Performance Bounds for Graphical Record Linkage

Rebecca C. Steorts
Departments of Statistical Science and Computer Science
Duke University
beka@stat.duke.edu

Matt Barnes
The Robotics Institute
Carnegie Mellon University
mbarnes1@cs.cmu.edu

Willie Neiswanger
Machine Learning Department
Carnegie Mellon University
willie@cs.cmu.edu

Abstract

Record linkage involves merging records in large, noisy databases to remove duplicate entities. It has become an important area because of its widespread occurrence in bibliometrics, public health, official statistics production, political science, and beyond. Traditional linkage methods directly linking records to one another are computationally infeasible as the number of records grows. As a result, it is increasingly common for researchers to treat record linkage as a clustering task, in which each latent entity is associated with one or more noisy database records. We critically assess performance bounds using the Kullback-Leibler (KL) divergence under a Bayesian record linkage framework, making connections to Kolchin partition models. We provide an upper bound using the KL divergence and a lower bound on the minimum probability of misclassifying a latent entity. We give insights for when our bounds hold using simulated data and provide practical user guidance.

1 Introduction

Record linkage (de-deduplication or entity resolution) involves identifying duplicate records in large, noisy databases [3]. Traditional linkage methods that directly link records to one another become computationally infeasible as the number of records grows [3, 18], and thus, it is increasingly common for researchers to treat linkage as a clustering task, in which latent entities are associated with one or more noisy database records, and the inferential goal is to identify the latent entity underlying each observed database record [14, 15, 16]. Although there are many probabilistic, generative models for clustering — of which several have been used for record linkage — the theoretical properties, such as performance bounds, have such not been critically assessed.

The work of [14, 15, 16] attempted to deconstruct distorted data by latent variable mixture models. The authors achieved this by clustering similar records to a hypothesized latent entity for each observed record, where their *linkage structure* kept track of which latent entity belongs to the same observed records. This is modeled through a latent variable mixture model with a distortion process on the data (sections 2.1 and 2.2). Thus, the main goal is to be able to take distorted data and uncover the underlying structure in the presence of noise. This is similar to signal processing, where a signal is received in the presence of some noise and often the goal is to understand if the underlying true (latent) signal can be recovered. We develop performance bounds under the framework proposed by [14, 15, 16].

We provide an upper bound on the Kullback-Leibler (KL) divergence between models with different linkage structures and use it to provide a lower bound on the minimum probability of misclassifying a latent entity. More precisely, under the categorical model of [15, 16] and string model of [14], we find the minimum probability of getting a latent entity incorrect. We make connections to Kolchin partition (KP) models [10], along with extending our overall KL bounds in general. Finally, we explore how our bounds perform in practice and describe their user practicality.

1.1 Prior work

Bayesian methods and latent variable modeling have become recently popularized in record linkage models. A major advantage of Bayesian methods is their natural handling of uncertainty quantification for the

Appearing in Proceedings of the 20^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2017, Fort Lauderdale, Florida, USA. JMLR: W&CP volume 54.

resulting estimates. The first notion of understanding a distortion process for record linkage is the hit-miss-model, which uses a binary distortion process on the data [4]. Within the Bayesian paradigm, most work has focused on specialized approaches related to linking two files [6, 17]. These contributions, while valuable, do not easily generalize to more than two files or to de-duplication within a single file. For a review of recent development in Bayesian methods, see [8].

The work of [15, 16] recently introduced a Bayesian model that simultaneously handled record linkage and de-duplication for categorical data. Their approach allowed for natural uncertainty quantification during analysis and post-processing. Finally, [12] recently extended the work of [16] to both categorical and string valued data using a coreference matrix or a partitioning approach. In the later paper, it was shown that the coreference matrix is a special case of the linkage structure, thus, we work with the linkage structure. Another advantage of [16] and similar approaches is that their linkage structure is amenable to an efficient MCMC inference algorithm. These models have become practically relevant as they have been shown to perform well on a variety of applications, including official statistics and medical data. In addition, extensions have been made to more general framework of models [12, 17, 19], which is incorporated into our framework in section 4.

Given the noted distortion process, deriving performance bounds seems natural to recover the underlying structure. For example, much work has been done in information theory for subset selection in graphical model selection, signal de-noising, compressive sensing, and others. In compressed sensing, one question recently addressed in [5], was directly measuring the part of the data from sounds and images that *will not* be thrown away. We make a connection here, as in record linkage we wish to take noisy, distorted data and recover this under the KL divergence. Divergence functions by [13, 7] are useful in many applications including recent statistical applications of clustering, as done in [1] for hard clustering to obtain optimal quantization by minimizing the Bregman divergence (motivated by rate distortion theory).

The rest of this paper proceeds as follows. Two recent record linkage models are given in section 2; Section 2.1 and section 2.2 review these models. Section 3 derives the respective performance bounds, while section 4 extends our general result to a wider class of models. Section 5 shows performance of the bounds in practice, discusses our findings and user practicality. Section 6 discusses future work.

2 Bayesian Record Linkage

We assume two Bayesian record linkage models, one dealing with categorical data and the other dealing with both categorical and noisy string data, such as names, addresses, etc. The first is that of [15, 16], and the second is that of [14].

2.1 Categorical Bayesian Record Linkage

We review common notation to both models.[1] Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ represent the data, with k databases, indexed by i. The ith list has n_i observed records, indexed by j. Each record corresponds to one of N latent entities, indexed by j'. Assume $N = \sum_{i=1}^{k} n_i$ without loss of generality. Each record or latent entity has values on p fields, indexed by ℓ, and are assumed be categorical and the same across all records and entities [15, 16]. M_ℓ denotes the number of possible categorical values for the ℓth field.

