Computational Statistics and Data Analysis 172 (2022) 107474

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

www.elsevier.com/locate/csda

A prior for record linkage based on allelic partitions

Brenda Betancourt [a,*], Juan Sosa [b], Abel Rodríguez [c]

[a] Department of Statistics, University of Florida, 220 Griffin-Floyd Hall, P.O. Box 118545, Gainesville, FL, United States of America
[b] Departamento de Estadística, Universidad Nacional de Colombia, Bogotá D.C, Colombia
[c] Department of Statistics, University of Washington, Seattle, WA, United States of America

ARTICLE INFO

Article history:
Received 20 March 2021
Received in revised form 9 March 2022
Accepted 10 March 2022
Available online 16 March 2022

Keywords:
Microclustering
Allelic partitions
Record linkage

ABSTRACT

In database management, record linkage aims to identify multiple records that correspond to the same individual. Record linkage can be treated as a clustering problem in which one or more noisy database records are associated with a unique latent entity. In contrast to traditional clustering applications, a large number of clusters with a few observations per cluster is expected in this context. Hence, a new class of prior distributions based on allelic partitions is proposed for the small cluster setting of record linkage. The proposed prior facilitates the introduction of information about the cluster size distribution at different scales, and naturally enforces sublinear growth of the maximum cluster size – known as the *microclustering property*. In addition, a set of novel microclustering conditions are introduced in order to impose further constraints on the cluster sizes a priori. The performance of the proposed class of priors is evaluated using simulated data and three official statistics data sets. Moreover, different loss functions for optimal point estimation of the partitions are compared using decision-theoretical based approaches recently proposed in the literature.

1. Introduction

With the current stream of data, collection and integration of information from multiple sources has become imperative. The process of merging databases and/or removing duplicate records is known as record linkage (RL) (Christen, 2012). This is a challenging problem considering that databases often contain corrupted data and lack common unique identifiers across files. Areas of application where RL tasks are prevalent, include public health (Gutman et al., 2013; Hof et al., 2017), human rights (Sadinle, 2014, 2017, 2018), official statistics (Winkler, 2014; Kaplan et al., 2018; Wortman, 2019), and fraud detection and national security (Vatsalan et al., 2017).

The seminal work of Fellegi and Sunter (1969) is the classical reference for a probabilistic approach to identifying links between two files, with a recent extension to three files introduced in Sadinle and Fienberg (2013). In particular, these approaches rely on record pair similarity weights to determine sets of matches and non-matches. Other work involving the merge of two files includes Belin and Rubin (1995), Fienberg et al. (1997), Larsen and Rubin (2001), Tancredi and Liseo (2011) and Gutman et al. (2013). A known caveat of these techniques is that they do not easily generalize to either multiple files or duplicate detection within files. In order to deal with more general scenarios, the RL problem can be viewed as a clustering task in which one or more noisy database records that possibly represent the same latent entity are grouped

* Corresponding author.
E-mail address: bbetancourt@ufl.edu (B. Betancourt).

https://doi.org/10.1016/j.csda.2022.107474
0167-9473/

together. From this point of view, an important feature of RL applications is that, generally, a large number of clusters with a few observations per cluster is expected. From a model-based perspective, popular choices for clustering include finite mixture models and Dirichlet/Pitman-Yor process mixture models (Müller and Rodriguez, 2013; Casella et al., 2014; Miller and Harrison, 2018). Although these models have been used in all sorts of applications, including RL (Bhattacharya and Getoor, 2006), they are not well suited for problems with small clusters. Unlike models exhibiting infinitely exchangeable clustering features, models specifically conceived for RL need to generate clusters with a small number of records, even as the size of the data increases (Miller et al., 2015). Within the Bayesian framework, recent advances in latent variable modeling and clustering methods for RL include those of Sadinle (2014), Steorts et al. (2015, 2016). These approaches, however, have the limitation of assuming a uniform prior on the linkage structure which requires strong parameter tuning to achieve sensible RL results.

In order to formulate more appropriate priors for the small cluster setting of RL, Miller et al. (2015) introduce the concept of *microclustering*, in which the size of the largest cluster of the partition is required to grow sublinearly with the number of records. Zanella et al. (2016) extended the work of Miller et al. (2015) by introducing a class of Kolchin partition priors (KPPs) for the linkage structure (or cluster assignments) as a way to enforce the microclustering property. However, this formulation is limited by issues of interpretability and identifiability, and also lacks a full characterization of its asymptotic properties. More recently, Betancourt et al. (2020) improved on the weaknesses of the KPP models by proposing a class of prior distributions on random partitions that displays the microclustering property and other desirable characteristics, while preserving computational tractability.

In this paper, we expand on the existing work of microclustering by proposing a new prior distribution based on allelic partitions. This approach is inspired by the structure of the Ewens's sampling formula (Crane et al., 2016), which in turn has strong connections with modern Bayesian nonparametric methods. Specifically, allelic partitions are an equivalent representation of partitions which summarizes the number of clusters of each size. In contrast to the previous microclustering approaches, the most appealing feature of this framework for RL applications is being able to handle directly the distribution of the cluster sizes in a natural fashion. Our proposed class of priors is general, however, and can be adapted and used in other microclustering problems (Bloem-Reddy et al., 2018; Klami and Jitta, 2016).

The remainder of the paper is organized as follows: Section 2 introduces notation and frames RL as a clustering problem. Section 3 discusses in detail the concept of microclustering, introduces two new microclustering properties that require stronger conditions, and presents a more detailed review of previous work. Section 4 discusses our approach based on allelic partitions including inference details. Then, Sections 5 and 6 explore the performance of our approach compared to the ESC models on five simulated data scenarios and three RL applications, respectively. For the applications, we also explore alternatives for optimal point estimation of the partitions. Finally, we discuss our findings and future work directions in Section 7.

