not across and within data files). In Sadinle (2014), the authors propose a duplicate detection approach borrowing approaches from (Fellegi and Sunter, 1969; Newcombe et al., 1959), the Bayesian literature, and the blocking literature. Blocking (filtering or indexing) is a way of reducing down the entire space of records, such that one only must compare similar records. Sadinle (2014) first reduces down the space of all records using deterministic blocking rules. Next, the authors propose a Bayesian model based upon comparison data, which is an input from the blocking stage. One benefit of this is there is a computational cost from filtering records pairs, however, there is a tremendous drawback in that there is no way to propagate the

uncertainty from the blocking mechanism in the duplicate detection step. The authors evaluate their proposed methodology for two municipalities, where ground truth is thought to be accurate. Further evaluations are performed in an entirely unsupervised fashion on the remaining municipalities.

In Steorts et al. (In press) a fully hierarchical-Bayesian approach to record linkage, using Dirichlet prior distributions over latent attributes and assuming a data distortion model. The authors derived an efficient hybrid (Metropolis-within-Gibbs) MCMC algorithm for fitting these models, SMERED. SMERED updates most of the latent variables and parameters using Gibbs sampling from conjugate conditional distributions. It updates the bipartite graph using a split-merge step, following Jain and Neal (2004). Thus, one has all the advantages of the Bayesian paradigm for both the latent entities and the linkage structure. Similar bipartite graph structures have been considered in the two-database scenario (Tancredi and Liseo, 2011; Fortini et al., 2001; Gutman et al., 2013; Sadinle, 2014, 2016; Matsakis, 2010; Larsen, 2002, 2005, 2012). The attributes of the latent entities, the number of latent entities, the edges linking records to latents, etc., all have posterior distributions, and it is easy to sample from these distributions for uncertainty quantification or error propagation. More recently, (Steorts, 2015) extended these approaches to both categorical and noisy string data using an empirically motivated prior, `blink`, which is available on `CRAN`. The authors illustrated on real and simulated data that the EB method beat supervised methods (e.g., random forests, Bayesian Adaptive Regression Trees, logistic regression) when the training data is 10 percent (or less) of the total amount of data. While SMERED and the EB method work on moderately sized data sets, there are potential limitations with scaling to industrial-sized data sets. For a review on recent developments in Bayesian methods, see Liseo and Tancredi (2013).

2.1 Overview of the Article

In this section, we provide an overview of the article. In this paper, we provide five contributions to the literature. First, we propose an extension to the `blink` model for end-to-end empirical Bayesian entity resolution (Steorts, 2015). Specifically, we provide to our knowledge the first use of subjective priors on the linkage structure for generalized entity resolution. We consider two non-parametric priors on the linkage structure, which are the Pitman-Yor Process prior and the Dirichlet Process prior. Second, we do not require any dimension reduction (such as blocking) to be applied to the data, which means that the only sources of error in our inferential methods comes from the data and the entity resolution task. Third, our extension using generalized entity resolution propagates the error of the entity resolution task exactly into our inferential task. Fourth, our method considers an application the synthetic data, where we can understand and evaluate our methodology rigorously. Finally, our method looks at the case study to the UNTC data set from a fully unsupervised point of view.

In Section 3, we review prior methodology that is used in this paper, followed by a description of our proposed methodology. In Section 4 and Section 5, we apply our proposed model to a synthetic data set. 6, we test our proposed methodology on our motivational data set from El Salvadoran Conflict. In section 7, we provide a discussion regarding our proposed work and directions for future research.

3 METHODOLOGY

In this section, we first give notation and assumptions that is used throughout the rest of the paper in Section 3.1. We then review prior work that we build upon in Section 3.2, before describing the attribute similarity measure in Section 3.3. In Section 3.4 we outline the proposed generative process of entity resolution. In Section 3.5 we describe the use of Bayesian nonparametric prior on the linkage structure. Finally, we provide the posterior distribution under our proposed model in Section 3.6.

3.1 Notation and Assumptions

Let $i \in \{1, \ldots, D\}$ index databases and $j \in \{1, \ldots, R_i\}$ index records within each database. Allow $j' \in \{1, \ldots, N\}$ index true individuals, where $N = \sum_{i=1}^{D} R_i$ without loss of generality. Our indexing allows for categorical or string-field data.[1] Given this, let $\ell \in \{1, \ldots, p_s\}$ index string-valued fields, and let $\ell \in \{p_s + 1, \ldots, p_s + p_c\}$ index categorical fields.

Using the same notation as Steorts (2015), $X_{ij\ell}$ denotes the observed value of the ℓth field for the jth record in the ith database and it is assumed to be a noisy observation of $Y_{j'\ell}$ denotes the true value of the ℓth field for the j'th latent individual. Additionally, we incorporate the possibility that some attributes $X_{ij\ell}$ may be missing at random through a corresponding observed indicator $O_{ij\ell}$. $O_{ijl} = 1$ implies that X_{ijl} is observed and $O_{ijl} = 0$ implies that X_{ijl} is missing. We also define $\boldsymbol{X}^{obs} = \{X_{ijl} : O_{ijl} = 1\}$ as the observed part and $\boldsymbol{X}^{miss} = \{X_{ijl} : O_{ijl} = 0\}$ as the missing part of $\boldsymbol{X}$. Let λ_{ij} denote the assigned latent individual to which the jth record in the ith database corresponds, i.e., $X_{ij\ell}$ and $Y_{j'\ell}$ represent the same individual if and only if $\lambda_{ij} = j'$. Finally, allow the distortion parameter to be $z_{ij\ell} = I(X_{ij\ell} \neq Y_{\lambda_{ij}\ell})$.

We next introduce notation for empirical distributions. For each $\ell \in \{1, \ldots, p_s + p_c\}$, let S_ℓ denote the set of *all* values for the ℓth field that occur anywhere in the data, i.e., $S_\ell = \{X_{ij\ell} : 1 \leq i \leq D, 1 \leq j \leq R_i\}$. Define $\alpha_\ell(v) = \frac{1}{N} \sum_{i=1}^{D} \sum_{j=1}^{R_i} I(X_{ij\ell} = v) =$ relative frequency of v in data for field ℓ.

For each $\ell \in \{1, \ldots, p_s\}$ and all possible values $v \in S_\ell$, let $F_\ell(v)$ denote the distribution defined as follows: If $W \sim F_\ell(v)$, then for every $w \in S_\ell$,

$$P(W = w) = \frac{\alpha_\ell(w) \exp[-c\, d(v, w)]}{\sum_{w \in S_\ell} \alpha_\ell(w) \exp[-c\, d(v, w)]} \propto \alpha_\ell(w) \exp[-c\, d(v, w)]\,,$$

where $d(\cdot, \cdot)$ is a string similarity measure and $c > 0$.

Remark. *F_ℓ is used to choose values proportional to their empirical frequency, while placing more weight on those that are more "similar" to w in terms of the similarity measure. This intuitively says that the distorted values are likely to be close to the truth. A more detailed discussion about the string similarity measure can be found in Section 3.2.*

For each $\ell \in \{1, \ldots, p_s + p_c\}$, let G_ℓ denote the empirical distribution of the data in the ℓth field from all records in all databases combined. In other words, if a random variable W has distribution G_ℓ, then for every $w \in S_\ell$,

$$P(W = w) = \alpha_\ell(w).$$

[1] For example, if the categorical data is thought to be reliable, we would like to avoid comparisons of gender. On the other hand, text-style data, such as name and address should be treated as strings.

Table 2: Summary of notation

Symbol	Description	Symbol	Description
$i \in 1 \dots D$	index over databases	$y_{j'\ell}$	attribute ℓ for entity j'
$j \in 1 \dots R_i$	index over records in db i	λ_{ij}	assigned entity for record j in db i
$j' \in 1 \dots N$	index over true individuals	$\beta_{i\ell}$	prob. attribute ℓ in db i is distorted
$\ell \in 1 \dots p_s + p_c$	index over attributes	a_ℓ, b_ℓ	distortion hyperparams. for attribute ℓ
$\ell \in 1 \dots p_s + p_c$	index over attributes	ϑ, σ	BNP hyperparams. for clustering
$v \in 1 \dots \lvert\mathcal{S}_\ell\rvert$	index over domain of attribute ℓ	$\mathcal{S}_\ell$	(empirical) domain of attribute ℓ
$R = \sum_i R_i$	total number of records	$\alpha_\ell(\cdot)$	distribution over domain of attribute ℓ
$X_{ij\ell}$	attribute ℓ for record j in table i	$\texttt{sim}_\ell(\cdot,\cdot)$	similarity measure for attribute ℓ
$z_{ij\ell}$	distortion indicator for $X_{ij\ell}$	$O_{ij\ell}$	observed indicator for $X_{ij\ell}$

Remark. *G_ℓ depends on the values of $\boldsymbol{X}$. However, the idea is that we construct G_ℓ* before *doing any computations with the model. So although G_ℓ "depends on" $\boldsymbol{X}$ when we construct it, we don't treat G_ℓ as if it depends on $\boldsymbol{X}$ when we plug it into the model. We construct G_ℓ using $\boldsymbol{X}$, but then we "forget" where G_ℓ came from when we use it in the model. This is exactly the same thing that happens—conceptually—in any empirical Bayesian procedure. As usual, let $\delta(v)$ denote the distribution of a random variable that takes the value v with probability* 1.

3.2 Background on Empirical Bayesian Entity Resolution

We review the end-to-end entity resolution framework of Steorts (2015) that lays the foundation. Assuming the notation defined above, the generative model can be written as:

$$
\begin{aligned}
X_{ij\ell} \mid \lambda_{ij},\, Y_{\lambda_{ij}\ell},\, z_{ij\ell} \;&\sim \begin{cases} \delta(Y_{\lambda_{ij}\ell}) & \text{if } z_{ij\ell} = 0 \\ F_\ell(Y_{\lambda_{ij}\ell}) & \text{if } z_{ij\ell} = 1 \text{ and } \ell \le p_s \\ G_\ell & \text{if } z_{ij\ell} = 1 \text{ and } \ell > p_s \end{cases} \\
Y_{j'\ell} \;&\sim G_\ell \\
z_{ij\ell} \mid \beta_{i\ell} \;&\sim \text{Bernoulli}(\beta_{i\ell}) \\
\beta_{i\ell} \;&\sim \text{Beta}(a, b) \\
\lambda_{ij} \;&\sim \text{DiscreteUniform}(1, \dots, N)
\end{aligned}
$$

with everything independent of everything else. Since duplication is allowed within databases, any record can correspond to any latent individual. Hence, we can specify the prior the linkage structure by specifying it independently for each λ_{ij} as shown above.