In both models, $X_{ij\ell}$ denotes the observed value of the ℓth field for the jth record in the ith list, and $Y_{j'\ell}$ denotes the true value of the ℓth field for the j'th latent entity. Then Λ_{ij} denotes the latent entity to which the jth record in the ith list corresponds, i.e., $X_{ij\ell}$ and $Y_{j'\ell}$ represent the same entity if and only if $\Lambda_{ij} = j'$. Then $\boldsymbol{\Lambda}$ denotes the Λ_{ij} collectively. Distortion is denoted by $z_{ij\ell} = I(X_{ij\ell} \neq Y_{\Lambda_{ij}\ell})$, where $I(\cdot)$ denotes the indicator function. As usual, I represents the indicator function (e.g., $I(x_{ij\ell} = m)$ is 1 when the ℓth field in record j in file i has the value m), and let δ_a denote the distribution of a point mass at a (e.g., $\delta_{y_{\Lambda_{ij}\ell}}$). The model of [15, 16] is:

$$
\begin{aligned}
X_{ij\ell} \mid \Lambda_{ij}, Y_{\Lambda_{ij}\ell}, z_{ij\ell}, \boldsymbol{\theta}_\ell &\overset{\text{ind}}{\sim} \begin{cases} \delta_{Y_{\Lambda_{ij}\ell}} & \text{if } z_{ij\ell} = 0 \\ \text{MN}(1, \boldsymbol{\theta}_\ell) & \text{if } z_{ij\ell} = 1 \end{cases} \\
z_{ij\ell} &\overset{\text{ind}}{\sim} \text{Bernoulli}(\beta_\ell) \\
Y_{j'\ell} \mid \boldsymbol{\theta}_\ell &\overset{\text{ind}}{\sim} \text{MN}(1, \boldsymbol{\theta}_\ell) \\
\boldsymbol{\theta}_\ell \overset{\text{ind}}{\sim} \text{Dirichlet}(\boldsymbol{\mu}_\ell) \;\text{ and }\; \beta_\ell &\overset{\text{ind}}{\sim} \text{Beta}(a_\ell, b_\ell) \\
\Lambda_{ij} &\overset{\text{ind}}{\sim} \text{Uniform}(1, \ldots, N), \qquad (1)
\end{aligned}
$$

where MN denotes the Multinomial distribution and $a_\ell, b_\ell, \boldsymbol{\mu}_\ell$ are all known. Guidance for the hyper-parameters and a justification of the (discrete) uniform prior are given in [15, 14, 16]. Model 1 assumes that different records are independent conditional on the deeper variables of the model. Moreover, it assumes the same conditional independence of different fields for the same record. Finally, observe that record linkage and de-duplication are both simply a question of whether

[1] For a toy example of the record linkage process, see the Supplementary Material.

$\Lambda_{i_1,j_1} = \Lambda_{i_2,j_2}$, where $i_1 \neq i_2$ for record linkage and $i_1 = i_2$ for de-duplication.

2.2 Empirical Bayesian Record Linkage

The work of [14] assumes fields $1, \ldots, p_s$ are string-valued, while fields $p_s + 1, \ldots, p_s + p_c$ are categorical, where $p_s + p_c = p$ is the total number of fields. They assume an empirical Bayesian distribution on the latent parameter. 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., $S_\ell = \{X_{ij\ell} : 1 \leq i \leq k, 1 \leq j \leq n_i\}$, and let $\alpha_\ell(w)$ equal the empirical frequency of value w in field ℓ. Let G_ℓ denote the empirical distribution of the data in the ℓth field from all records in all databases combined. So, if a random variable W has distribution G_ℓ, then for every $w \in S_\ell$, $P(W = w) = \alpha_\ell(w)$. Hence, let G_ℓ be the prior for each latent entity $Y_{j'\ell}$. The distortion process changes such that

$$P(X_{ij\ell} = w \mid \Lambda_{ij}, Y_{\Lambda_{ij}\ell}, z_{ij\ell}) = \frac{\alpha_\ell(w)\, \exp[-c\, d(w, Y_{\Lambda_{ij}\ell})]}{\sum_{w \in S_\ell} \alpha_\ell(w)\, \exp[-c\, d(w, Y_{\Lambda_{ij}\ell})]},$$

where $c > 0$ is a fixed normalizing constant corresponding to an arbitary distance metric $d(\cdot, \cdot)$. Denote this distribution by $F_\ell(Y_{\Lambda_{ij}\ell})$. The model becomes

$$\begin{aligned} X_{ij\ell} \mid \Lambda_{ij}, Y_{\Lambda_{ij}\ell}, z_{ij\ell} &\overset{\text{ind}}{\sim} \begin{cases} \delta(Y_{\Lambda_{ij}\ell}) & \text{if } z_{ij\ell} = 0 \\ F_\ell(Y_{\Lambda_{ij}\ell}) & \text{if } z_{ij\ell} = 1, \ell \leq p_s \\ G_\ell & \text{if } z_{ij\ell} = 1, \ell > p_s \end{cases} \\ Y_{j'\ell} &\overset{\text{ind}}{\sim} G_\ell \\ z_{ij\ell} \mid \beta_{i\ell} &\overset{\text{ind}}{\sim} \text{Bernoulli}(\beta_{i\ell}) \\ \beta_{i\ell} &\overset{\text{ind}}{\sim} \text{Beta}(a, b) \\ \Lambda_{ij} &\overset{\text{ind}}{\sim} \text{Uniform}(1, \ldots, N), \end{aligned} \tag{2}$$

where all distributions are also independent of each other; assume that a, b, N are assumed known. This framework was shown to work well in applications and simulation studies, however, it was quite sensitive to the choice of the hyperparameters. This method beat supervised methods, such as random forests when the amount of training data input into the supervised methods was $< 10\%$.