2. Record linkage as a clustering task

In this section, we introduce some notation and describe RL from a clustering perspective using a bipartite graph representation of the problem (Steorts et al., 2016). Consider a collection of $J \geq 2$ files. Let $\boldsymbol{x}_{i,j} = (x_{i,j,1}, \ldots, x_{i,j,L})$ be the attribute data associated with the i-th record in file j, and let $\mathbf{X}_j = [x_{i,j,\ell}]$ be the corresponding $n_j \times L$ array for every j. For simplicity, we assume that every record contains L fields in common, field ℓ having D_ℓ levels. Attribute data of this sort may be considered as either categorical or string-valued but here we focus on a model for categorical data. Let us say, for instance, that data about gender, state of residency, and race regarding n_j individuals in file j are available; in this scenario, $\boldsymbol{x}_{i,j}$ is a categorical vector with dimension $L = 3$ whose entries have $D_1 = 2$ (male and female), $D_2 = 51$ (there are 51 states in the United States including DC), and $D_3 = 6$ (White, Black or African-American, American Indian or Alaska Native, Asian, Native Hawaiian or Other Pacific Islander, and some other race) levels, respectively. Hence, we can think of records as L dimensional vectors storing attribute information (L fields), while the j-th file is composed of n_j records.

Now, let $\boldsymbol{y}_k = (y_{k,1}, \ldots, y_{k,L})$ be the vector of "true" attribute values for the k-th latent individual, $k = 1, \ldots, K$, where K is the total number of unique individuals in the J files (K could be as small as 1 if every record in every file refers to the same entity or as large as $n = \sum_j n_j$ if files do not share records at all). Hence, $\boldsymbol{Y} = [y_{k,\ell}]$ is an unobserved $K \times L$ attribute matrix whose k-th row stores the attribute data associated with the k-th latent individual. Next, we define the linkage structure $\boldsymbol{\xi} = (\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_J)$, where $\boldsymbol{\xi}_j = (\xi_{1,j}, \ldots, \xi_{n_j,j})$. Here, $\xi_{i,j}$ is an integer from 1 to K indicating which latent individual the i-th record in file j refers to, which means that $\boldsymbol{x}_{i,j}$ is a possibly-distorted measurement of $\boldsymbol{y}_{\xi_{i,j}}$. Such structure unequivocally defines a partition $\mathcal{C}_{\boldsymbol{\xi}}$ on $\{1, \ldots, n\}$. To see this, notice that by definition, two records (i, j) and (i^*, j^*) correspond to the same individual if and only if $\xi_{i,j} = \xi_{i^*,j^*}$. Therefore, $\mathcal{C}_{\boldsymbol{\xi}}$ is nothing more than a set composed of K disjoint non-empty subsets $\{C_1, \ldots, C_K\}$ such that $\cup_k C_k = \{1, \ldots, n\}$, where each C_k is defined as the set of all records pointing to latent individual k. Hence, the total number of latent individuals $K = K(\boldsymbol{\xi})$ is a function of the linkage structure; specifically, $K = \max\{\xi_{i,j}\}$, since without loss of generality we label the cluster assignments with consecutive integers from 1 to K. Cluster assignments $\xi_{i,j}$ play a fundamental roll in our approach since they define a linkage structure between files.

Fig. 1 shows the linkage structure $\boldsymbol{\xi}$ as a bipartite graph in which each edge links a record to a latent individual. For instance, this figure shows that the sets of records $\boldsymbol{x}_{3,1}$, $\boldsymbol{x}_{4,2}$, $\boldsymbol{x}_{5,2}$ and $\boldsymbol{x}_{1,3}$ correspond to the same individual ($\boldsymbol{y}_4$). This toy example makes clear that linking records to a hypothesized latent entity is at its core a clustering problem where the main goal is to make inferences about the cluster assignments $\boldsymbol{\xi}$. In contrast to other clustering tasks, however, we

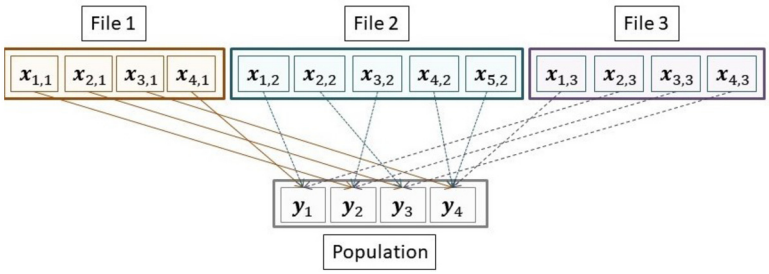

Fig. 1. Bipartite graph representation of RL as a clustering task including records $\boldsymbol{x}_{i,j}$, latent true attributes $\boldsymbol{y}_k$, and the linkage structure (edges) $\boldsymbol{\xi}$.

aim to develop an approach that lets the number of records in each cluster be small even for large data sets – known as *microclusters*, which is characteristic of RL applications (Miller et al., 2015; Zanella et al., 2016). Note that the bipartite graph representation allows for duplicates across and within databases. In practical terms this implies that multiple files can be combined into a single file of size $n = \sum_j n_j$, and we can treat the problem as one of deduplication. Hence, for the remainder of the paper, we drop the file subindex in the notation and simply refer to the attribute data associated with record i as $\boldsymbol{x}_i$, and the linkage structure as $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)$.

As far as the linkage structure $\boldsymbol{\xi}$ is concerned, previous approaches have assumed a uniform prior on $\boldsymbol{\xi}$ conditional on the size of the latent population which is always unknown (Steorts et al., 2016). The uniform prior on $\boldsymbol{\xi}$ induces a prior distribution over partitions where any two partitions with the same number of latent entities are equally likely a priori. Although this prior is convenient because it greatly simplifies computation of the posterior, it requires strong tuning of the latent population size which in turn results in a highly informative prior. For this reason, we devote Sections 3.1 and 4 (the latter introduces our proposal) to characterize prior distributions on partitions that can be used as default priors and intrinsically induce the microclustering behavior desired for RL tasks.

3. Microclustering

Finite mixture models and Dirichlet/Pitman-Yor process mixture models are widely used in many clustering applications (Miller and Harrison, 2018). These models, however, display a sublinear growth of the number of clusters with respect to the number of records. Such a property is unappealing in the context of RL problems because we need to generate a large number of clusters, each with a negligible number of records. In order to formulate more realistic models for de-duplication, Miller et al. (2015) introduce the *microclustering property*. Formally, the definition states the following:

Definition 1. A random partition $\mathcal{C}_{\boldsymbol{\xi}}$ of n elements is said to satisfy the microclustering property if $\frac{M_n}{n} \xrightarrow{\text{p}} 0$ as $n \to \infty$, where $M_n = \max\left\{|C| : C \in \mathcal{C}_{\boldsymbol{\xi}}\right\}$ represents the size of the largest element in $\mathcal{C}_{\boldsymbol{\xi}}$.