We now give the joint posterior and full conditionals. For each $v \in S_\ell$, define

$$
h_\ell(v) = \left\{ \sum_{w \in S_\ell} \exp[-c\, d(v, w)] \right\}^{-1},
$$

i.e., $h_\ell(v)$ is the normalizing constant for the distribution $F_\ell(v)$. We can compute $h_\ell(v)$ in advance for each possible $v \in S_\ell$. After some simplification, the joint posterior of Steorts

(2015) becomes (where the full conditional distributions are derived in Appendix A):

$$
\begin{aligned}
&\pi(\boldsymbol{\lambda}, \boldsymbol{Y}, \boldsymbol{z}, \boldsymbol{\beta} \mid \boldsymbol{X}) \\
&\propto \prod_{i=1}^{D} \prod_{j=1}^{R_i} \left\{ \left[\prod_{\substack{\ell=1 \\ z_{ij\ell}=1}}^{p_s+p_c} \alpha_\ell(X_{ij\ell}) \right] \left[\prod_{\substack{\ell=1 \\ z_{ij\ell}=1}}^{p_s} h_\ell(Y_{\lambda_{ij}\ell}) \right] \exp\left[-c \sum_{\ell=1}^{p_s} z_{ij\ell}\, d(X_{ij\ell}, Y_{\lambda_{ij}\ell}) \right] \right\} \\
&\quad \times \left[\prod_{j'=1}^{N} \prod_{\ell=1}^{p_s+p_c} \alpha_\ell(Y_{j'\ell}) \right] \left[\prod_{i=1}^{D} \prod_{\ell=1}^{p_s+p_c} \beta_{i\ell}^{\sum_{j=1}^{n_i} z_{ij\ell}+a-1} (1-\beta_{i\ell})^{n_i - \sum_{j=1}^{n_i} z_{ij\ell}+b-1} \right] \\
&\quad \times I(X_{ij\ell} = Y_{\lambda_{ij}\ell} \text{ for all } i, j, \ell \text{ such that } z_{ij\ell} = 0).
\end{aligned}
$$

3.3 Attribute Similarity Measures

In this section, we review the attribute similarity measures defined by Marchant et al. (2019).

Definition (Attribute similarity measure)**.** *Let $\mathcal{V}$ be the domain of an attribute. An* attribute similarity measure *on $\mathcal{V}$ is a function $\texttt{sim} : \mathcal{V} \times \mathcal{V} \to [0, s_{\max}]$ that satisfies $0 \leq s_{\max} < \infty$ and $\texttt{sim}(v, w) = \texttt{sim}(w, v)$ for all $v, w \in \mathcal{V}$.*

These similarity measures is used to quantify the likelihood that some value v in the empirical distribution gets chosen as a distortion of the true value w. Although the parameterization of attribute similarity is different from the distance measure of Steorts (2015), Marchant et al. (2019) proved that the two parameterization is in fact equivalent, as long as the distance measure is bounded and symmetric. We refer the readers to Marchant et al. (2019) for detailed proofs of this result.

During the process of inference, these similarities for the attributes may be expensive to evaluate on-the-fly, so Marchant et al. (2019) consider caching and truncation of attribute similarities. Only similarities for pairs of values that fall above a cut-off $S_{cut;\ell}$ are being stored. This is achieved through the following truncation transformation to the raw attribute similarity $\texttt{sim}_\ell(v, w)$:

$$
\underline{\texttt{sim}}_\ell(v, w) = \max\left(0,\ \frac{\texttt{sim}_\ell(v, w) - s_{\text{cut};\ell}}{1 - s_{\text{cut};\ell}/s_{\max;\ell}}\right). \tag{1}
$$

Pairs of values not present in the cache have a truncated similarity of zero by default. We refer readers to Section 6.2 of the paper for discussions about this efficiency consideration.

In this section we also discuss what the appropriate distance functions for the UNTC data would be like. Distances such as the Levenshtein distance would perform poorly because of situations like a dropped name or a re-ordered name would imply a large edit distance. Instead, we consider the Monge-Elkan distance of Monge and Elkan (1997), a hybrid similarity measure, that seems more appropriate as it is insensitive to the variations above. We define Monge-Elkan distance as

$$
\texttt{sim}_\ell^{\text{M-E}}(A, B) = \frac{1}{|A|} \sum_{a \in A} \max_{b \in B} \texttt{sim}'_\ell(a, b) \tag{2}
$$

where A, B are attributes with several words (e.g. JOSE TITO), $a \in A, b \in B$ are words in an attribute (e.g. JOSE), and $\texttt{sim}'$ is a base similarity measure (such as the normalized edit

similarity). With this formulation, we thus compare average similarity between all existing words in the two attributes, ignoring the ordering of them. We then define the asymmetric similarity function that we use for the string attributes in the UNTC data:

$$\texttt{sim}_\ell(A, B) = \begin{cases} 0 & \text{if } |A| < |B| \\ \texttt{sim}_\ell^{\text{M-E}}(A, B) & \text{otherwise.} \end{cases}$$

If attribute A contains fewer words than attribute B, then we immediately treat A as not similar to B. Otherwise we use the Monge-Elkan distance to measure their similarity.

3.4 Model Specification

We now describe the generative process of our proposed model.

Latent entities. The model assumes a total of N latent entities whose attributes have the true values. The value of attribute ℓ from the j' latent entity is to be drawn independently from the empirical distribution:

$$Y_{j'\ell} \sim G_\ell.$$

Distortions. We draw a distortion probability for each attribute ℓ in database i assuming

$$\beta_{i\ell}|a_\ell, b_\ell \sim Beta(a_\ell, b_\ell),$$

where a_ℓ, b_ℓ are hyperparameters that we tune.

Records. We assume the records are generated one after another in an iterative fashion. Different from Steorts (2015), we no longer do so by selecting a latent entity uniformly at random. Instead, we incorporate subjective, more flexible priors on the linkage structure Λ. The generative process is described below.

(a) Draw a latent entity assignment from a Bayesian nonparametric prior. Specifically, we consider the Pitman Yor Process prior and the Dirichlet Process prior (generalized as BNP Prior here):

$$\lambda_{ij} \sim \text{BNP Prior}(\vartheta, \sigma),$$

where ϑ and σ are the hyperparameters of these two BNP priors. We provide details about the two priors in Section 3.5.

(b) For attribute ℓ of record j in database i, draw a distortion indicator $z_{ij\ell}$:

$$z_{ij\ell}|\beta_{i\ell} \sim Bernoulli(\beta_{i\ell}).$$

(c) Draw the record value $X_{ij\ell}$ from a hit-or-miss model (different from the model of Steorts (2015), we also incorporate attribute similarity measures to categorical fields):

$$X_{ij\ell}|\lambda_{ij},\, Y_{\lambda_{ij}\ell},\, z_{ij\ell} \sim (1 - z_{ij\ell})\delta(Y_{\lambda_{ij}\ell}) + z_{ij\ell}\phi(X_{ij\ell}|Y_{\lambda_{ij}\ell}),$$

where

$$\phi(X_{ij\ell} = w|Y_{\lambda_{ij}\ell}) = \frac{\alpha_\ell(w)\, \exp\big[-c\, d(w, Y_{\lambda_{ij}\ell})\big]}{\sum_{w \in S_\ell} \alpha_\ell(w)\, \exp\big[-c\, d(w, Y_{\lambda_{ij}\ell})\big]} \propto \alpha_\ell(w)\, \exp\big[-c\, d(w, Y_{\lambda_{ij}\ell})\big]$$

for all attributes string-valued and categorical. If $z_{ij\ell} = 0$ or no distortion, then the value of $X_{ij\ell}$ is exactly that of the corresponding latent entity. If $z_{ij\ell} = 1$ or distortion, $X_{ij\ell}$ is then drawn from a weighted empirical distribution, with similarity measures. Equivalently, given the result of Marchant et al. (2019), we can write

$$\phi(X_{ij\ell} = w | Y_{\lambda_{ij}\ell}) = \frac{\alpha_\ell(w)\, \exp\big[\texttt{sim}_\ell(w, Y_{\lambda_{ij}\ell})\big]}{\sum_{w \in S_\ell} \alpha_\ell(w)\, \exp\big[\texttt{sim}_\ell(w, Y_{\lambda_{ij}\ell})\big]} \propto \alpha_\ell(w)\, \exp\big[\texttt{sim}_\ell(w, Y_{\lambda_{ij}\ell})\big]\,.$$

3.5 Entity Resolution with the Bayesian Nonparametric Priors

The empirical Bayesian approach (EB) of (Steorts, 2015) uses a uniform prior to model the prior distribution of the linkage structure $\mathbf{\Lambda}$. The uniform prior assumes that every legitimate configuration of the λ_{ij} is equally likely a priori, and this implies a default prior on related quantities, such as the number of individuals in the data (Steorts et al., 2016). Moreover, each record is assumed to be equally likely to correspond to any of the N possible latent individuals a priori. While the choice of uniform prior is convenient and simplifies the computation of the posterior, there are several weaknesses that should be addressed. Firstly, the uniform prior is constructed under the assumption that the N total records are randomly sampled with replacement from a population of N total latent individuals. This turns out to be quite a strong assumption on the linkage structure. We assume that the total number of latent individuals has the same size as the sample. We also restrict the latent "population" size to be maximum N, and not considering the case that the latent "population" size being greater than N. Secondly, (Steorts et al., 2016) showed that even though the uniform prior is often regarded as an "non-informative" prior, it is actually highly informative in the EB model because under certain conditions the data will not be able to overwhelm the prior, which defeats the purpose of developing a Bayesian model.

For the above reasons, we consider more well-principled, subjective priors for the linkage prior. More specifically, we consider the Pitman-Yor prior (PYP). We first present notation that is used throughout the remainder of the paper and then derive the full conditional distributions. In terms of inference, we implement a standard Gibbs sampler. (We also consider the case of the Dirichlet process prior in our experiments given that this prior is a special case of the PYP).

PYP prior. The PYP prior is adapted from (Pitman, 2006). Assume a total of D databases. Assume that $\lambda_{1,1}, ..., \lambda_{i,j-1}$ are already classified into $k_{i,j-1}$ clusters identified by the population labels $j'_1, ..., j'_{k_{i,j-1}}$. The clusters have sizes $n_1, ..., n_{k_{i,j-1}}$ respectively. In our context this means λ's that belong to the same cluster *coreference* the same latent entity. Let $N_{i,j-1}$ denote the total number of these records. We now consider the classification of λ_{ij}, the label of the latent individual to which the jth record in the ith database corresponds. Recall that the PYP has three parameters, a *concentration* parameter ϑ, a *discount* parameter σ, and a base distribution H_0. Under the PYP prior, λ_{ij} will either identify a new cluster with probability

$$P(\lambda_{ij} \sim H_0 | \lambda_{1,1}, ..., \lambda_{i,j-1}, \vartheta, \sigma, H_0) = \frac{k_{i,j-1}\sigma + \vartheta}{N_{i,j-1} + \vartheta},$$

or identify with an existing cluster with probability

$$P(\lambda_{ij} = j'_g \in \{j'_1, ..., j'_{k_{i,j-1}}\} | \lambda_{1,1}, ..., \lambda_{i,j-1}, \vartheta, \sigma, H_0) = \frac{n_g - \sigma}{N_{i,j-1} + \vartheta},$$

where the admissible values for the parameters are $\sigma \in [0, 1)$ with $\vartheta > -\sigma$ or $\sigma < 0$ with $\vartheta = m|\sigma|$ for some positive integer m. Together ϑ and σ control the formation of new cluster. The discount parameter σ reduces the probability of adding a new record into the existing cluster. The PYP prior yields power-law behavior in terms of cluster behavior when $0 < \sigma < 1$. In addition, there is an obvious characteristic of the PYP prior, which is that the probability of a new record joining an existing cluster is proportional to the size of that cluster. So new records are more likely to join existing large clusters rather than a new cluster. This is often referred to as the "rich-get-richer" characteristic (Wallach, 2010).