Figure 1 contains a graphical representation of models 1-2.

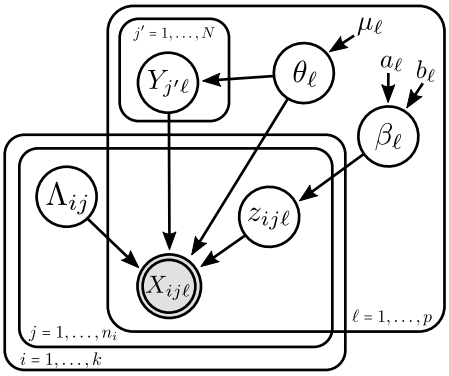

Figure 1: Graphical representation of models 1-2.

3 Performance Bounds of Record Linkage

Recall the connection to KL divergence in the sense that for any two distributions P and Q, the maximum power for testing P versus Q is $\exp\{-nD_{\text{KL}}(P||Q)\}$. Hence, a low value of D_{KL} means that we need many samples to distinguish P from Q. A natural question is how does changing $\boldsymbol{Y}$ (latent entity) or $\boldsymbol{\Lambda}$ (linkage structure) change the distribution of $\boldsymbol{X}$ (observed records)? We search for both meaningful upper and lower bounds, since an upper bound will say that P and Q are never more than so far apart, whereas a lower bound says how easy it is to tell P and Q apart. Moreover, we investigate how well can we recover $\boldsymbol{Y}$ (latent entity) and $\boldsymbol{\Lambda}$ (linkage structure) from $\boldsymbol{X}$ (data).

Assuming the conditions of [15, 14], let $\mathcal{P} = \{f(X \mid \boldsymbol{Y}, \Lambda_{ij}, \boldsymbol{\theta}, \boldsymbol{\beta}) : \forall \Lambda_{ij} \in \{1, \ldots, N\}.\}$ We know that $X_1, X_2, \ldots, X_N$ are all independent given $(\boldsymbol{Y}, \boldsymbol{\Lambda}, \boldsymbol{\theta}, \boldsymbol{\beta})$ under both $P, Q \in \mathcal{P}$. This implies that $D_{X_1,X_2,\ldots,X_N}(P\|Q) = \sum_i D_{X_i}(P\|Q)$. We first provide a theorem under the model of [15], which assumes categorical data and a hierarchical model. In Theorem 1, we find the minimum probability of getting a latent entity wrong. Moreover, we are able to say that with growing distortion of the data, there is no difference between two latent entities and the bound becomes infinite and non-informative in this case. Next, under the model of [14] we provide a general theorem, which assumes both categorical and noisy text data. This theorem provides an upper bound on the KL divergence of arbitrary distributions P and Q.

3.1 Kullback-Leibler Divergence under Categorical Data

We use Fano's inequality [11] to bound the probability of misclassification, as a function of the KL divergence between P and Q, as defined in the previous section. Assume that $\boldsymbol{\Lambda}$ and $\hat{\boldsymbol{\Lambda}}$ are two distinct linkage structures that correspond to the same latent entity $(\boldsymbol{y})$. Let $r + 1$ be the cardinality of $\mathcal{P}$, i.e. $r + 1 = N$.

Theorem 1. *This result finds an upper bound on the KL divergence and a lower bound for the prob-*

ability that model 1 gets the linkage structure incorrect. Let $\gamma = \max_{\Lambda_{ij} \neq \Lambda'_{ij}} 2\sum_{ij\ell} I(Y_{\Lambda_{ij}\ell} \neq Y_{\Lambda'_{ij}\ell})(1-\beta_\ell) \ln\left\{\frac{1}{\min_m \theta_{\ell m}\beta_\ell}\right\}$.

i) The KL divergence is bounded above by γ*. That is,* $D_X(P||Q) \leq \gamma \ \ \forall P, Q \in \mathcal{P}$.

ii) The minimum probability of getting a latent entity wrong is $Pr(\Lambda_{ij} \neq \Lambda'_{ij}) \geq 1 - \frac{\gamma + \ln 2}{\ln r}, \ \ \forall i, j$

That is, as the latent entities become more distinct, γ increases. On the other hand, as the latent entities become more similar, $\gamma \to 0$.

Remark 1. *Consider Theorem 1 (i). Suppose* $\beta_\ell \to 1$*. Then* $D_{\boldsymbol{X}} \geq 0$*. If instead* $\beta_\ell \to 0$*, then* $D_{\boldsymbol{X}} \geq 1$*. The lower bound is only informative when* $\beta_\ell \to 0$*. We have more information when the latent entities are separated.*

Proof. To show this, we simply apply Pinsker's inequality, where for all $P, Q \in \mathcal{P}$: $D(P\|Q) \geq 2||P-Q||_1^2 \implies D(P\|Q) \geq I(Y_{\Lambda_{ij}\ell} \neq Y_{\Lambda'_{ij}\ell})(1-\beta_\ell)^2 \implies D_{\boldsymbol{X}}(P\|Q) \geq \sum_{ij\ell} I(Y_{\Lambda_{ij}\ell} \neq Y_{\Lambda'_{ij}\ell})(1-\beta_\ell)^2$. □