That is, the size of the largest cluster in the partition grows sublinearly with n, which in turn implies that the number of clusters grows linearly. Miller et al. (2015) and Zanella et al. (2016) argue that no model on partitions can exhibit the microclustering property, unless its parameters are allowed to vary with n. In addition, the authors show that in order to obtain nontrivial models exhibiting the microclustering property, we must sacrifice either finite exchangeability or projectivity. In the context of random partitions, projectivity means that the distribution of a partition of n elements is the same as the distribution of a partition of m elements restricted to n, for $1 \leq n < m$. In Section 4, we follow Zanella et al. (2016) by sacrificing projectivity, which is less restrictive in the RL context where records are naturally exchangeable. A model for microclustering that sacrifices exchangeability in the context of data with a temporal component is presented in Di Benedetto et al. (2017).

Note, however, that Definition 1 does not necessarily imply that the size of the largest cluster is finite. Indeed, if for example $\mathsf{E}[M_n] \sim \mathcal{O}(\log n)$, a simple application of Markov's inequality shows that

$$\lim_{n\to\infty} \Pr\left[\frac{M_n}{n} > \epsilon\right] \leq \lim_{n\to\infty} \frac{1}{\epsilon}\frac{\mathsf{E}[M_n]}{n} = \frac{1}{\epsilon}\lim_{n\to\infty}\frac{\log n}{n} = 0,$$

i.e., the microclustering property as initially defined in Miller et al. (2015) is satisfied even though the size of the clusters is allowed to grow unboundedly (both a priori and a posteriori). Hence, in the sequel we refer to this as the *weak microclustering property*.

In order to impose further constraints on the cluster sizes a priori, we define the *strong microclustering property* as follows:

Definition 2. A random partition $\mathcal{C}_{\boldsymbol{\xi}}$ is said to satisfy the strong microclustering property if for any $\epsilon > 0$, there exists finite $M, N > 0$ such that $\Pr[M_n > M] < \epsilon$ for all $n > N$, where M_n represents the size of the largest element in $\mathcal{C}_{\boldsymbol{\xi}}$.

Evidently, the strong microclustering property implies the weak microclustering property (again, by a simple application of Markov's inequality), but not viceversa. However, one shortcoming of this definition is that controlling the size of the largest cluster a priori does not necessarily imply that we have controlled its size a posteriori. In RL applications, where we may have prior information about the size of the clusters, we might want to employ priors that impose stronger constraints. Therefore, we introduce the *bounded microclustering property*:

Definition 3. A random partition $\mathcal{C}_{\boldsymbol{\xi}}$ of n elements is said to satisfy the bounded microclustering property if, for some constant M^*, $\Pr[M_n > M^*] = 0$, for all n, where M_n represents the size of the largest element in $\mathcal{C}_{\boldsymbol{\xi}}$.

By definition, $0 < M_n \leq M^*$ almost surely for all n, such that the bounded microclustering property implies both the strong and weak microclustering properties, and ensures that the same behavior holds a posteriori i.e. $\Pr[M_n > M^* \mid \boldsymbol{X}] = 0$. This definition is related to the notion of size-constrained microclustering for finite mixtures discussed in Klami and Jitta (2016), which also assumes that the clusters sizes are bounded in a deterministic fashion. In the remainder of the paper we focus on defining priors that satisfy the bounded microclustering property.

3.1. Existing models for microclustering

The work of Zanella et al. (2016) introduced the idea of Kolchin partition priors (KPPs) as a way to enforce the weak microclustering property (Kolchin, 1971). This approach consists of placing a prior on the number of clusters, $K \sim \boldsymbol{\kappa}$, and then, given K, the cluster sizes $S_1, \ldots, S_K$ with $S_k = |C_k|$ are modeled directly as $S_1, \ldots, S_K \mid K \overset{\text{iid}}{\sim} \boldsymbol{\mu}$. Here $\boldsymbol{\kappa} = (\kappa_s)_{s=1}^{\infty}$ and $\boldsymbol{\mu} = (\mu_s)_{s=1}^{\infty}$ are probability distributions over $\mathbb{N} = \{1, 2, \ldots\}$. In particular, the authors proposed two models: (a) the NBNB model where both $\boldsymbol{\kappa}$ and $\boldsymbol{\mu}$ belong to the Negative-Binomial family, and a more flexible specification (b) the NBD model where $\boldsymbol{\kappa}$ belongs to the Negative-Binomial family and $\boldsymbol{\mu}$ is modeled as a random probability vector with a Dirichlet distribution prior. Conditional on $n = \sum_{k=1}^{K} S_k$, it is straightforward to generate a set of cluster assignments $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)$, which in turn induces a random partition $\mathcal{C}_{\boldsymbol{\xi}} = \{C_1, \ldots, C_K\}$.

One potential issue with this formulation is that the conditioning on n drastically effects the interpretability of $\boldsymbol{\kappa}$ and $\boldsymbol{\mu}$, making the elicitation process difficult when information is available a priori. Additional caveats of the KPPs also include a lack of identifiability and of a clear characterization of their asymptotic properties. In order to overcome these limitations, Betancourt et al. (2020) assumes an Exchangeable Sequence of Clusters (ESC) rather than an exchangeable sequence of data points. Under this construction, the prior distribution on a random partition $\mathcal{C}_{\boldsymbol{\xi}}$ only depends on $\boldsymbol{\mu} = (\mu_s)_{s=1}^{\infty}$ by implicitly conditioning the sequence of exchangeable clusters on the following event

$$E_n = \left\{ \text{there exists } k \in \mathbf{N} \text{ such that } \sum_{j=1}^{k} S_j = n \right\}. \tag{1}$$

Conditional on the occurrence of the event E_n, the random variable K is a function of $(S_1, S_2, \ldots)$ defined as the unique positive integer such that $\sum_{j=1}^{K} S_j = n$. In this case, in contrast to the KPPs, the interpretation of $\boldsymbol{\mu}$ as the distribution of the size of a randomly chosen cluster is not distorted by the direct conditioning on $n = \sum_{k=1}^{K} S_k$. However, generating samples a priori from the ESC models requires the use of a rejection sampler as described in Betancourt et al. (2020, Section 3.3.3). The authors specify two versions of the ESC models similar to those of Zanella et al. (2016): (a) the ESCNB model where $\boldsymbol{\mu} = \mathsf{NegBin}(a, q)$; and (b) the ESCD model where $\boldsymbol{\mu} \sim \mathsf{Dir}(\alpha, \boldsymbol{\mu}^{(0)})$, for α fixed and $\boldsymbol{\mu}^{(0)} = \mathsf{NegBin}(a, q)$. The ESCD is inherently more flexible as it models $\boldsymbol{\mu}$ as a random probability vector with a Dirichlet distribution prior. In both cases, the parameters $a > 0$ and $q \in (0, 1)$ are assigned Gamma and Beta priors, respectively.