Note that under the PYP framework, we allow the latent "population" size to be greater than N, which will be more applicable to real world scenarios. In addition, the results of this process are exchangeable, meaning the order in which the λ's identify with the clusters does not affect the probability of the final distribution, which is a desirable property of non-uniform priors.

The above probabilities induce a prior on the set of all possible partitions of the N records which is

$$P(Z(\lambda) = z) = \frac{(\vartheta + \sigma)_{k-1,\sigma}}{(\vartheta + 1)_{N-1,1}} \prod_{g=1}^{k} (1 - \sigma)_{n_g - 1,1},$$

where $\{n_1, ...n_k\}$ are the cluster sizes of a particular partition z, and $x_{r,s} = x(x + s)...(x + (r - 1)s)$ (Pitman, 2006). It can also be proved that under this prior setup, the expected value of the number of clusters in partition z, $k(z)$, is

$$E(k(z)) = \sum_{i=1}^{N} \frac{(\vartheta + \sigma)_{(i-1)\uparrow}}{(\vartheta + 1)_{(i-1)\uparrow}} = \frac{\vartheta}{\sigma} \left[\frac{(\vartheta + \sigma)_{N\uparrow}}{(\vartheta)_{N\uparrow}} - 1 \right] \tag{3}$$

and the variance is

$$Var(k(z)) = \frac{\vartheta(\vartheta + \sigma)}{\sigma^2} \frac{(\vartheta + 2\sigma)_{N\uparrow}}{(\vartheta)_{N\uparrow}} - \frac{\vartheta^2}{\sigma^2} \left(\frac{(\vartheta + \sigma)_{N\uparrow}}{(\vartheta)_{N\uparrow}} \right)^2 - \frac{\vartheta}{\sigma} \frac{(\vartheta + \sigma)_{N\uparrow}}{(\vartheta)_{N\uparrow}} \tag{4}$$

with $x_{s\uparrow} = \Gamma(x + s)/\Gamma(x)$.

We use the equations of expectation and variance for prior elicitation by selecting ϑ and σ to have $E(k(z))$ equal to a rough prior guess of the number of clusters and $Var(k(z))$ equal to a specific amount of prior variability in the number of clusters.

DP prior. The Dirichlet Process (DP) is a special case of the Pitman-Yor Process when the discount parameter $\sigma = 0$. Recall the definition of Dirichlet Process:

$$G \sim DP(\vartheta, H_0)$$

if for any partition $(A_1, ..., A_k)$ of $\mathbb{X}$:

$$(G(A_1), ..., G(A_k)) \sim \text{Dirichlet}\left(\vartheta H_0(A_1), ..., \vartheta H_0(A_k)\right)$$

where H_0 is the base distribution, and ϑ is the concentration parameter. Under a DP prior, similar to the PYP prior, the predictive probability of cluster membership of all records

constructs a partition of these records sequentially. Under the DP prior, λ_{ij} will either identify a new cluster with probability (Wallach, 2010).

$$P(\lambda_{ij} \sim H_0 | \lambda_{1,1}, ..., \lambda_{i,j-1}, \vartheta, H_0) = \frac{\vartheta}{N_{i,j-1} + \vartheta},$$

or identify with an existing cluster with probability

$$P(\lambda_{ij} = j'_g \in \{j'_1, ..., j'_{k_{i,j-1}}\} | \lambda_{1,1}, ..., \lambda_{i,j-1}, \vartheta, H_0) = \frac{n_g}{N_{i,j-1} + \vartheta}.$$

3.6 Posterior Joint Distribution

For the context of our problem, we will assume that the data is missing at random (MAR), that is, $\boldsymbol{O}$ and $\boldsymbol{X}$ are statistically independent and that the distribution of $\boldsymbol{O}$ does not depend on the hyperparameters. Let $\boldsymbol{X}^{obs} = \{X_{ijl} : O_{ijl} = 1\}$ and $\boldsymbol{X}^{miss} = \{X_{ijl} : O_{ijl} = 0\}$. Let $\boldsymbol{\Theta} = \{a, b, \vartheta, \sigma\}$. We obtain the following expression by integrating out the missing attributes

$$\begin{aligned}
&p(\boldsymbol{\Lambda}, \boldsymbol{Y}, \boldsymbol{z}, \boldsymbol{\beta}, \boldsymbol{\Theta}, \boldsymbol{X}^{miss} | \boldsymbol{X}^{obs}, \boldsymbol{O}) \\
&\propto p(\boldsymbol{\beta}|\boldsymbol{\Theta}) \times p(\boldsymbol{z}|\boldsymbol{\beta}, \boldsymbol{\Theta}) \times p(\boldsymbol{\Lambda}|\boldsymbol{\Theta}) \times p(\boldsymbol{Y}) \times p(\boldsymbol{X}^{miss}|\boldsymbol{\Lambda}, \boldsymbol{Y}, \boldsymbol{z}) \times p(\boldsymbol{X}^{obs}|\boldsymbol{\Lambda}, \boldsymbol{Y}, \boldsymbol{z}) \\
&\propto \left[\prod_{i}^{D} \prod_{\ell}^{p_s+p_c} \beta_{i\ell}^{a-1} (1-\beta_{i\ell})^{b-1} \right] \times \left[\prod_{i}^{D} \prod_{j}^{R_i} \prod_{\ell}^{p_s+p_c} \beta_{i\ell}^{z_{ij\ell}} (1-\beta_{i\ell})^{1-z_{ij\ell}} \right] \times \left[\prod_{j'}^{N} \prod_{\ell}^{p_s+p_c} \alpha_\ell(Y_{j'\ell}) \right] \\
&\quad \times \left[\prod_{ij\ell \text{ s.t. } O_{ij\ell}=1} z_{ij\ell} \cdot \phi(X_{ij\ell}|Y_{\lambda_{ij}\ell}) + (1 - z_{ij\ell}) \mathbb{1}(X_{ij\ell} = Y_{\lambda_{ij}\ell}) \right] \\
&\quad \times \left[\prod_{i}^{D} \prod_{j}^{R_i} \mathbb{1}(\lambda_{ij} = \text{“new”}) \frac{k_{ij}\sigma + \vartheta}{N_{ij} + \vartheta} + \mathbb{1}(\lambda_{ij} = j'_g \in \{j'_1, ..., j'_{k_{i,j-1}}\}) \frac{n_g - \sigma}{N_{ij} + \vartheta} \right].
\end{aligned} \tag{5}$$

We now derive the full conditional distribution of $\boldsymbol{\Lambda}$ under the PYP prior:

$$\begin{aligned}
&p(\boldsymbol{\Lambda}|\boldsymbol{Y}, \boldsymbol{z}, \boldsymbol{\beta}, \boldsymbol{X}^{miss}, \boldsymbol{X}^{obs}) \\
&\propto p(\boldsymbol{\Lambda}|\boldsymbol{\Theta}) \times p(\boldsymbol{X}^{obs}|\boldsymbol{\Lambda}, \boldsymbol{Y}, \boldsymbol{z}) \\
&\propto \left[\prod_{i}^{D} \prod_{j}^{R_i} \mathbb{1}(\lambda_{ij} = \text{“new”}) \frac{k_{ij}\sigma + \vartheta}{N_{ij} + \vartheta} + \mathbb{1}(\lambda_{ij} = j'_g \in \{j'_1, ..., j'_{k_{i,j-1}}\}) \frac{n_g - \sigma}{N_{ij} + \vartheta} \right] \\
&\quad \times \left[\prod_{ij\ell \text{ s.t. } O_{ij\ell}=1} (1 - z_{ij\ell}) \delta(Y_{\lambda_{ij}\ell}) + z_{ij\ell} \cdot \phi(X_{ij\ell}|Y_{\lambda_{ij}\ell}) \right].
\end{aligned} \tag{6}$$

The full conditional distributions for other parameters remain the same as in (Steorts, 2015) and we show them in Appendix A. The full conditional distributions under the DP framework is only different from that under the PYP framework in the cluster assignment probabilities $P(\lambda_{ij} \sim H_0 | \lambda_{1,1}, ..., \lambda_{i,j-1}, \vartheta, H_0)$ and $P(\lambda_{ij} = j'_g \in \{j'_1, ..., j'_{k_{i,j-1}}\} | \lambda_{1,1}, ..., \lambda_{i,j-1}, \vartheta, H_0)$.

4 APPLICATION TO SYNTHETIC DATA

In this section, we first apply our proposed methodology to synthetic data sets before applying our methodology in section 6 to El Salvador. We describe the synthetic data, describe the evaluation metrics, and then provide our results on the synthetic data set. In our experiments on RLdata500, we compared our proposed methodology of the PYP and DP priors to the uniform prior of Steorts (2015). We elaborate on the settings of each of these below.

4.1 RLdata500 data set

The `RLdata500` synthetic data set is available in the RecordLinkage package in CRAN. It consists of 500 records, where 10 percent of the records are duplicates. That is, 450 individuals of this data set are unique entities. Features available in this data set are first and last German name and full date of birth. In addition, there are ground truth identifiers available for all records in this data set, so that we can easily ascertain the sensitivity and robustness of our proposed methodology.

4.2 Parameter Settings

We first provide all settings that are used in our experiments on RLdata500. Since we have ground truth in this particular application, we know that the true number of unique entities is 450. This helps us in choosing the hyper-parameters of the DP and the PYP distributions. As we explain below, for each prior, such as the PYP, we test several different hyper-parameter settings, and run the Gibbs sampler for each experiment.

PYP prior parameters For example, we assume that the prior mean of the latent population is 450. Next, we choose three different prior variances of the latent population: 2840, 1610, and 584. For the PYP prior, we then solve for the parameters using (3) and (4) to arrive at three sets of PYP prior hyper-parameters

$$(\vartheta, \sigma) : (0.4, 0.98), (2, 0.975), (10, 0.965).$$

Next, we chose three sets of prior hyper-parameters for the Beta distortion prior

$$(a, b) : (0.5, 50), (1, 99), (3, 97).$$

These settings all satisfy the constraints on the prior, and correspond to prior mean of 0.001, 0.01, and 0.03 for the distortion probability.[2] Finally, we choose S_{max} and S_{cut} in the string similarity measure of first name and last name to be $(20, 10)$ for the PYP prior setting. This means that we will treat similarity below 10 as zero.

DP prior parameters Since DP is a special case of the PYP, we start with the hyperparameter settings described above for the PYP. We also perform a random search over the space of concentration parameter ϑ, as well as the string similarity measure.

[2]These settings were recommended in Steorts (2015), and thus, this motivates our choice of these parameter settings here.

Uniform prior parameters For the uniform prior on the linkage structure, given the sensitivity analysis completed in Steorts (2015), we utilize the best configuration for the parameters, which is $(a, b) = (1, 99)$. We refer the reader to this paper for a full review regarding the sensitivity analysis. We also perform a random search over the string similarity measure.

4.3 Evaluation Metrics

In order to assess our model performance, we consider the following entity resolution metrics (a) precision, (b) recall, (c) posterior mean (estimated population size), (d) posterior standard error, and (e) runtime. The precision and recall are defined in the following way:

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP}) \text{ and Recall} = \text{TP}/(\text{TP} + \text{FN}),$$

where TP = # of ground truth matching pairs that are also predicted matches, FP = # of ground truth non-matching pairs that are predicted matches, and FN = # of ground truth matching pairs that are predicted non-matches.