Proof. We assume the model of [15, 16], which assumes that data is categorical. We assume model 1 holds in section 2.1. We first prove (i). Consider $f(X \mid \boldsymbol{Y}, \boldsymbol{\Lambda}, \boldsymbol{\theta}, \boldsymbol{\beta})$. Then

$$\begin{aligned} &Pr(X_{ij\ell} = m \mid \boldsymbol{Y}, \boldsymbol{\Lambda}, \boldsymbol{\theta}, \boldsymbol{\beta}) \\ &= 1(Y_{\Lambda_{ij}\ell} = m)(1-\beta_\ell) + \theta_{\ell m}\beta_\ell. \end{aligned} \tag{3}$$

It follows from equation 3 that

$$\begin{aligned} D_{X_{ij\ell}}(P\|Q) = &\sum_{m=1}^{M_\ell} I(Y_{\Lambda_{ij}\ell} = m)(1-\beta_\ell) + \theta_{\ell m}\beta_\ell\} \\ &\times \log\left[\frac{I(Y_{\Lambda_{ij}\ell} = m)(1-\beta_\ell) + \theta_{\ell m}\beta_\ell}{I(Y_{\Lambda'_{ij}\ell} = m)(1-\beta_\ell) + \theta_{\ell m}\beta_\ell}\right]. \end{aligned}$$

It directly follows that

$$\begin{aligned} D_{\boldsymbol{X}}(P\|Q) = &\sum_{ij\ell m} \{I(Y_{\Lambda_{ij}\ell} = m)(1-\beta_\ell) + \theta_{\ell m}\beta_\ell\} \\ &\times \log\left[\frac{I(Y_{\Lambda_{ij}\ell} = m)(1-\beta_\ell) + \theta_{\ell m}\beta_\ell}{I(Y_{\Lambda'_{ij}\ell} = m)(1-\beta_\ell) + \theta_{\ell m}\beta_\ell}\right]\Bigg\}. \end{aligned}$$

If $Y_{\Lambda_{ij}\ell} \neq Y_{\Lambda'_{ij}\ell}$, then

$$\begin{aligned} \|P-Q\|_1 = &\sum_m \Big| I(Y_{\Lambda_{ij}\ell} = m)(1-\beta_\ell) + \theta_{\ell m}\beta_\ell \\ &\quad - I(Y_{\Lambda'_{ij}\ell} = m)(1-\beta_\ell) - \theta_{\ell m}\beta_\ell \Big| \\ = &\ 2(1-\beta_\ell). \end{aligned} \tag{4}$$

Equation 4 holds since $P(m) = Q(m)$ unless $m = Y_{\Lambda_{ij}\ell}$ or $m = Y_{\Lambda'_{ij}\ell}$. If $Y_{\Lambda_{ij}\ell} = Y_{\Lambda'_{ij}\ell}$, then $P = Q$ and $\|P-Q\|_1 = 0$. The reverse Pinsker inequality of [2] relates the KL divergence to the L_1 norm in the following way: $D(P\|Q) \leq \|P-Q\|_1 \ln\{(\min Q)^{-1}\}$. Using this, we find that (if $Y_{\Lambda_{ij}\ell} \neq Y_{\Lambda'_{ij}\ell}$), then

$$\begin{aligned} D(P\|Q) &\leq 2(1-\beta_\ell) \\ &\quad \times \ln\left\{\frac{1}{\min_m I(Y_{\Lambda'_{ij}\ell} = m)(1-\beta_\ell) + \theta_{\ell m}\beta_\ell}\right\} \\ &\leq 2(1-\beta_\ell)\ln\left\{\frac{1}{\min_m \theta_{\ell m}\beta_\ell}\right\}. \end{aligned}$$

Hence,

$$\begin{aligned} \max_{P,Q\in\mathcal{P}} D_{\boldsymbol{X}}(P\|Q) \leq &\max_{\Lambda_{ij}\neq\Lambda'_{ij}} 2\sum_{ij\ell} I(Y_{\Lambda_{ij}\ell} \neq Y_{\Lambda'_{ij}\ell})(1-\beta_\ell) \\ &\times \ln\left\{\frac{1}{\min_m \theta_{\ell m}\beta_\ell}\right\} := \gamma. \end{aligned}$$

This proves (i). We now prove (ii). Using Fano's inequality [11], the minimum probability of getting a latent entity wrong is $Pr(\Lambda_{ij} \neq \Lambda'_{ij}) \geq 1 - \frac{\gamma + \ln 2}{\ln r}$, where $r+1$ is the cardinality of $\mathcal{P}$, i.e. $r+1 = N$. As the latent entities become more distinct, γ increases. On the other hand, as the latent entities become more similar, $\gamma \to 0$. □

3.2 KL Divergence Bounds for String and Categorical Data

We now consider P and Q under [14] for both categorical and noisy string data. Recall that β_ℓ tunes the amount of distortion as defined in equation 2. Recall that $d(\cdot,\cdot)$ denotes any arbitrary distance metric between an observed string and a latent string as seen in equation 2, and $c > 0$ is a fixed normalizing constant corresponding to the distance metric d.

In Theorem 2, for any distinct linkage structures, the minimum probability of getting a latent entity wrong is governed by a lower bound, which is growing at a rate $c \to \infty$ that is determined by the moment generating function of the distances between an observed string in data and a latent string.