In a similar fashion to truncated implementations of traditional Dirichet/Pitman-Yor process mixtures, posterior computations with the ESC models are carried out by generating only the first M components of $\boldsymbol{\mu}$ i.e. $\boldsymbol{\mu} = (\mu_s)_{s=1}^{M}$, for M large. Hence, from a practical perspective, the ESC priors have a similar flavor to the allelic partition priors that we introduce next. Moreover, it is important to note that even though the ESC models were introduced under the scope of the weak microclustering property, they also satisfy the strong microclustering property when the expectation of $\boldsymbol{\mu}$ is finite (i.e. $\sum_{s=1}^{\infty} s\mu_s < \infty$). Our proposal, however, satisfies the bounded microclustering property provided in Definition 3 which is more desirable in practical applications.

In addition to these previous approaches constructed under the microclustering setting, the work of Aleshin-Guendel and Sadinle (2021) recently proposed a Bayesian model for the general setting of multifile record linkage and duplicate detection by constructing a structured prior for partitions that incorporates file information. This prior satisfies the bounded microclustering property but, unfortunately, direct interpretation of the prior on partitions is difficult because the construction is done conditional on the sizes of the data files. In this work, we evaluate the performance of our proposed prior for microclustering, introduced in Section 4, and compare it to the ESC models using both simulated and real data scenarios (see Sections 5 and 6). We limit our comparisons to the ESC models because, similar to our prior proposal, they belong to the class of defaults priors for partitions with microclustering properties.

4. Allelic partition prior

In this section, we introduce a new class of prior distributions on the cluster assignments $\boldsymbol{\xi}$ based on allelic partitions. Let $\mathcal{C}_{\boldsymbol{\xi}} = \{C_1, \ldots, C_K\}$ be the partition implicitly represented by $\boldsymbol{\xi}$ and let $\boldsymbol{r} = (r_1, \ldots, r_n)$ be the allelic partition induced by $\mathcal{C}_{\boldsymbol{\xi}}$, where r_i denotes the number of clusters of size i in $\mathcal{C}_{\boldsymbol{\xi}}$. For example, the set $\{1, 2, 3\}$ yields five possible partitions: $\{\{1, 2, 3\}\}$, $\{\{1\}, \{2, 3\}\}$, $\{\{1, 2\}, \{3\}\}$, $\{\{1, 3\}, \{2\}\}$, $\{\{1\}, \{2\}, \{3\}\}$; which correspond to three possible allelic partitions: $(0, 0, 1)$, $(1, 1, 0)$, $(3, 0, 0)$. This example makes evident that, in general, each partition $\mathcal{C}_{\boldsymbol{\xi}}$ corresponds uniquely to an allelic partition $\boldsymbol{r}$, but the converse is not true. Therefore, allelic partitions define equivalence classes on the space of partitions. The notion of allelic partitions will allow us to construct a flexible model for microclustering by assigning appropriate prior distributions on r_i. The most appealing feature of this framework for RL applications is being able to explicitly calibrate the maximum cluster size and control the distribution of the cluster sizes.

Note that, from the definition of allelic partition, it follows directly that $\sum_{i=1}^{n} i\, r_i = n$ and $\sum_{i=1}^{n} r_i = K$. Similarly to the KPP models (Zanella et al., 2016), the construction of the model based on allelic partitions entails conditioning of n. However, the limitations that arose in that case from this conditioning are overcome in this context by allowing the parameters of the prior distribution on r_i to vary with n in a natural fashion (see Section 4.1). To further illustrate the concept of allelic partition, consider the Ewens-Pitman Prior (EPP, McCullagh and Yang, 2006), which is intrinsically related to the Dirichlet process. The probability mass function for the EPP is given by

$$p(\boldsymbol{\xi} \mid \theta) = \frac{\Gamma(\theta)}{\Gamma(n+\theta)}\, \theta^K \prod_{k=1}^{K} \Gamma(S_k), \tag{2}$$

where θ is an unknown positive parameter. Note that this prior can be factorized as

$$p(\boldsymbol{\xi} \mid \theta) = p(\boldsymbol{\xi} \mid \boldsymbol{r})\, p(\boldsymbol{r} \mid \theta), \tag{3}$$

where $p(\boldsymbol{\xi} \mid \boldsymbol{r}) = \frac{1}{n!} \prod_{i=1}^{n} i!^{r_i}\, r_i!$ is the uniform distribution on all partitions that belong to the equivalence class represented by $\boldsymbol{r}$, and

$$p(\boldsymbol{r} \mid \theta) = \frac{n!}{\theta(\theta+1)\cdots(\theta+n-1)} \prod_{i=1}^{n} \frac{\theta^{r_i}}{i^{r_i}\, r_i!},$$

has support on all possible allelic partitions of the set $\{1, \ldots, n\}$. This representation of the EPP directly motivates the structure of our allelic priors for microclustering. In particular we preserve the same structure for $p(\boldsymbol{\xi} \mid \boldsymbol{r})$ (which ensures that the prior is finitely exchangeable for any n), and replace $p(\boldsymbol{r})$ with a distribution that places its probability on the kind of allelic partitions that are consistent with microclustering applications.