4.4 RLdata500 Results

We now provide the experimental results of our proposed methodology using the PYP and DP priors, where we also make a comparison with the uniform prior of Steorts (2015).

PYP prior on RLdata500 First, we consider the PYP prior on the linkage structure. The best parameter setting used was $(a, b) = (0.5, 50)$, $(S_{max}, S_{cut}) = (20, 10)$, $(\vartheta, \sigma) = (2, 0.975)$, where best is determined based upon the evaluation metrics defined in section 4.3.

Figure 1 (right) illustrates the posterior density of population size under the model, with lines indicating predicted mean, true mean, and 95% credible interval. Table 3 presents evaluation metrics for all our experiments under the PYP prior. As one can observe, under the PYP prior, when the prior expectation is set to the true number of clusters (450), the recall and precision remain both above 0.9, regardless of how we set the degree of the prior variability. Turning to inference, the posterior mean was typically close to the truth, although typically over-estimating the true value. The posterior errors remain quite small.

Figure 1 (left) illustrates the number of linked latent entities versus the number of Gibbs iterations (or rather a trace plot). From the trace plot we do not see any apparent issues with convergence after 30000 iterations of the Gibbs sampler. As an additional diagnostic, Figure 2 illustrates the number of distortions for each attribute along the Gibbs sampler, and we also do not see any apparent sign of non-convergence. We notice that the number of distortions remain below 10% in the categorical fields, birth year, month, and day. On the other hand, the number of distortions are higher for the string fields, which are between 20% and 30%. This is consistent with our prior belief that string fields are more prone to distortions than categorical fields.

DP prior on RLdata500 Second, we consider the DP prior on the linkage structure, which is a special case of the PYP where the discount parameter $\sigma = 0$. The best parameter setting used was $(a, b) = (1, 99)$, $(S_{max}, S_{cut}) = (30, 10)$, and $\alpha = 12$.

Figure 3 shows the posterior density of the population size under the model, with lines indicating predicted mean, true mean, and 95% credible interval. Table 4 shows the

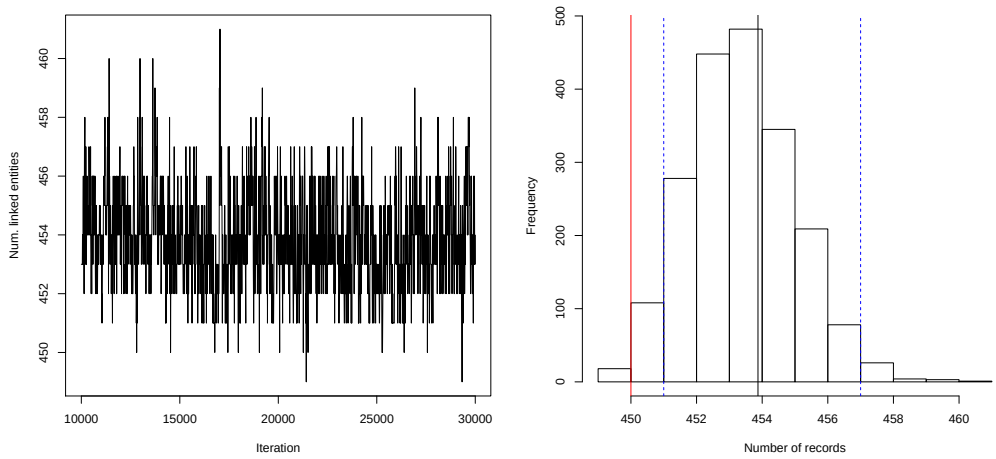

Figure 1: **PYP prior on RLdata500:** Posterior diagnostic plots for our proposed method with PYP clustering prior with $(a, b) = (0.5, 50)$, $(S_{max}, S_{cut}) = (20, 10)$, and $(\vartheta, \sigma) = (2, 0.975)$. The left plot shows the trace plot of the latent "population" size that are estimated for 30000 iterations of the Gibbs sampler. The right plot shows the posterior density of the number of distinct individuals in the sample for the `RLdata500` data set under the proposed methodology, along with the posterior mean of 453.87 (black line), true value of 450 (red line) and 95% credible interval of [451, 457] (blue dashed line).

evaluation metrics for all the experiments that we conducted under the DP prior. The model with DP prior was able to achieve perfect recall in all experiments, while the highest precision achieved was 0.909. The estimated posterior mean of the observed population was father from the truth compared to the PYP model. In addition, the DP prior consistently underestimates the observed population size, whereas the PYP overestimates the observed population size. In both cases, the posterior standard error are low.

Turning to convergence diagnostics, we did not see any signs of a lack of convergence under the DP prior after 100000 iterations of the Gibbs sampler. Note that the DP prior took longer to converge than the PYP prior. Figure 4 illustrates the number of distortions for each attribute versus the number of Gibbs iterations. We notice that the number of distortions remain below 10% in the categorical fields, birth year, month, and day. The number of distortions are still higher for the string fields, but under the DP prior, the numbers were reduced to between 15% and 22%, which is an improvement compared to under the PYP prior.

Turning to computational speed, the runtimes varied depending on the maximum similarity allowed in the Levenshtein similarity function and the threshold for truncation. Higher thresholds, which correspond to the maximum similarity can result in memory storage, which results in faster sampling, and thus, a faster runtime. In addition, if the maximum similarity is also set to a low value, the the model allows for more distortion in the sampling process, which will lead to longer run times.

Uniform prior on RLdata500 Third, we consider the uniform prior on the linkage structure, as in (Steorts, 2015). The best parameter setting used was $(a, b) = (1, 99)$ and

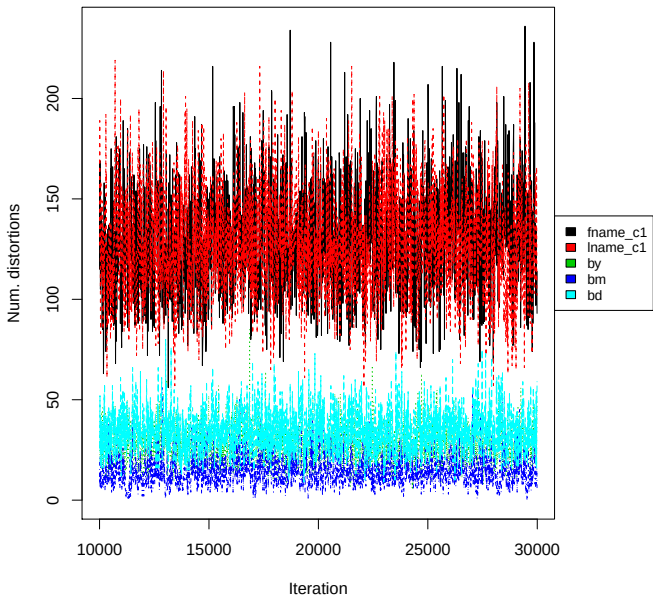

Figure 2: **PYP prior on RLdata500:** Convergence diagnostic plot for the number of distortions in each attribute along the Markov Chain.

$(S_{max}, S_{cut}) = (30, 10)$. Figure 5 shows the posterior density of the population size under the model, with lines indicating predicted mean, true mean, and 95% credible interval. Table 5 shows the evaluation metrics for all the experiments that we conducted under the uniform prior.

Turning to convergence diagnostics, we did not see any signs of a lack of convergence under the Uniform prior after 90,000 iterations of the Gibbs sampler. Notice that the Uniform prior also converged much slower than the PYP prior in the application to this data set. Figure 6 shows the number of distortions for each attribute versus the Gibbs sampler. We notice that the number of distortions remain below 10% in the categorical fields, birth year, month, and day. The number of distortions are still higher for the string fields, but under the uniform prior, the numbers were even reduced from that of the DP prior. The model with the uniform prior was able to achieve high precision and recall in all experiments. The highest precision and recall achieved was 0.909 and 1 under the best configuration. The posterior mean population was relatively close to the truth, and the model consistently underestimated the true value. Posterior standard errors remained low. Runtime again varied depending on the maximum similarity allowed in the Levenshtein similarity function and the threshold for truncation.

Table 3: **PYP prior on RLdata500**: Performance of proposed method (PYP) on `RLdata500`. 11,000 iterations of Gibbs sampler were executed.

a	b	S_{max}	S_{cut}	ϑ	σ	Precision	Recall	Posterior mean	SE	Runtime
0.5	50	20	10	0.4	0.98	1	0.94	454.53	1.77	550.79
				2	0.975	**1**	**0.98**	453.87	1.64	543.74
				10	0.965	1	0.96	453.13	1.45	548.63
1	99	20	10	0.4	0.98	1	0.94	455.45	1.77	672.91
				2	0.975	1	0.94	454.80	1.70	659.22
				10	0.965	1	0.96	453.86	1.55	664.02
3	97	20	10	0.4	0.98	1	0.94	454.64	1.76	673.06
				2	0.975	1	0.94	454.05	1.60	336.74
				10	0.965	1	0.96	453.19	1.43	342.92

Table 4: **DP prior on RLdata500**: Performance of proposed method (DP) on `RLdata500`. 100,000 iterations of Gibbs sampler were executed.

a	b	S_{max}	S_{cut}	ϑ	Precision	Recall	Posterior mean	SE	Runtime (s)
1	99	30	10	11	0.88	1	442.51	1.59	1452.44
				12	**0.909**	1	444.54	1.22	1447.15
				13	0.877	1	441.98	1.85	1457.86
		40	28	10	0.862	1	441.17	1.91	143.36
				11	0.877	1	441.84	1.59	291.34
				12	0.877	1	442.50	1.75	287.28
				13	0.893	1	442.92	2.28	294.06

Table 5: **Uniform prior**: Performance of proposed method (Uniform) on `RLdata500`. 90,000 iterations of Gibbs sampler were executed.

a	b	S_{max}	S_{cut}	Precision	Recall	Posterior mean	SE	Runtime (s)
0.5	50	30	10	0.893	1	444.74	1.47	4692.72
		40	28	0.909	1	444.43	1.47	875.15
		50	35	0.904	0.94	448.48	1.63	691.11
1	99	30	10	0.909	1	446.15	1.26	4730.12
		40	28	0.907	0.98	447.65	1.38	854.90
		50	35	0.904	0.94	450.01	1.78	678.79
3	97	30	10	0.909	1	445.59	1.35	3537.60
		40	28	0.909	1	445.82	1.47	623.52
		50	35	0.906	0.96	448.20	1.82	694.33

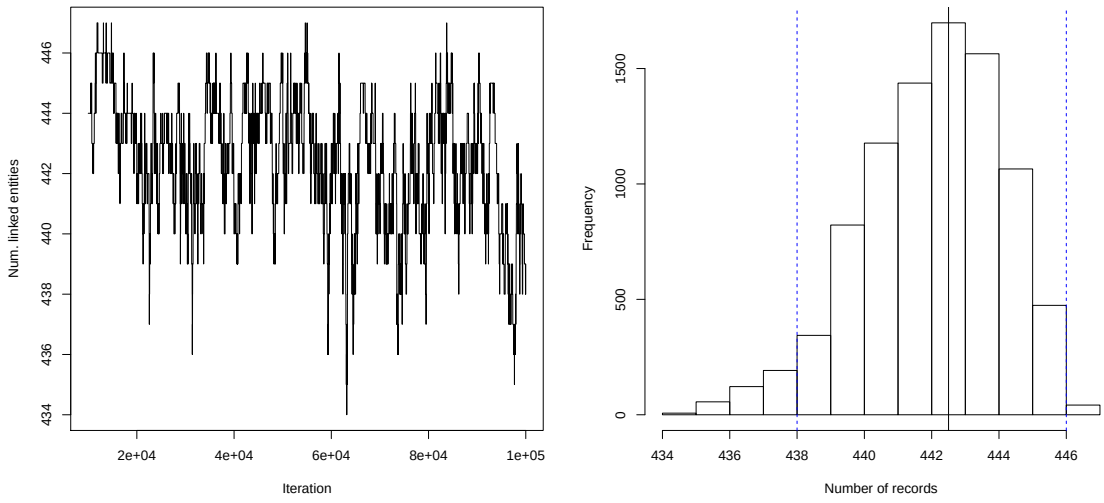

Figure 3: **DP prior on RLdata500:** Posterior diagnostic plots for our proposed method with DP clustering prior with $(a, b) = (1, 99)$, $(S_{max}, S_{cut}) = (30, 10)$, and $\alpha = 12$. The left plot shows the trace plot of the latent "population" size that are estimated for 100000 Gibbs samples for the `RLdata500` data set. The right plot shows the posterior density of the number of distinct individuals in the sample for the `RLdata500` data set under the proposed methodology, along with the posterior mean of 442.50(black line), true value of 450 (red line) and 95% credible interval of [438, 446] (blue dashed line).