Theorem 2. *Assume data* $\boldsymbol{X}$*, and distributions* $P, Q \in \mathcal{P}$ *defined in section 3. Assume two distinct linkage structures, denoted by* $Y_{\Lambda_{ij}\ell}, Y_{\Lambda'_{ij}\ell}$*.*

i) There is an upper bound on the KL divergence between any $P, Q \in \mathcal{P}$ *given by* κ*, that is* $D_X(P||Q) \leq \kappa$*.*

ii) $Pr(\Lambda_{ij} \neq \Lambda'_{ij}) \geq 1 - \dfrac{\kappa + \ln 2}{\ln r}$, *where*

$$\kappa = \max_{\Lambda_{ij} \neq \Lambda'_{ij}} \Bigg[2 \sum_{\ell} (1 - \beta_\ell) I(Y_{\Lambda_{ij}\ell} \neq Y_{\Lambda'_{ij}\ell}) + \sum_{\ell m} I(Y_{\Lambda_{ij}\ell} \neq Y_{\Lambda'_{ij}\ell}) \left(1 - e^{-cd(Y_{\Lambda_{ij}\ell}, Y_{\Lambda'_{ij}\ell})}\right) \times E[e^{-cd(m, Y_{\Lambda_{ij}\ell})}] \Bigg] \ln\{(\min Q)^{-1}\}$$

and $r + 1$ *is the cardinality of* $\mathcal{P}$.

The proof of Theorem 2 can be found in the Supplementary Material.

4 Kolchin Partition Models

The models in sections 2.1 and 2.2 assume discrete uniform priors on the linkage structure. We extend this to a more general class of models from Bayesian nonparametrics known as KP models [10]. Special cases include the work of [19, 17, 12]. We provide notation, examples, and then provide a general theorem.

The prior structure on $\mathbf{\Lambda}$ can instead be viewed on the set of labelings. Specifically, let z denote the partition of the observed records determined by $\mathbf{\Lambda}$, and $\mathcal{B}$ denote the set containing all the possible partitions of the r observed records. Then a distribution on the sample labels $\mathbf{\Lambda}$ induces a distribution on $\mathcal{B}$. That is, matches and duplicates are completely specified given the knowledge of z, which is invariant with respect to the labelings of the partition blocks.

4.1 Special Cases of KP Models

We give some special cases of KP models that extend the class of models that we consider.

4.2 A Uniform Prior on the Label Space

Let $z(\mathbf{\Lambda})$ denote the partition identified by $\mathbf{\Lambda}$ and let $n(z)$ denote the number of distinct entities considering all the observed records, i.e. the number of blocks of the partition. One has

$$N!/(N - n(z))! = (N)_{n(z)}$$

different labelings which identify the same partition $z(\mathbf{\Lambda})$. Then

$$P(z(\mathbf{\Lambda}) = z) = \left(\frac{1}{N}\right)^r \frac{N!}{(N - n(z))!} \quad \forall z \in \mathcal{B}.$$

Note also that $N^r = \sum_{n=0}^{r} (N)_n S(r, n)$ where $S(r, n)$ is the Stirling number of second kind, that is the number of possible partitions of the r records into n non empty sets, which implies

$$P(z(\mathbf{\Lambda}) = z) = \frac{(N)_n}{N^r} \quad \forall z \in \mathcal{B} \tag{5}$$

where $n = n(z)$. Following [10], equation (5) is a special case of a KP model. Moreover, the distribution of the number of distinct elements $n(z)$ is given by $P(n(z) = n) = \dfrac{(N)_n S(r, n)}{N^r}$. (A similar prior was considered in [17]).

4.3 The Uniform Prior on the Partition Space

Assuming one database, [12] focused mainly on the partitions of the r records induced by $\mathbf{\Lambda}$ and proposed a flat prior on the partition space, that is a prior which assigns equal probability to each different partition of the r observed records. Assume that

$$p(\mathbf{\Lambda}) = \frac{1}{B_r \, (N)_{n(z)}}$$

where $B_r = \sum_{n=0}^{r} S(r, n)$ is the r-th Bell number. In terms of partitions, the prior used by [12] can be written as $p(z(\mathbf{\Lambda}) = z) = \frac{1}{B_r}$ and $p(n(z) = n) = \frac{S(r,n)}{B_r}$. This prior is also a special case of a KP model.

Moreover, the discrete uniform priors of [15, 16, 14] can also be represented easily as KP models. We refer to [16] for details.

We now provide a general theorem, which gives a relationship regarding priors (or partitions or blocks) of KP models to our KL divergence bounds. Suppose the prior on the linkage structure can be represented as a KP model [10]. Then a wide class of priors is able to be considered and compared.

Theorem 3. *Consider model 1 or 2. Then the error bounds behave like the corresponding bounds in Theorem 1 or 2.*

Proof. The results directly follows from the proofs of Theorems 1, 2 and the representation of KP models [10]. Specifically, all bounds in Theorems 1 and 2 depend upon the linkage structure $\mathbf{\Lambda}$, which in the proofs, is agnostic to its form. □

4.4 Microclustering and Record Linkage

There has been early work in Bayesian nonparametrics to push forward record linkage. The work of [9, 19] recently pointed out that most clustering tasks assume the cluster sizes grow linearly with the number of the data points. Such examples include infinitely exchangeable clustering models, including finite mixture models,

Dirichlet process (DP) mixture models, and Pitman–Yor process (PYP) mixture models. However, in record linkage, such an assumption is undesirable since linkage methods require models that yield clusters whose sizes grow sublinearly with the total number of data points (records). Due to this, [19] defined the microclustering property as well as a new model exhibiting such growth, where their models outperformed or performed as well as the PYP and DP in terms of standard record linkage evaluation metrics on data for official statistics, medical data, and human rights data. Furthermore, the authors proved that one of their models satisfies the microclustering property under very weak assumptions. We refer to their paper for further details.