In particular, in the sequel we focus on the bounded microclustering property. Let $M^* = \max\{i \in [n] : r_t = 0, \text{ for all } t > i\}$, $M^* \ll n$, be the size of the largest cluster in $\mathcal{C}_{\boldsymbol{\xi}}$, i.e., let M^* represent the maximum number of times any one unique record can be repeated in the data set. Our strategy consists in fixing M^* to a reasonable value, and then, placing a distribution on $\boldsymbol{r}$ that reflects our prior believes, such that $\Pr[r_t = 0] = 1$ for all $t > M^*$. It should be clear that, by fixing M^*, this approach satisfies the bounded microclustering property, and consequently the strong and weak properties as well. This type of hard constraint could be of particular practical use in RL scenarios where, due to the data collection mechanism, it is known a priori that there are no duplicates within databases. In that case, the maximum cluster size is expected to be restricted to the number of databases available for deduplication. In cases where there is no strong prior information about the size of the clusters or one wishes to be less restrictive a priori, the value of M^* can be chosen to be relatively large to allow for more flexibility (see section 6 for illustrations). Moreover, the number of singletons and the number of latent individuals are easy to calibrate, which is very appealing for RL settings where prior information is available at such a scale.

4.1. Beta binomial allelic prior (BBAP)

In this section, we describe one possible specification of the distribution of the allelic partition for bounded microclustering. In order to specify $p(\boldsymbol{r})$, we first factorize the joint distribution as

$$p(\boldsymbol{r}) = p(r_{M^*})\, p(r_{M^*-1} \mid r_{M^*})\, p(r_{M^*-2} \mid r_{M^*-1}, r_{M^*}) \ldots p(r_1 \mid r_2, \ldots, r_{M^*}).$$

Moreover, we assume conditional Binomial distributions for the cluster sizes,

$$r_{M^*} \sim \mathsf{Bin}(\lfloor n/M^* \rfloor, \theta_{M^*}) \quad \text{and} \quad r_t \mid r_{t+1}, \ldots, r_{M^*} \sim \mathsf{Bin}(Q_t(r_{t+1}, \ldots, r_{M^*}), \theta_t),$$

where the number of trials follow the recursive specification

$$Q_t(r_{t+1}, \ldots, r_{M^*}) = \left\lfloor \left(n - \sum_{i=t+1}^{M^*} i\, r_i \right) \Big/ t \right\rfloor,$$

for $t = 2, \ldots, M^* - 1$. Finally, $r_1 = n - \sum_{i=2}^{M^*} i\, r_i$ which means that $r_1 \mid r_2, \ldots, r_{M^*} \sim \delta_{Q_1}$. It is important to note that this particular specification yields cluster size distributions that are consistent with the definitions $\sum_{i=1}^{n} i\, r_i = n$ and $\sum_{i=1}^{n} r_i = K$. For instance, for $M^* = 2$, we can at most observe $\lfloor n/2 \rfloor$ clusters of size two in a data set of size n.

In addition, the parameters θ_t control the proportion of clusters of size t that we expect to observe in the partition. Because the parameters $\theta_2, \ldots, \theta_{M^*}$ play such a critical role in the model, we increase the versatility of the prior by letting $\theta_t \sim \mathsf{Beta}(a_t, b_t)$, allowing greater control on both the prior mean and the prior variance of each r_t. We refer to this prior formulation as the Beta Binomial Allelic Prior (BBAP).

As an example, consider the case of $M^* = 2$. Here, it is straightforward to see that the corresponding allelic partition becomes $\boldsymbol{r} = (n - 2r_2, r_2, 0, 0, \ldots, 0)$, which allow us to formulate a hierarchical prior for $\boldsymbol{\xi}$ only in terms of the number of clusters of size two (r_2). Thus, if $M^* = 2$ and we denote $a_2 = a$ and $b_2 = b$, we have that

$$p_{BBAP}(\boldsymbol{\xi} \mid a, b) = \frac{(n - 2r_2)!\, 2^{r_2}\, r_2!}{n!} \frac{\Gamma(\lfloor n/2 \rfloor + 1)}{\Gamma(r_2 + 1)\, \Gamma(\lfloor n/2 \rfloor - r_2 + 1)} \frac{\Gamma(r_2 + a)\, \Gamma(\lfloor n/2 \rfloor - r_2 + b)}{\Gamma(\lfloor n/2 \rfloor + a + b)} \frac{\Gamma(a + b)}{\Gamma(a)\, \Gamma(b)}, \tag{4}$$

where $\Gamma(\cdot)$ represents the gamma function. In this case, the expected number of singletons a priori is

$$\mathsf{E}[r_1] = n - 2 \left\lfloor \frac{n}{2} \right\rfloor \frac{a}{a + b} \approx \frac{bn}{a + b},$$

with variance

$$\mathsf{Var}[r_1] = 4 \left\lfloor \frac{n}{2} \right\rfloor \left(a + b + \left\lfloor \frac{n}{2} \right\rfloor \right) \frac{ab}{(a + b)^2 (a + b + 1)}.$$

As we discussed before, the number of singletons is one of the quantities for which there is often strong prior information in RL problems. Therefore, these expressions are key for prior calibration. In fact, more generally

$$\mathsf{E}[r_{M^*}] = \frac{a_{M^*}}{a_{M^*} + b_{M^*}} \left\lfloor \frac{n}{M^*} \right\rfloor$$

and

$$\mathsf{E}[r_t] = \frac{a_t}{a_t + b_t} \sum_{s_{t+1}=0}^{Q_{t+1}} \cdots \sum_{s_{M^*}=0}^{Q_{M^*}} Q_t\, q\,(s_{t+1}, \ldots, s_{M^*}),$$

where

$$q\,(s_{t+1}, \ldots, s_{M^*}) = \mathsf{BetaBin}(s_{M^*} \mid \lfloor n/M^* \rfloor, a_{M^*}, b_{M^*}) \prod_{k=t+1}^{M^*-1} \mathsf{BetaBin}(s_k \mid Q_k, a_k, b_k), \tag{5}$$

for $Q_k \equiv Q_k(s_{k+1}, \ldots, s_{M^*})$ and $t = 2, \ldots, M^* - 1$. These expressions are too convoluted to be of real practical utility but could be easily computed via simulations. In the following section, we provide some practical guidelines to calibrate the hyperparameters of the model to prior knowledge in a simple manner.