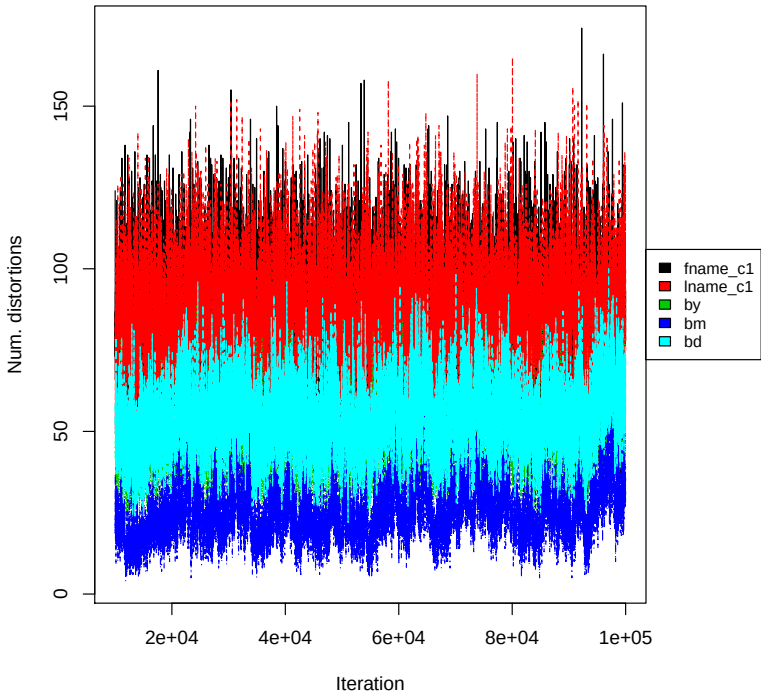

Figure 4: **DP prior on RLdata500:** Convergence diagnostic plot for the number of distortions in each attribute versus the number of Gibbs iterations.

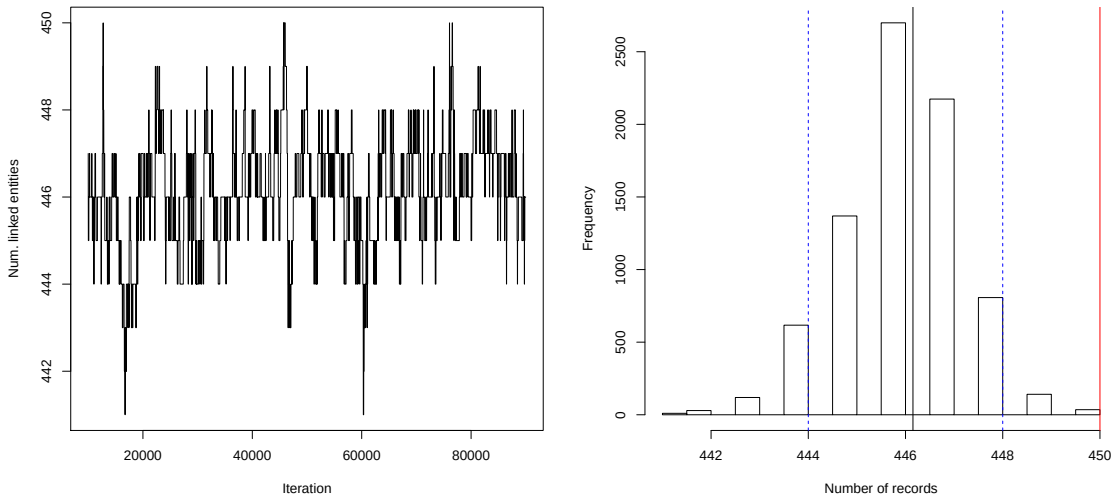

Figure 5: **Uniform prior.** Posterior diagnostic plots for our proposed method with uniform clustering prior with $(a, b) = (1, 99)$ and $(S_{max}, S_{cut}) = (30, 10)$. The left plot shows the trace plot of the latent "population" size that are estimated for 90000 Gibbs samples for the `RLdata500` data set. The right plot shows the posterior density of the number of distinct individuals in the sample for the `RLdata500` data set under the proposed methodology, along with the posterior mean of 446.15 (black line), true value of 450 (red line) and 95% credible interval of [444, 448] (blue dashed line).

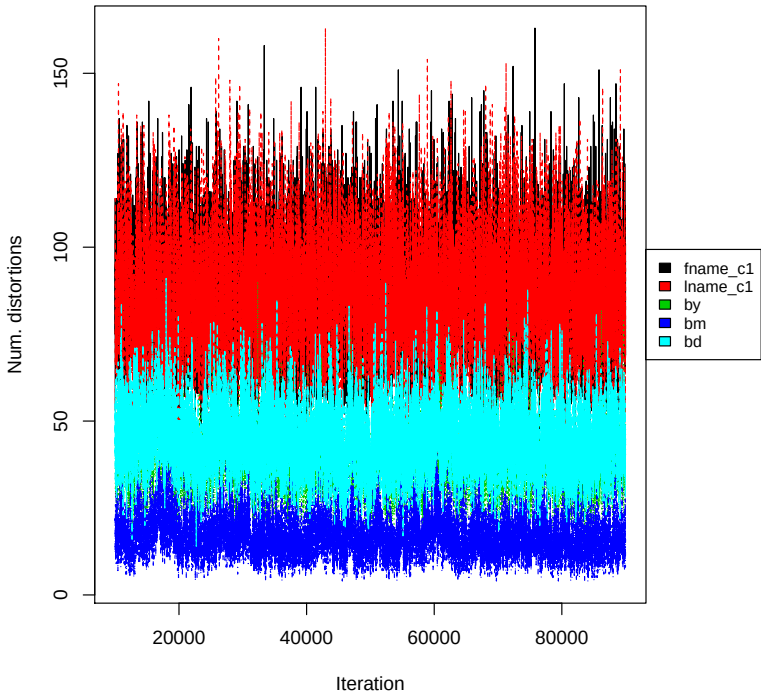

Figure 6: **Uniform prior on RLdata500:** Convergence diagnostic plot for the number of distortions in each attribute versus the number of Gibbs iterations.

5 DISCUSSION OF APPLICATION ON RLDATA500

In this section, we briefly summarize our results on the RLdata500 data set. Under the best settings described in Section 4.4, the PYP prior overall performs the best in terms of the precision (1), recall (0.98) with a 95 percent credible interval of [451, 457]. There is a tendency for the PYP to slightly overestimate the true value of number of unique records, whereas the DP and Uniform prior consistently underestimate the true value. All methods tend to have a balance between the recall and precision, which is a desired property in the field of entity resolution. In terms of computational complexity, the uniform prior is the slowest to mix, which is not unexpected given the fact that we are starting from a very unlikely configuration of the linkage structure. This results in the Gibbs sampler to need to be run for much longer compared to the PYP or DP prior. For example, if we have some idea about the true number of clusters of the existing population apriori, we are able to provide guidance regarding the two unknown parameters of the PYP prior. We are able to provide similar guidance for the DP prior. This in turn allows for faster mixing and a faster computational time of the Gibbs sampler.

6 APPLICATION TO EL SALVADOR

In this section, we consider a case study to entity resolution by applying our proposed methods to the UNTC data set from El Salvadorian Civil War, introduced in Section 1.1. The evaluation metrics in this section are the same as in Section 4.3.

6.1 Parameter Settings

We first provide all settings that are used in our experiments on the UNTC data set. There is no ground truth information available for this data set except for two departments, namely 1 and 7. Hand matching was done by Sadinle (2014) for both of these departments, and we refer to the details of this paper regarding how the hand matching was performed. Given that we do not know the true cluster size of departments 1 or 7, we will assume that the true cluster sizes are close to the posterior mean of both departments 1 and 7 reported in Sadinle (2014), which is 680 unique clusters in total. We investigate intensively the hyperparameters of prior distortion, the concentration and discount parameters of the linkage structure, as well as the attribute similarity measures using the prior knowledge and random search. We also perform random search in a "coordinate" descent fashion along each of the attributes due to the large space of similarity measures of each attribute.

6.2 UNTC Results

In this section, we provide the experimental results of our proposed methodology using the PYP and DP priors and the Uniform prior of Steorts (2015) on the UNTC data set.

PYP Prior on UNTC First, we consider the PYP prior on the linkage structure. The best parameter setting used was $(a, b) = (1, 99)$, $(\vartheta, \sigma) = (4.6017, 0.9875)$ and the similarity configuration for each attribute is listed in Table 6. Figure 7 shows the posterior density of population size under the model, with lines indicating predicted mean, true mean, and 95% credible interval. In addition, we provide evaluation metrics for all the experiments we conducted in Table 7. The model with PYP prior was able to achieve very high precision in all experiments, however the highest recall achieved was around 0.16. The posterior mean of the observed population was quite far from the truth. Under the PYP, we see that the model consistently overestimates the true value. Posterior standard error was low.

From the trace plot we assess the convergence of our Gibbs sampler under the PYP prior. We do not see any apparent sign for non-convergence after 20000 iterations. Figure 8 shows the number of distortions for each attribute along the Markov chain, and we also do not see any apparent sign of non-convergence. We notice that the number of distortions remain below 5% in the department field, is about 15-20% for firstname, lastname and month fields, and is about 30-40% for the municipality, day and year fields. This is very different from our findings on the `RLdata500` data set, where in general we observed low distortion in the categorical fields and higher distortion in the string fields. Real world data sets, such as the UNTC data set, are often much more noisy, with higher degree of distortion on average.

Turning to computational speed, we found that our model with PYP prior was still able to converge in a relatively short amount of time, even with increased data size and more complex attributes. Runtime also varied with the changing concentration and discount hyper-parameters, but the difference is not too significant.