One insight of this paper was the fact that their class of microclustering models considered can always be written as a KP model. Combining this clustering approach with the likelihood in equation 1 or 2 immediately allows one to perform record linkage inference. Furthermore, Theorems 1 and 2 are immediately satisfied since the prior on the linkage structure can be represented as a KP model.

5 Simulation Study and Discussion

We consider how the bounds in Sections 3.1 and 3.2 hold for two simulated experiments. In our experiments (**Experiment I** and **Experiment II**), synthetic categorical data are generated according to either model 1 or 2 using the parameters shown in Table 1 and 2, respectively. In order to consider a realistic set of strings for S, we consider the set of 20 most popular female baby names from 2014, according to the United States Census. Then for the distance d, we consider the generalized Levenshtein edit distance.

We then generate both categorical and string records according to either model 1 or 2. For each experiment, we vary exactly one of the parameters to demonstrate its impact of the linkage error rate $Pr((\hat{\Lambda}_{ij}, \boldsymbol{Y}) \neq (\Lambda_{ij}, \boldsymbol{Y}))$. We choose the other values such that the performance is neither extremely low nor extremely high. We set the distortion parameter β_ℓ to the same value for each ℓ, i.e. $\beta_\ell = 0.6$ denotes a distortion probability of 0.6 for every field. $\beta_\ell = 0.0$ to 1.0 means we started with $\beta_\ell = 0$ for all ℓ and swept the values until $\beta_\ell = 1$ for all ℓ. Recall p is the number of fields, and thus the maximum value of ℓ. We also set each $\theta_{\ell m}$ to the same value, i.e. $\theta_{\ell m} = 0.1$ denotes $\theta_{\ell m} = 0.1$ for all ℓ and all m. This further implies each field ℓ takes on exactly $M_\ell = 1/\theta_{\ell m}$ values in order for θ_ℓ to be a valid probability distribution.

We compare the bound in Theorem 1 to two record linkage algorithms [15, 16, 14]. The first is an exact sampler, which samples directly from $Pr(\Lambda_{ij}|X_{ij}, \boldsymbol{Y}, \boldsymbol{z})$.

Experiment	N	β_ℓ	$p = p_c$	$\theta_{\ell m}$
Fig. 1(a)	10 to 500	0.6	3	0.1
Fig. 1(b)	100	0 to 1	3	0.1
Fig. 1(c)	100	0.6	1 to 8	0.25
Fig. 1(d)	100	0.8	5	$\frac{1}{46}$ to 1

Table 1: Categorical Experiments

Experiment	N	β_ℓ	$p = p_s$	c
Fig. 2(a)	100 to 500	0.6	1	1.0
Fig. 2(b)	100	0.2 to 1	1	1.0
Fig. 2(c)	100	0.6	1 to 10	1.0
Fig. 2(d)	100	0.6	1	0 to 2

Table 2: String Experiments

The second is a more realistic Gibbs sampler with empirically motivated priors proposed by [14]. We run the Gibbs sampler for 10,000 iterations on all experiments to ensure proper mixing. There is some difficulty in comparing Λ to $\hat{\Lambda}$, as there are multiple equally correct modes due to arbitrary re-orderings of the latent individuals $\hat{\boldsymbol{Y}}$ and corresponding linkage structure $\hat{\Lambda}$. Even though the Gibbs sampler may infer the correct latent individuals $\boldsymbol{Y}$ and linkage structure, because the ordering is arbitrary, it is unlikely that $\Lambda = \hat{\Lambda}$. To avoid such an issue of label switching, we fix $\hat{\boldsymbol{Y}}$ during the sampling process.

Specifically, we compare the bound to the empirical error rate of the Gibbs sampler proposed by [14]. In order to compute the empirical probability $Pr(\hat{\Lambda}_{ij} \neq \Lambda_{ij})$, we hold $\boldsymbol{Y}$ fixed during Gibbs sampling to ensure errors in $\hat{\boldsymbol{\Lambda}}$ are not due to arbitrary changes in the ordering of the labels of $\boldsymbol{Y}$. In addition, we compare the linkage error rate to an exact sampler, which samples directly from $Pr(\boldsymbol{\Lambda}|X, \boldsymbol{Y}, \boldsymbol{z})$.

Results of Experiment I In Figures 2 (a)-(d) we vary the number of records N, distortion parameter β, number of fields p and number of values each field takes M_ℓ, respectively. The empirical results demonstrate Theorem 1 captures the dependence between the error rate and the all relevant latent parameters θ, N and β. Specifically, linking records becomes more difficult as N increases, the distortion parameter β increases, the number of fields p decreases or the number of values each field can take M_ℓ decreases. The bound nicely captures the logarithmic increase in error with respect to N in Figure 2 (a), which gives hope for linking records in very large databases. Other terms appear to be $\bar{O}(n)$ when not near extreme error values, implying low noise and a larger feature space are essential to performing high quality record linkage.