4.2. BBAP calibration

In general, for RL applications where the percentage of duplication is low, we would like θ_t to decrease fast with t to reflect the fact that we expect most items to be singletons. On the other hand, when attempting to combine J files in which we expect substantial overlap, we would typically pick $M^* \geq J$ and use relatively large values of θ_J. For example, in the case $M^* = 2$, given a prior probability of duplication π (often less than 0.3 in many deduplication settings) along with a corresponding coefficient of variation γ (e.g., $\gamma = 0.5$ for vague levels of precision), it is straightforward to see that by letting

$$a_2 = \frac{1 - \pi(1 - \gamma^2)}{\gamma^2} \quad \text{and} \quad b_2 = a_2 \frac{(1 - \pi)}{\pi},$$

we obtain the desired prior calibration. For $M^* > 2$, a similar procedure can be implemented using numerical computations that leverage the recursive nature of the prior. More specifically, after providing a vector of prior probabilities for the cluster sizes $\boldsymbol{\pi} = (\pi_2, \ldots, \pi_{M^*})$ based on prior knowledge, the elicitation of the hyperparameters a_t and b_t can be done recursively according to the coefficient of variation chosen by the practitioner. See Appendix A for details.

Considering that many RL applications display a distribution of cluster sizes with a 'geometric like' decay i.e. a large number of singleton clusters is expected (Sadinle, 2014, 2017; Steorts et al., 2016), we explore a default calibration of the BBAP that exhibits this behavior. The prior is calibrated assuming values for the prior probabilities of the clusters of each size from a truncated Geometric distribution, $\boldsymbol{\pi} = \mathsf{Geom}(p)$. We also consider a truncated Negative Binomial, $\boldsymbol{\pi} = \mathsf{NegBin}(r, p)$

with the purpose of assessing the sensitivity of the results to the prior calibration. Furthermore, in cases where the data collection mechanism naturally informs the maximum cluster size, for example merging J databases known to have no duplication within, we can choose $M^* = J$ to obtain sensible RL results. Note, however, that records within the same file can still be assigned to the same cluster. In particular, for the case of $M^* = 2$ with no duplicates within, Sadinle (2017) proposes a bipartite matching which imposes a one-to-one correspondence of records across the two files and eliminates the possibility of duplicates within files a priori. Because of the latter, the approach of Sadinle (2017) is more appropriate for this setting. Otherwise, for the multifile case when there is no strong prior information about the size of the clusters or one wishes to be less restrictive a priori, the value of M^* can be chosen to be relatively large to allow the maximum cluster size to be estimated from the data without risk of truncation a priori. The choice of M^* is important but its effect on posterior results also depends on the prior calibration. Sections 5 and 6 include illustrations of different prior calibrations and their effects on posterior inference. Refer to Appendix B for comparisons of the BBAP model to the approach of Sadinle (2017) and a detailed description of sensitivity of posterior results to values of M^* for different prior calibrations.

4.3. Posterior inference for BBAP model

In order to obtain samples from the BBAP model a posteriori, we derive the probability distribution of a record being assigned to an existing or new cluster conditional on the current partition of the data and the prior parameters. This type of assignment rule has been widely used in the context of Dirichlet/Pitman-Yor processes and it is especially useful for computational tractability in sampling of random partitions. For non-projective models like the BBAP model, we refer to these cluster assignment probabilities as *reallocation probabilities*. Given the conditional EPPF in equation (3) and that

$$p(\xi_i \mid \boldsymbol{\xi}_{-i}, \boldsymbol{r}) = \frac{p(\boldsymbol{\xi} \mid \boldsymbol{r})}{p(\boldsymbol{\xi}_{-i} \mid \boldsymbol{r}_{-i})} \frac{p(\boldsymbol{r})}{p(\boldsymbol{r}_{-i})},$$

the reallocation probabilities for the BBAP model are given by

$$p(\xi_i = k \mid \boldsymbol{\xi}_{-i}, \boldsymbol{r}_{-i}) \propto \begin{cases} (|k|+1)\, \dfrac{r_{-i,|k|+1}+1}{r_{-i,|k|}}\, \dfrac{p(\boldsymbol{r})}{p(\boldsymbol{r}_{-i})} & \text{if } k = 1, \ldots, K_{-i}, \\ (r_{-i,1}+1)\, \dfrac{p(\boldsymbol{r})}{p(\boldsymbol{r}_{-i})} & \text{if } k = K_{-i}+1\,, \end{cases} \tag{6}$$

where $|k| = 1, \ldots, M^* - 1$ is the size of cluster k, and $r_{-i,|k|}$ and K_{-i} are the number of clusters of size $|k|$ and the total number of clusters in $\mathcal{C}_{\boldsymbol{\xi}} \setminus i$, respectively. While the term $p(\boldsymbol{r})/p(\boldsymbol{r}_{-i})$ cannot be readily simplified, its evaluation is straightforward and has a low computational cost.

Posterior inference using the BBAP is performed by introducing the corresponding likelihood terms of the RL model in the reallocation probabilities. Given that standard Gibbs sampling algorithms are too slow for large data sets with many small clusters, we utilize a modified version of the Chaperones Algorithm initially proposed in Miller et al. (2015) to obtain samples from the full conditional distribution of $\boldsymbol{\xi}$. The Chaperones algorithm is similar in spirit to existing split–merge Markov chain sampling algorithms (Jain and Neal, 2004) but exhibits better mixing properties in microclustering settings. The modified version that we implement accelerates the convergence of the algorithm by using a non-uniform proposal to select the 'chaperone records' that favors records with common field values while still assigning probabilities greater than zero to all possible record pairs (Betancourt et al., 2020). In the following section, we describe a specific RL model formulation used to illustrate our prior proposal. Note, however, that our allelic partition approach is general and can be used with other RL models or adapted to other microclustering applications beyond RL.

4.3.1. Record linkage model

For the simulations and applications presented in the remainder of the paper, we follow the RL model proposed by Steorts et al. (2016). Here, each field is modeled depending on whether it is distorted or not. If $x_{i,\ell}$ is not distorted, that particular field is left intact by giving it a point mass distribution at the true value; otherwise, a categorical (multinomial) distribution is placed over all the categories of that particular field. In summary, assuming that the attribute data $x_{i,\ell}$ are conditionally independent given the cluster assignments ξ_i and the true population attributes $y_{n,\ell}$, we have that:

$$x_{i,\ell} \mid y_{\xi_i,\ell}, w_{i,\ell}, \boldsymbol{\vartheta}_\ell \overset{\text{ind}}{\sim} \begin{cases} \delta_{y_{\xi_i,\ell}}, & w_{i,\ell} = 0; \\ \text{Cat}\,(\boldsymbol{\vartheta}_\ell)\,, & w_{i,\ell} = 1, \end{cases} \tag{7}$$

where δ_a is the distribution of a point mass at a, $w_{i,\ell}$ are distortion indicators, and $\boldsymbol{\vartheta}_\ell$ is a D_ℓ-dimensional vector of multinomial probabilities. We simply let $w_{i,\ell} \mid \psi_\ell \overset{\text{ind}}{\sim} \text{Ber}(\psi_\ell)$ where ψ_ℓ represents the distortion probabilities of the fields, and fix $\boldsymbol{\vartheta}_\ell$ at the empirical distribution of the data. By integrating $w_{i,\ell}$ out, the likelihood in equation (7) is now:

$$x_{i,\ell} \mid y_{\xi_i,\ell}, \psi_\ell, \boldsymbol{\vartheta}_\ell \overset{\text{ind}}{\sim} (1-\psi_\ell)\delta_{y_{\xi_i,\ell}} + \psi_\ell \boldsymbol{\vartheta}_\ell. \tag{8}$$