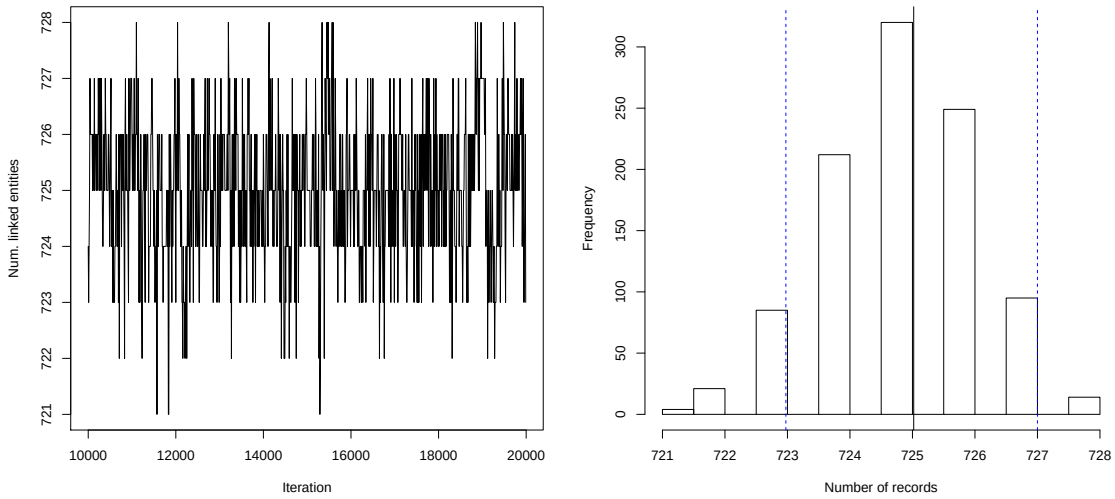

Figure 7: **PYP prior.** Posterior diagnostic plots for our proposed method with PYP clustering prior with $(a, b) = (1, 99)$, $(\vartheta, \sigma) = (4.6017, 0.9875)$, and similarity configuration for each attribute specified in Table 6. The left plot shows the trace plot of the latent "population" size that are estimated for 20000 Gibbs samples for the `UNTC` data set. The right plot shows the posterior density of the number of distinct individuals in the sample for the `UNTC` data set under the proposed methodology, along with the posterior mean of 725.02 (black line), "true" value of 680 (red line) and 95% credible interval of [723, 727] (blue dashed line).

DP Prior on UNTC Secondly, we consider the DP prior on the linkage structure, which is a special case of the PYP when the discount parameter $\sigma = 0$. The best parameter setting used was $(a, b) = (1, 99)$, $\vartheta = 2$ and the similarity configuration for each attribute listed in Table 8. As in the PYP experiment, Figure 9 shows the posterior density of population size under the model, with lines indicating predicted mean, true mean, and 95% credible interval. Table 9 shows the evaluation metrics for all the experiments that we conducted under the DP prior. The model with DP prior was able to achieve precision and recall around 0.8. The posterior mean of the observed population was very close to the "truth" of 680 we calculated Sadinle (2014). Posterior standard errors remain low.

From the trace plot we assess the convergence of our Gibbs sampler under the DP prior. We do not see any apparent sign for non-convergence after 50,000 iterations. Figure 10 shows the number of distortions for each attribute along the Markov chain, and we also do not see any apparent sign of non-convergence. We notice that the number of distortions remain below 5% in the department field, is about 15-20% for firstname, lastname and month fields, and is about 44-50% for the municipality, day and year fields.

Turning to computational speed, the runtimes varied depending on the value of ϑ, but the difference was not too significant.

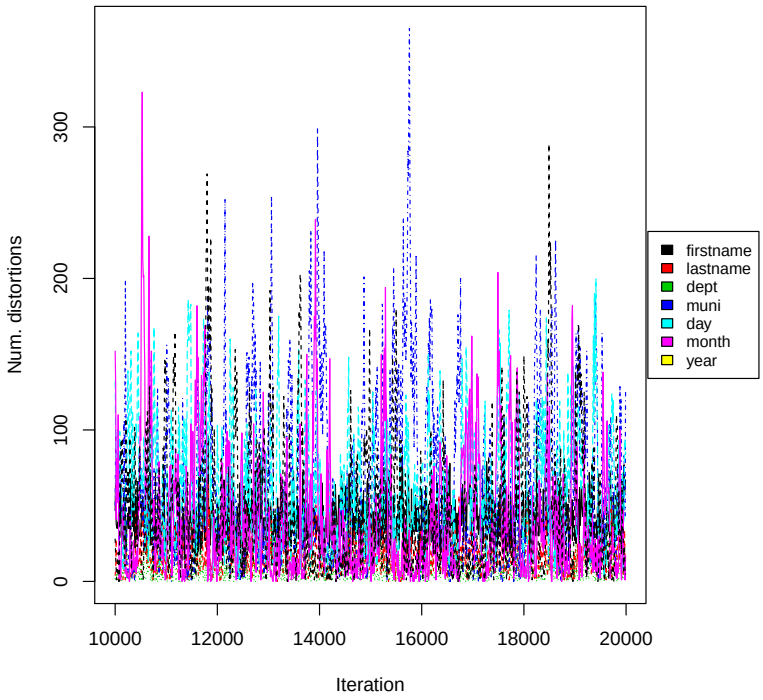

Figure 8: **PYP prior on UNTC:** Convergence diagnostic plot for the number of distortions in each attribute versus the number of Gibbs iterations.

Uniform Prior on UNTC Lastly, we consider the uniform prior on the linkage structure, as in (Steorts, 2015). The best parameter setting used was $(a, b) = (1, 73.5)$ and the similarity configuration for each attribute listed in Table 10. Figure 11 shows the posterior density of the population size under the model, with lines indicating predicted mean, true mean, and 95% credible interval. Table 11 shows the evaluation metrics for all the experiments that we conducted under the uniform prior. The best precision and recall resulted from model with uniform prior was 0.867 and 0.661. The posterior mean population was relatively close to the truth, and the model consistently overestimated the value. Posterior standard error remained low. Runtime again varied depending on the maximum similarity allowed in the Levenshtein similarity function and the threshold for truncation.

From the trace plots we assess the convergence of our Gibbs sampler under the uniform prior. We do not see any apparent sign of non-convergence after 130000 iterations. Figure 12 shows the number of distortions for each attribute along the Markov chain, and we also do not see any apparent sign of non-convergence. We notice that the number of distortions remain below 5% in the department field, is about 15-20% for firstname, lastname and month fields, and is about 44-50% for the municipality, day and year fields.

As shown in Table 11, computational speed varied depending on the maximum similarity allowed in the Levenshtein similarity function and the threshold for truncation.

Table 6: **PYP prior**: Best configuration of attribute similarity measures for `UNTC` data set. Recall that similarities below S_{cut} are effectively treated as zero (truncation) in order to speed up the sampling.

Field	S_{max}	S_{cut}
firstname	40	28
lastname	40	28
dept	20	10
muni	20	10
day	10	5
month	10	5
year	20	10

Table 7: **PYP prior**: Performance of proposed method (PYP) on `UNTC`. 20000 iterations of Gibbs sampler were executed.

a	b	ϑ	σ	Precision	Recall	Mean	SE	Runtime (s)
1	99	1.7272	0.9890	0.900	0.153	725.45	1.27	892.17
1	99	2.5663	0.9885	0.900	0.153	725.58	1.64	894.91
1	99	4.6017	0.9875	0.900	0.153	728.21	1.27	803.36
10	90	1.7272	0.9890	0.833	0.164	723.15	1.52	635.98
10	90	2.5663	0.9885	0.833	0.164	723.03	1.51	642.34
10	90	4.6017	0.9875	0.833	0.164	722.77	1.40	1156.16
3	97	2.5663	0.9885	0.889	0.131	725.38	1.45	587.36
40	60	4.6017	0.9875	0.833	0.164	720.52	1.92	731.06
40	500	2.5663	0.9885	0.778	0.115	724.13	1.55	676.96

7 DISCUSSION OF APPLICATION ON EL SALVADOR

In this section, we briefly summarize our results on the UNTC data set. Under the best setting given in Section 6.2, the DP prior overall performs the best in terms of precision (0.8), recall (0.8), with a 95 percent credible interval of [677,684]. The DP prior was able to generate estimates very tightly around the true cluster size, while the PYP and Uniform prior consistently overestimate the true value. All methods tend to have a balance between the precision and recall, which is desired in the field of entity resolution. In terms of computational complexity, the uniform prior is again the slowest to mix, as we have observed in RLdata500 experiments. Because the DP and PYP prior are more subjective and flexible than the Uniform prior, we are able to provide prior guidance in the selection of hyper-parameters, if we have some idea about the true cluster size apriori. This allows for faster mixing and faster computational time of the Gibbs sampler.

In applying our proposed method with Bayesian nonparametric priors to the UNTC data set, we make note again that the potential users of our methodology are various human rights groups who need accurate estimates and evaluations of the number of documented identifiable deaths. It is always important, therefore, to inform them about the choice of priors and the uncertainty of the Bayesian framework by nature. In terms of estimated

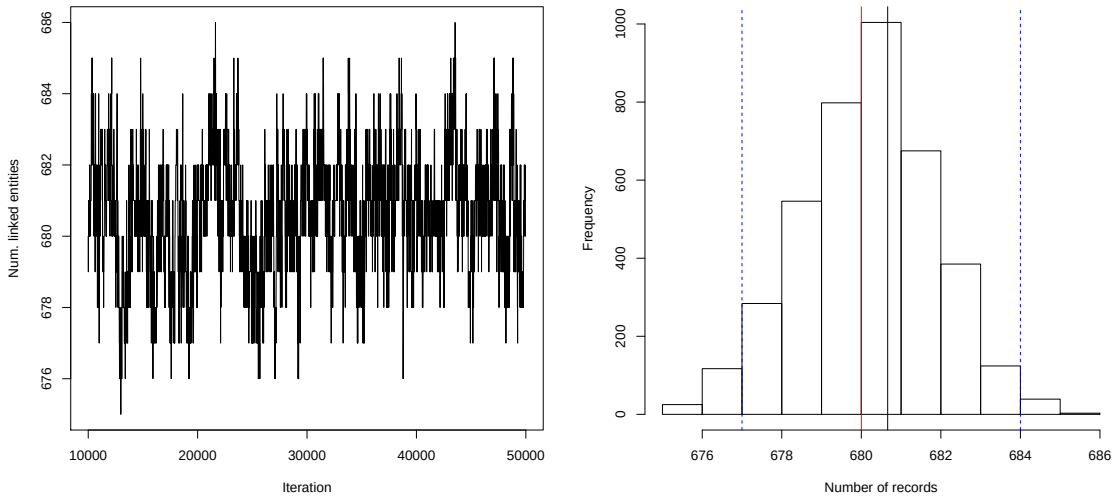

Figure 9: **DP prior.** Posterior diagnostic plots for our proposed method with DP clustering prior with $(a, b) = (1, 99)$, $\vartheta = 2$, and similarity configuration for each attribute specified in Table 8. The left plot shows the trace plot of the latent "population" size that are estimated for 50000 Gibbs samples for the `UNTC` data set. The right plot shows the posterior density of the number of distinct individuals in the sample for the `UNTC` data set under the proposed methodology, along with the posterior mean of 680.66 (black line), "true" value of 680 (red line) and 95% credible interval of [677, 684] (blue dashed line).

number of deaths and the standard error, the three priors we considered were able to produce estimates in the reasonable range, with small variation. Depending on these groups' expert knowledge of the data set, in the actual application we could choose one prior over another, after experimenting with tuning parameters and doing a thorough sensitivity analysis. In terms of applying entity resolution to a new real world conflict data set, we believe that our Bayesian model needs to be adjusted in a few ways. First, depending on the language of the country, it is likely that different similarity measures than the one we use here needs to be proposed. Secondly, different priors may have to be incorporated. And finally, the parameters of priors will have to be tuned specifically for the data set at hand.