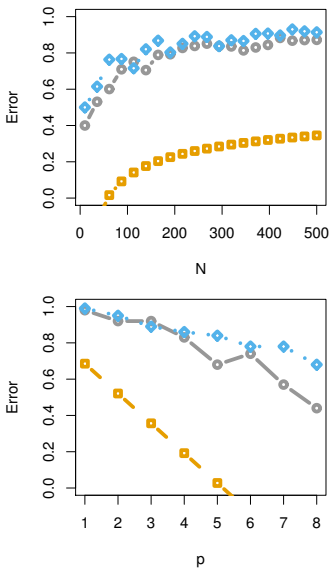

Figure 2: Theorem 2 (gold squares) holds on simulated categorical records compared to exact sampling (grey circles) and Gibbs sampler (blue diamonds).

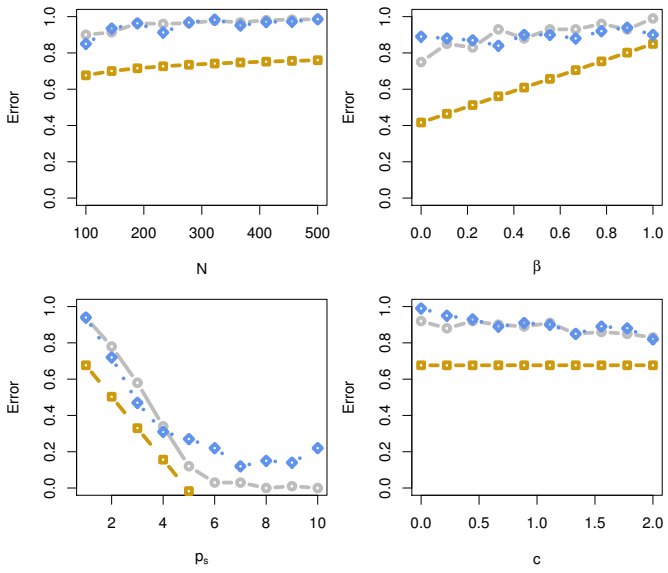

Figure 3: Theorem 2 (gold squares) holds on simulated noisy string records compared to exact sampling (grey circles) and Gibbs sampler (blue diamonds).

Results of Experiment II Figures 3 (a)-(d) show Theorem 2 is tight to the true performances on string data when varying N, β, number of string fields p_s and c, respectively. As expected, and similarly to the categorical results, linking records becomes more difficult as N increases, the distortion parameter β increases and the parameter c decreases. The effects of parameter variation is less noticeable in the string experiments due to the fact that linking string fields is easier than ones that have been anonymized, i.e., categorical fields.

The Gibbs sampler (blue diamonds) performs almost as well as the exact sampler (grey circles). In fact, due to the conditional entropy version of Fano's inequality and the fact that $H(X|Y) \leq H(X)$, any Gibbs sampler cannot perform better in expectation than an exact sampler. Thus, we believe the gap between the bound (gold squares) and the exact sampler does not necessarily indicate the existence of a better algorithm, but perhaps only some unnecessary slack due to the application of Pinsker's and then reverse Pinsker's inequalities.

5.1 Discussion of Results

As illustrated in Theorems 1 and 2 we have derived an upper bound on the KL divergence as well as lower bounds for misclassifying a latent entity. In Theorem 1 (i), we showed that the latent entities become more distinct when γ is increasing. This is in contrast to when γ gets closer to 0, since then the latent entities become more similar. In Theorem 1 (ii), we showed that as the distortion parameter $\beta_\ell \to 1$, then the upper bound γ is infinite. In practice, as illustrated in [15], the latent entities are difficult to distinguish when the amount of distortion is more than 5% at every field value. Thus, this corresponds to when the bound is too loose. On the other hand, as $\beta_\ell \to 0$, the latent entities become more separated.

We discuss how separated the latent entities are under choices of β_ℓ, θ_ℓ and N, providing guidance to the user in this setting given our simulation results. As practical guidance when the distortion is between 0 to 5% at every feature value, the latents will be more separated and the bound will be be loose. On the other hand, as β_ℓ increases, the bound becomes tighter. The choice of β_ℓ can be made using subjective information about the underlying data and tuned using the hyper-parameters a, b. (See [15, 14] for choosing such values). On the other hand, we can see that for more realistic values of the distortion parameter in Figure 2 (a), (b), and (d) , the bound is quite loose when the distortion parameter β_ℓ is large. Thus, a loose bound here is warranted due to the amount of noise or model-misspecification being placed into the model as well as the fact that all of the fields being used are categorical. Such results match the intuition given in [15].

In Theorem 2 (ii), we derived a lower bound where the minimum probability of getting a latent entity wrong is controlled by c, which is determined by the moment generating function of the distances between an observed string and a latent string. This bound has the same type of form as the bound in Theorem 1, however, since we now have string-valued data, we see that the minimum probability of getting a latent entity wrong is dominated by the string-valued variables and specifically, the distances functions used and the constants used. In comparison to [14], this completely matches up with the sensitivity that was seen to the choice of the distance functions as well as the choice of c as this will completely dominate the posterior, and hence, the ability to tell latent entities apart under this posterior.

In practice, the driving force of the tightness of the bound is c, the steepness parameter of the string distribution in equation 2. As c increases, it is less likely for a string-valued record's features to be distorted to values that are far from that of their latent feature values. This is verified in Figure 3(c), where linkage error decreases as c increases. The work of [14] gave practical choices for c, which were [0,2]. Similarly, we can speak to the tightness of d, which relies on the distortion parameter β_ℓ not being too small in practice, as verified in Figure 3(b). In terms of the bounds found in Theorem 1 and 2, the empirical Gibbs sampler has tight bounds in almost all situations, except when the number of features is large, N is too small, or β_ℓ is too small (and similarly for θ_ℓ). This coincides with exactly what we would expect in practice from the real experiments of [14].