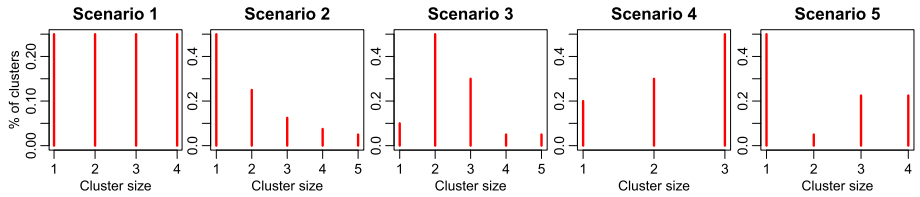

Fig. 2. True allelic partitions for five simulated scenarios with $K = 200$ and $n = 500, 385, 490, 460, 435$, respectively.

In order to complete the model specification, we let $y_{k,\ell} \mid \boldsymbol{\vartheta}_\ell \overset{\text{ind}}{\sim} \text{Cat}(\boldsymbol{\vartheta}_\ell)$ and assign independent priors for the distortion probabilities of the fields, $\psi_\ell \overset{\text{ind}}{\sim} \text{Beta}(c_\ell, d_\ell)$. Finally, we utilize the ESC and BBAP microclustering priors for the linkage structure $\boldsymbol{\xi}$. The distortion parameters ψ_ℓ capture the noise of the data and their values are expected to remain small (usually below 10%) to obtain sensible RL results.

5. Simulation study

In this section, we explore the behavior of the proposed BBAP models, compared to the ESC models, in five different simulation scenarios. These scenarios are chosen to explore the flexibility of the microclustering priors and their ability to recover the true partition beyond the conventional 'geometric like' behavior assumed in many RL applications. In order to evaluate the sensitivity of the results to prior calibration, we use a default Geometric calibration, $\boldsymbol{\pi} = \text{Geom}(p)$ with $p = 0.5$ – denoted as BBAPG, as well as a truncated Negative-Binomial specification, $\boldsymbol{\pi} = \text{NegBin}(r, p)$ with $r = 4$ and $p = 0.5$ – denoted as BBAPNB. The parameter values of the Negative-Binomial specification reflect a prior mode for the distribution of cluster sizes between 2 and 3. In both cases, we use a coefficient of variation of $\gamma = 0.25$ to reflect relatively vague levels of precision in the calibrations (recall the discussion in Section 4.2). For the ESC models, we set $\alpha = 1$, $a \sim \text{Gamma}(1, 1)$, and $q \sim \text{Beta}(2, 2)$. These values have been previously suggested as defaults and shown to work well (Betancourt et al., 2020). We also assume a Beta prior distribution with mean 0.01 and standard deviation of 0.01 for the distortion probabilities of the fields, ψ_ℓ. For computational and comparison purposes, we work with a truncated version of the ESC models in which only the first M^* components of $\boldsymbol{\mu}$ are generated.

For the simulation, we generate ten data sets obtained from the combination of five different partitions of $K = 200$ clusters and two fixed values of the distortion probabilities of the fields – $\psi_\ell = 0.01$ and 0.05. The RL task is expected to become more challenging for higher levels of distortion of the fields. Fig. 2 displays the five partitions which yield datasets of sizes $n = 500, 385, 490, 460, 435$, respectively. All the data sets contain five fields of information with ten categories each generated uniformly at random.

For all the prior distributions, we set M^* to be one and a half times the true maximum cluster size to generate the prior samples. Fig. 3 displays the true data partitions and prior samples from the two BBAP and ESC models for the data sets simulated with $\psi_\ell = 0.05$ (similar behavior is observed for $\psi_\ell = 0.01$). As Fig. 3 shows, scenarios (rows) 1 and 2 display uniform and geometric behavior of the partition, respectively, while scenarios 3 and 4 are more unconventional in that the proportion of singleton clusters is low compared to other cluster sizes. Finally, the cluster size distribution of the last scenario can be thought of as a mixture of the previous ones. The behavior of ESCNB and BBAPG in terms of the number of clusters of each size is quite similar, although the rate of decay for BBAPG seems to be faster. Furthermore, the behavior of the ESCD prior is quite different from that of the alternatives. In particular, ESCD induces very skewed marginal priors for the proportion of clusters of any given size, and favors configurations in which the most frequent cluster size is between 3 and 4. Finally, BBAPNB distributes the proportion of clusters of any given size more evenly but favors cluster sizes between 2 and 3 as expected from the parameter values used for the calibration.

Fig. 4 displays the posterior distribution over allelic partitions for each prior and simulated scenario, compared to the true cluster size distributions, for a distortion probability value of $\psi_\ell = 0.05$. Results for $\psi_\ell = 0.01$ are shown in Appendix E. In addition, Table 2 displays the posterior average Jensen-Shannon (JS) distance between the MCMC samples of the partitions and the true partition, as well as more traditional RL classification error rates, namely, False Negative Rate (FNR) and False Discovery Rate (FDR). The JS distance metric is based on a symmetrization of the Kullback–Leibler divergence, and allows us to evaluate how well the different models recover the true distribution of the allelic partition (Lin, 1991). This is in contrast to the FNR and FDR values, which focus exclusively on pairwise classification. The JS distance values range between 0 and 1, so that values closer to zero are preferred. See Appendix C for more details.