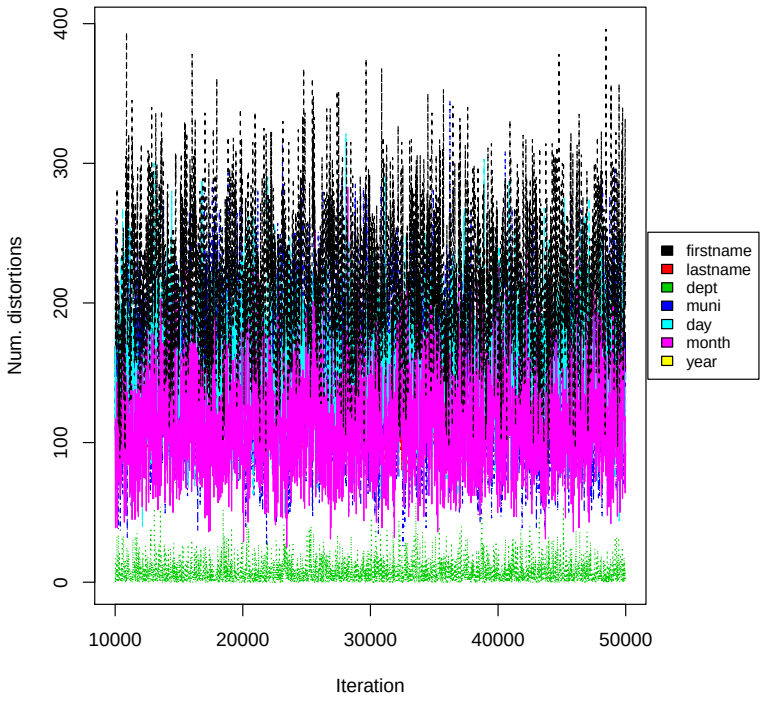

Figure 10: **DP prior on UNTC:** Convergence diagnostic plot for the number of distortions in each attribute versus the number of Gibbs iterations.

Table 8: **DP prior**: Best configuration of attribute similarity measures for `UNTC` data set. Recall that similarities below S_{cut} are effectively treated as zero (truncation) in order to speed up the sampling.

Field	S_{max}	S_{cut}
firstname	40	28
lastname	40	28
dept	20	10
muni	20	10
day	10	5
month	10	5
year	20	10

Table 9: **DP prior**: Performance of proposed method (DP) on `UNTC`. 50000 iterations of Gibbs sampler were executed.

a	b	ϑ	Precision	Recall	Posterior mean	SE	Runtime (s)
		1	0.762	0.814	675.22	2.05	2003.64
1	73.5	2	0.797	0.797	677.26	1.91	1969.29
		3	0.797	0.797	677.93	2.01	1943.78
		1	0.770	0.797	678.237	1.38	1663.35
1	99	2	0.797	0.797	680.08	1.79	1757.02
		3	0.793	0.780	682.18	1.74	1737.98

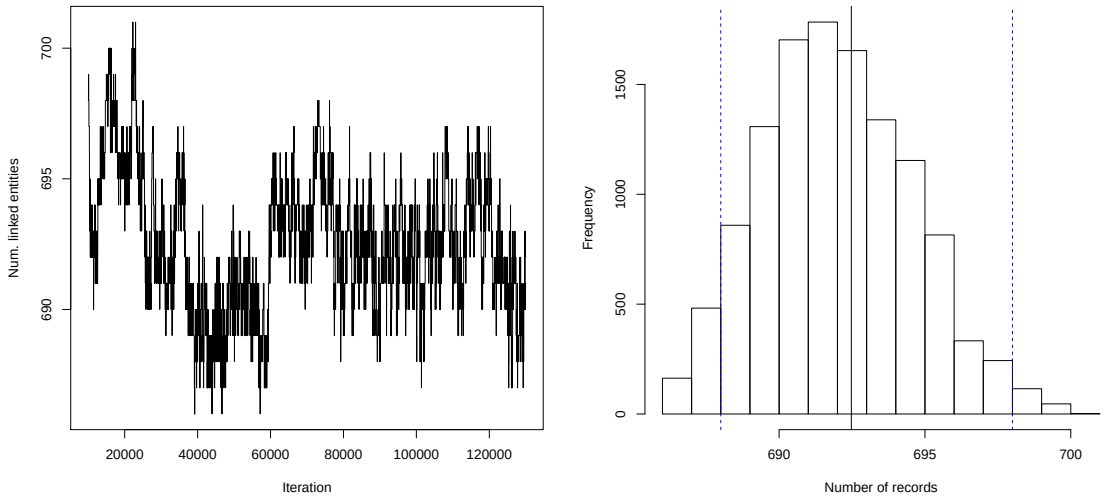

Figure 11: **Uniform prior.** Posterior diagnostic plots for our proposed method with uniform clustering prior with $(a, b) = (1, 73.5)$ and similarity configuration for each attribute specified in Table 10. The left plot shows the trace plot of the latent "population" size that are estimated for 130000 Gibbs samples for the `UNTC` data set. The right plot shows the posterior density of the number of distinct individuals in the sample for the `UNTC` data set under the proposed methodology, along with the posterior mean of 692.47 (black line), "true" value of 680 (red line) and 95% credible interval of [688, 698] (blue dashed line).

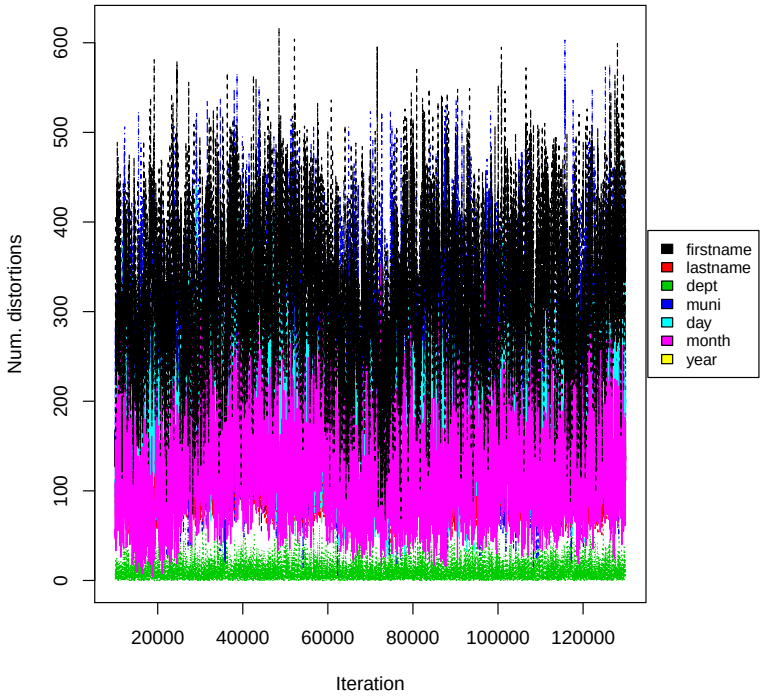

Figure 12: **Uniform prior on UNTC:** Convergence diagnostic plot for the number of distortions in each attribute versus the number of Gibbs iterations.

Table 10: **Uniform prior**: Best configuration of attribute similarity measures for `UNTC` data set. Recall that similarities below S_{cut} are effectively treated as zero (truncation) in order to speed up the sampling.

Field	S_{max}	S_{cut}
firstname	40	28
lastname	40	28
dept	20	10
muni	20	10
day	10	5
month	10	5
year	20	10

Table 11: **Uniform prior**: Performance of proposed method (Uniform) on `UNTC`. 130,000 iterations of Gibbs sampler were executed.

a	b	Precision	Recall	Posterior mean	SE	Runtime (s)
1	73.5	0.867	0.661	692.47	2.58	3490.09
1	99	0.826	0.644	688.84	2.18	3280.19

8 CONCLUSION

In this paper, we have provided five novel contributions to the literature. First, we have introduced to our knowledge, the first subjective priors (the Pitman Yor and Dirichlet Process Priors) for entity resolution with both categorical and string valued data. Second, we have introduced missing values into our model, making the model more realistic to real data situations. Third, we have derived the conditional distributions and implemented a Gibbs sampler for our proposed model. Fourth, we have illustrated the strength and weakness of our model on both synthetic and real data. For the synthetic data, our model performs better than the uniform prior, where performance is measure by standard entity resolution comparisons. For the real data (UNTC data set), our model does well with respect to inference of the underlying population here, however, the precision and recall suffer. Perhaps there may be better similarity metrics that might work to adapt to how names appear in this dataset here, however, this seems to be a difficult task. It seems a centroid or latent variable model may not be the best approach for such data, and this is still under exploration and left for future work.

REFERENCES

Ball, P. "The Salvadoran Human Rights Commission: Data Processing, Data Representation, and Generating Analytical Reports." In Ball, P., Spirer, H. F., and Spirer, L. (eds.), *Making the Case: Investigating Large Scale Human Rights Violations Using Information Systems and Data Analysis*. AAAS (2000).

Belin, T. R. and Rubin, D. B. "A Method for Calibrating False-Match Rates in Record Linkage." *Journal of the American Statistical Association*, 90(430):694–707 (1995).

Bhattacharya, I. and Getoor, L. "A Latent Dirichlet Model for Unsupervised Entity Resolution." In *SDM*, volume 5. SIAM (2006).

Bilenko, M. and Mooney, R. J. "Adaptive Duplicate Detection Using Learnable String Similarity Measures." In *KDD '03*, 39–48. ACM (2003).

Christen, P. "Automatic Record Linkage Using Seeded Nearest Neighbour and Support Vector Machine Classification." In *KDD '08*, 151–159. ACM (2008).

—. "A survey of indexing techniques for scalable record linkage and deduplication." *IEEE Transactions on Knowledge and Data Engineering*, 24(9):1537–1555 (2012).

Cohen, W., Ravikumar, P., and Fienberg, S. "A Comparison of String Metrics for Matching Names and Records." In *KDD Workshop on Data Cleaning and Object Consolidation*, volume 3, 73–78 (2003).

Copas, J. and Hilton, F. "Record Linkage: Statistical Models for Matching Computer Records." *Journal of the Royal Statistical Society, Series A*, 153(3):287–320 (1990).

Dai, A. M. and Storkey, A. J. "The Grouped Author-Topic Model for Unsupervised Entity Resolution." In *Artificial Neural Networks and Machine Learning–ICANN 2011*, 241–249. Springer (2011).

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

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

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

Hsu, W., Lee, M. L., Liu, B., and Ling, T. W. "Exploration Mining in Diabetic Patients Databases: Findings and Conclusions." In *KDD '00*, 430–436. ACM (2000).

Jain, S. and Neal, R. "A Split-Merge Markov Chain Monte Carlo Procedure for the Dirichlet Process Mixture Model." *Journal of Computational and Graphical Statistics*, 13:158–182 (2004).