For all applications in both categorical and string data, we expect the bounds to be as loose in practice (corresponding to easier record linkage), when the distortion parameter is small $(0-4)$ and when the the number of fields is large ($p \geq 5$) or the number of values that each field can take, M_ℓ, increases (this will be application specific). Finally, the bounds should be tighter, corresponding to more difficult record linkage, as the total number of records N increases (see Figure 2). These parameter values match almost exactly with two real data experiments (corresponding ranges of parameters) as well as a simulation study from [15, 16].

6 Future Directions

First, we have derived general performance bounds for record linkage, making connections to KP models and other related Bayesian models. More specifically, we have drawn connections to a wide class of models from Bayesian record linkage. Second, our bound for the categorical Bayesian record linkage model is easily interpretable and matches the intuition of the generative model. Third, our bound for the categorical and noisy string model, takes a similar form to that of the categorical model. We are also able to interpret this bound in a way that aligns with the interpretations [14, 15, 16] as well as show the practicality of our bounds to the aforementioned papers. More specifically, our bounds are empirically loose for categorical data, which is not unexpected since there is little information available to match on. This contrasts the empirical tight bounds for both categorical and noisy string data. As illustrated in our experiments, with just one string variable, our bounds become much tighter, and as the number of strings increases, the bound becomes more tighter when compared to exact and Gibbs sampling.

In addition, there has been early work in Bayesian nonparametrics to push forward record linkage. The work of [9] pointed out that most clustering tasks assume cluster sizes grow linearly with the size of the database. Such examples include infinitely exchangeable clustering models, including finite mixture models, Dirichlet process mixture models, and Pitman–Yor process mixture models, which all make this linear growth assumption. However, in record linkage such an assumption is undesirable since linkage methods require models that yield clusters whose sizes grow sublinearly with the total number of data points. This observation led the authors to define the microclustering property as well as a new model exhibiting such growth. Our work has been able to provide bounds for the aforementioned work since the prior consider is a KP model. In future work it would be helpful to try and draw connections between those proposed in [9] and [14, 15, 16] in order to generalize such bounds and provide tighter bounds using conditional entropy or other sophisticated bounding methods.

Acknowledgments

We thank David Choi, David Dunson, David Banks, and the reviewers for improving the ideas that led to publication of this paper. This work was supported in part by NSF grants SES-1534412 and SES-1131897, DARPA FA8750-12-2-0324 and FA8750-14-2-0244.

References

[1] Arindam Banerjee, Srujana Merugu, Inderjit S Dhillon, and Joydeep Ghosh. Clustering with bregman divergences. *The Journal of Machine Learning Research*, 6:1705–1749, 2005.

[2] Daniel Berend, Peter Harremoës, and Aryeh Kontorovich. Minimum KL-divergence on complements of l_1 balls. *IEEE Transactions on Information Theory*, 60(6):3172–3177, 2014.

[3] P. Christen. *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Springer, 2012.

[4] J. Copas and F.J. Hilton. Record linkage: Statistical models for matching computer records. *Journal of the Royal Statistical Society, Series A*, 153(3):287–320, 1990.

[5] David L Donoho. Compressed sensing. *Information Theory, IEEE Transactions on*, 52(4):1289–1306, 2006.

[6] R. Gutman, C. Afendulis, and A. Zaslavsky. A Bayesian procedure for file linking to analyze end-of-life medical costs. *Journal of the American Statistical Association*, 108(501):34–47, 2013.

[7] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[8] B. Liseo and A. Tancredi. Some advances on Bayesian record linkage and inference for linked data. *Technical Report*, 2013.

[9] Jeffrey Miller, Brenda Betancourt, Abbas Zaidi, Hanna Wallach, and Rebecca Steorts. The Microclustering Problem: When the Cluster Sizes Don't Grow with the Number of Data Points. *NIPS Bayesian Nonparametrics: The Next Generation Workshop Series*, 2015.

[10] Jim Pitman. *Combinatorial Stochastic Processes: Ecole D'Eté de Probabilités de Saint-Flour XXXII-2002*. Springer, 2006.

[11] Vyacheslav Valer'evich Prelov and Edward C. van der Meulen. Mutual information, variation, and fano's inequality. *Problems of Information Transmission*, 44(3):185–197, 2008.

[12] Mauricio Sadinle. Detecting duplicates in a homicide registry using a Bayesian partitioning approach. *The Annals of Applied Statistics*, 8(4):2404–2434, 2014.

[13] Claude E Shannon. A note on the concept of entropy. *Bell System Tech. J*, 27:379–423, 1948.

[14] R. C. Steorts. Entity resolution with empirically motivated priors. *Bayesian Analysis*, 10(4):849–875, 2015.

[15] R. C. Steorts, R. Hall, and S. E. Fienberg. SMERED: A Bayesian approach to graphical record linkage and de-duplication. *Journal of Machine Learning Research*, 33:922–930, 2014.

[16] R. C. Steorts, R. Hall, and S. E. Fienberg. A Bayesian approach to graphical record linkage and de-duplication. *Journal of the American Statistical Society*, In press.

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

[18] W. E. Winkler. Overview of record linkage and current research directions. Technical report, U.S. Bureau of the Census Statistical Research Division, 2006.

[19] Giacomo Zanella, Brenda Betancourt, Jeffrey W Miller, Hanna Wallach, Abbas Zaidi, and Rebecca Steorts. Flexible models for microclustering with application to entity resolution. In *Advances in Neural Information Processing Systems*, pages 1417–1425, 2016.