Jewell, N. P., Spagat, M., and Jewell, B. L. "MSE and Casualty Counts: Assumptions, Interpretation, and Challenges." In Seybolt, T. B., Aronson, J. D., and Fischhoff, B. (eds.), *Counting Civilian Casualties: An Introduction to Recording and Estimating Nonmilitary Deaths in Conflict*. Oxford, UK: Oxford University Press (2013).

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

—. "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 (2005).

—. "An Experiment with Hierarchical Bayesian Record Linkage." Preprint in arXiv: http://arxiv.org/abs/1212.5203 (2012).

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

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

Lum, K., Price, M. E., and Banks, D. "Applications of Multiple Systems Estimation in Human Rights Research." *The American Statistician*, 67(4):191–200 (2013).

Marchant, N. G., Steorts, R. C., Kaplan, A., Rubinstein, B. I. P., and Elazar, D. N. "d-blink: Distributed End-to-End Bayesian Entity Resolution." (2019).

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

McCallum, A. and Wellner, B. "Conditional Models of Identity Uncertainty with Application to Noun Coreference." In *Advances in Neural Information Processing Systems (NIPS '04)*, 905–912. MIT Press (2004).

Miller, P. L., Frawley, S. J., and Sayward, F. G. "IMM/Scrub: A Domain-Specific Tool for the Deduplication of Vaccination History Records in Childhood Immunization Registries." *Computers and Biomedical Research*, 33(2):126–143 (2000).

Monge, A. and Elkan, C. "An Efficient Domain-Independent Algorithm for Detecting Approximately Duplicate Datadata Records." (1997).

Murphy, J., Brackbill, R. M., Thalji, L., Dolan, M., Pulliam, P., and Walker, D. J. "Measuring and Maximizing Coverage in the World Trade Center Health Registry." *Statistics in Medicine*, 26(8):1688–1701 (2007).

Murray, J. S. "Probabilistic Record Linkage and Deduplication after Indexing, Blocking, and Filtering." *Journal of Privacy and Confidentiality*, 7(1):3–24 (2016).

Newcombe, H. B., Kennedy, J. M., Axford, S. J., and James, A. P. "Automatic Linkage of Vital Records Computers can be used to extract" follow-up" statistics of families from files of routine records." *Science*, 130(3381):954–959 (1959).

Sadinle, M. "Detecting Duplicates in a Homicide Registry Using a Bayesian Partitioning Approach." *Annals of Applied Statistics*, 8(4):2404–2434 (2014).

—. "Bayesian Estimation of Bipartite Matchings for Record Linkage." *Journal of the American Statistical Association*, (just-accepted):1–35 (2016).

Sariyar, M. and Borg, A. "The RecordLinkage Package: Detecting Errors in Data." *The R Journal*, 2(2):61–67 (2010).

Sariyar, M., Borg, A., and Pommerening, K. "Active Learning Strategies for the Deduplication of Electronic Patient Data Using Classification Trees." *Journal of Biomedical Informatics*, 45(5):893–900 (2012).

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

Steorts, R. C., Hall, R., and Fienberg, S. E. "A Bayesian Approach to Graphical Record Linkage and Deduplication." *Journal of the American Statistical Association*, 111(516):1660–1672 (2016).

—. "A Bayesian Approach to Graphical record Linkage and De-duplication." *Journal of the American Statistical Society* (In press).

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

APPENDIX

A DERIVATION OF FULL CONDITIONAL DISTRIBUTIONS

It is straightforward to show that

$$\pi(\boldsymbol{\beta} \mid \boldsymbol{\Lambda}, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{X}) \propto \prod_{i=1}^{k} \prod_{\ell=1}^{p_s+p_\ell} \boldsymbol{\beta}_{i\ell}^{\sum_{j=1}^{n_i} z_{ij\ell}+a-1}(1-\boldsymbol{\beta}_{i\ell})^{n_i-\sum_{j=1} \boldsymbol{z}_{ij\ell}+b-1},$$

which means

$$\boldsymbol{\beta}_{i\ell} \mid \boldsymbol{\Lambda}, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{X} \overset{iid}{\sim} \text{Beta}(\sum_{j=1}^{n_i} z_{ij\ell} + a, n_i - \sum_{j=1}^{n_i} z_{ij\ell} + b).$$

Consider the distribution of $\boldsymbol{z} \mid \boldsymbol{\Lambda}, \boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{X}$. We find that

- If $\boldsymbol{X}_{ij\ell} \neq Y_{\lambda_{ij}\ell}$, then $\boldsymbol{z}_{ij\ell} = 1$.
- If $\boldsymbol{X}_{ij\ell} = Y_{\lambda_{ij}\ell}$, then
  - if $\ell \leq p_s$ (meaning that we have a string field), then

$$\boldsymbol{z}_{ij\ell} = \begin{cases} 1 \text{ w.p. proportional to } \boldsymbol{\beta}_{i\ell}\ \alpha_\ell(\boldsymbol{X}_{ij\ell})\ h_\ell(Y_{\lambda_{ij}\ell})\ \exp\{-cd(\boldsymbol{X}_{ij\ell}, Y_{\lambda_{ij}\ell})\} \\ 0 \text{ w.p. proportional to } 1-\boldsymbol{\beta}_{i\ell} \end{cases}$$

    This implies that

$$\boldsymbol{z} \mid \boldsymbol{\Lambda}, \boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{X} \sim \text{Bernoulli}\left(\frac{\boldsymbol{\beta}_{i\ell}\ \alpha_\ell(\boldsymbol{X}_{ij\ell})\ h_\ell(Y_{\lambda_{ij}\ell})\ \exp\{-cd(\boldsymbol{X}_{ij\ell}, Y_{\lambda_{ij}\ell})\}}{\boldsymbol{\beta}_{i\ell}\ \alpha_\ell(\boldsymbol{X}_{ij\ell})\ h_\ell(Y_{\lambda_{ij}\ell})\ \exp\{-cd(\boldsymbol{X}_{ij\ell}, Y_{\lambda_{ij}\ell})\} + (1-\boldsymbol{\beta}_{i\ell})}\right).$$

  - If $\ell > p_s$ (meaning that ℓ is not a string field), then there is no h_ℓ term and hence

$$\boldsymbol{z} \mid \boldsymbol{\Lambda}, \boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{X} \sim \text{Bernoulli}\left(\frac{\boldsymbol{\beta}_{i\ell}\ \alpha_\ell(\boldsymbol{X}_{ij\ell})\}}{\boldsymbol{\beta}_{i\ell}\ \alpha_\ell(\boldsymbol{X}_{ij\ell}) + (1-\boldsymbol{\beta}_{i\ell})}\right).$$

Remark: $\boldsymbol{z}_{ij\ell}$ are all independent conditional on $\boldsymbol{\Lambda}, \boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{X}$.

We now turn to the conditional distribution of $\boldsymbol{y} \mid \boldsymbol{\Lambda}, \boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{X}$. First, note that each $\boldsymbol{y}_{j\prime\ell}$ takes values in the set S_ℓ, which consists of all values for the ℓth field that appear anywhere in the data. Then the distribution of $\boldsymbol{y}_{j\prime\ell} \mid \boldsymbol{\Lambda}, \boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{X}$ takes the form $P(\boldsymbol{y}_{j\prime\ell} = w \mid \boldsymbol{\Lambda}, \boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{X}) = A_\phi \phi_w$ for all $w \in S_\ell$, where $A_\phi = \left(\sum_{w \in S_\ell} \phi_w\right)^{-1}$.[3]

Let $R'_j = \{(i,j) : \lambda_{ij} = j'\}$ be the set of all records that correspond to individual j'. Then if $\ell \leq p_s$,

$$\phi_w = \prod_{\substack{(i,j) \in R'_j \\ z_{ij\ell}=1}} h_\ell(w) \exp\left\{-c \sum_{\substack{(i,j) \in R'_j \\ z_{ij\ell}=1}} d(X_{ij\ell}, w)\right\} \alpha_\ell(w) \times \prod_{\substack{(i,j) \in R'_j \\ z_{ij\ell}=0}} I(X_{ij\ell} = w).$$

[3]Both ϕ_w and A_ϕ depend on both $j\prime$ and ℓ, however, for convenience we leave this off.

Simplifying,

$$\phi_w = \begin{cases} \alpha_\ell(w) \prod\limits_{\substack{(i,j)\in R'_j \\ z_{ij\ell}=1}} \{h_\ell(w)\ \exp\{-c\ d(X_{ij\ell}, w)\}\} \text{ if } X_{ij\ell} = w\ \forall (i,j) \in R'_j \ni z_{ij\ell} = 0 \\ 0 \text{ otherwise.} \end{cases}$$

Hence, if $\ell \leq p_s$, then $Y_{j'\ell} \mid \boldsymbol{\Lambda}, \boldsymbol{z}, \boldsymbol{\beta}, \boldsymbol{X}$ has the distribution

$$P(Y_{j'\ell} = w \mid \boldsymbol{\Lambda}, \boldsymbol{z}, \boldsymbol{\beta}, \boldsymbol{X}) = \frac{\alpha_\ell(w) \prod_{(i,j)\in R'_j\, z_{ij\ell}=1} \{h_\ell(w)\ \exp\{-c\ d(X_{ij\ell}, w)\}\}}{\sum_{w\in S_\ell} \left(\alpha_\ell(w) \prod_{(i,j)\in R'_j\, z_{ij\ell}=1} \{h_\ell(w)\ \exp\{-c\ d(X_{ij\ell}, w)\}\} \right)}.$$

If instead, $\ell > p_c$, then we find that $P(Y_{j'\ell} = w \mid \boldsymbol{\Lambda}, \boldsymbol{z}, \boldsymbol{\beta}, \boldsymbol{X}) = \alpha_\ell(w) \left(\sum_{w\in S_\ell} \alpha_\ell(w) \right)^{-1}$.

Regarding, the linkage structure $\boldsymbol{\Lambda}$, its full conditional is as follows: $P(\lambda_{ij} = v \mid \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\beta}, \boldsymbol{X}) = 0$ if there exists ℓ such that $z_{ij\ell} = 0$ and $X_{ij\ell} \neq Y_{v\ell}$. Otherwise,

$$P(\lambda_{ij} = v \mid \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\beta}, \boldsymbol{X}) \propto \prod_{\substack{\ell=1 \\ z_{ij\ell}=1}}^{p_s} \{h_\ell(Y_{v\ell}) \exp\{-c\ d(X_{ij\ell}, Y_{v\ell})\}\}.$$

Define $\Omega_{ij} = \{j' : X_{ij\ell} = Y_{j'\ell}\ \forall\ \ell \ni z_{ij\ell} = 0\}$. Then this implies

$$P(\lambda_{ij} = v \mid \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\beta}, \boldsymbol{X}) = \frac{\prod\limits_{\substack{\ell=1 \\ z_{ij\ell}=1}}^{p_s} \{h_\ell(Y_{v\ell}) \exp\{-c\ d(X_{ij\ell}, Y_{v\ell})\}\}}{\sum_{v'\in\Omega_{ij}} \left\{ \prod\limits_{\substack{\ell=1 \\ z_{ij\ell}=1}}^{p_s} \{h_\ell(Y_{v\ell}) \exp\{-c\ d(X_{ij\ell}, Y_{v\ell})\}\} \right\}}.$$