From Table 1, we observe that for the lowest level of distortion of the fields ($\psi_\ell = 0.01$) all the models perform relatively well with FNR and FDR values between 0.1% and 5.5% in all cases. For $\psi_\ell = 0.05$, where the RL task is expected to be more challenging, the error rates range between 4.6% and 13.8% for all simulated scenarios. Overall, ESCNB is the model with the worst performance across all three metrics, specially for simulated scenarios 3 to 5 which have less conventional cluster size distributions. Focusing on the results for $\psi_\ell = 0.01$, where the clustering signal is stronger, we observe that ESCD seems to have the best performance in terms of the JS distance for all scenarios with very similar results from BBAPNB for scenarios 1 and 3. The performance in terms of FNR and FDR of the two BBAP calibrations is very similar to the ESCD for scenarios 1 and 3, while ESCD performs slightly better for the other scenarios. As the noise increases ($\psi_\ell = 0.05$), the

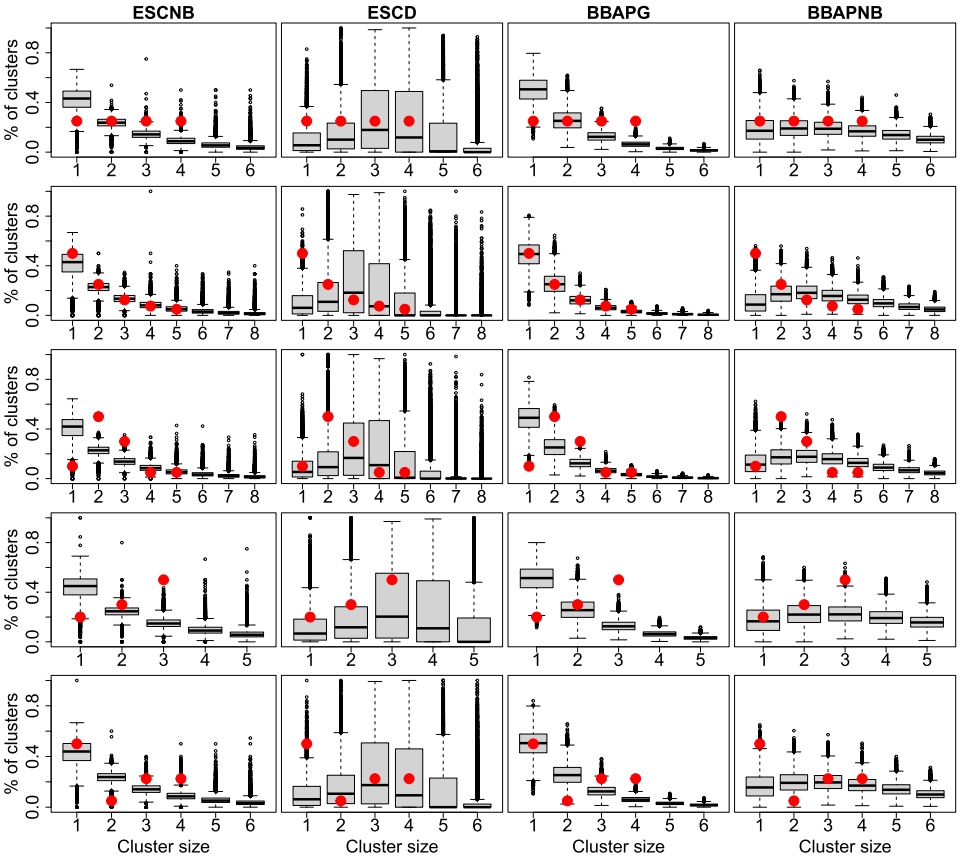

Fig. 3. Prior distribution of the allelic partition (boxplots) for ESC and BBAP models, and true data partition (red dots) for five simulated scenarios with $K = 200$ clusters and distortion probability of the fields $\psi_\ell = 0.05$ (rows). (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

Table 1
Posterior average Jensen-Shannon (JS) distance, FNR and FDR (in percentages) for ESC and BBAP models for five scenarios simulated with distortion levels $\psi_\ell = 0.01$ and 0.05.

		Scenario 1			Scenario 2			Scenario 3			Scenario 4			Scenario 5		
	Prior	JS	FNR	FDR	JS	FNR	FDR	JS	FNR	FDR	JS	FNR	FDR	JS	FNR	FDR
$\psi_\ell = 0.01$	ESCNB	0.026	2.6	1.1	0.037	3.6	3.5	0.043	2.9	0.1	0.056	4.4	1.4	0.063	5.5	2.1
	ESCD	0.019	2.0	1.3	0.019	3.3	2.8	0.016	1.4	0.1	0.013	1.8	0.8	0.024	3.3	0.8
	BBAPG	0.023	2.3	1.2	0.037	3.3	3.7	0.018	1.5	0.1	0.029	2.2	1.1	0.041	4.0	1.4
	BBAPNB	0.020	2.0	1.3	0.039	2.9	4.2	0.015	1.1	0.2	0.026	1.8	1.2	0.040	3.2	1.5
$\psi_\ell = 0.05$	ESCNB	0.087	12.1	5.4	0.045	13.8	9.4	0.115	9.8	8.6	0.122	12.5	7.3	0.141	9.8	8.6
	ESCD	0.050	9.5	4.6	0.040	13.2	9.5	0.053	7.8	8.5	0.041	8.1	5.9	0.080	7.8	8.5
	BBAPG	0.068	10.7	4.8	0.045	12.9	9.7	0.062	8.1	8.1	0.073	8.9	6.8	0.101	8.1	8.1
	BBAPNB	0.059	9.3	5.4	0.048	11.4	11.0	0.051	6.7	9.3	0.068	8.0	7.4	0.090	6.7	9.3

performance of ESCD and the two BBAP models becomes more similar. Overall, we found that posterior results are robust to prior specifications of ψ_ℓ. Note that all the models fail to capture the true maximum cluster size for all scenarios. In particular, the overall lower values of the JS distance for ESCD can be in part explained by its more accurate recovery of the true maximum cluster size compared to the other models (see Fig. 4). Motivated by this and the natural data collection mechanisms of real applications, we explore the performance of a BBAP calibration that incorporates the maximum cluster size as prior information in section 6. Finally, when compared to each other, the results for BBAPG and BBAPNB display a trade-off between FNR and FDR. Although BBAPNB has a slight edge over BBAPG in terms of the JS distance, there is no clear outperforming model across all scenarios. This behavior highlights robustness of the results to different BBAP calibrations.

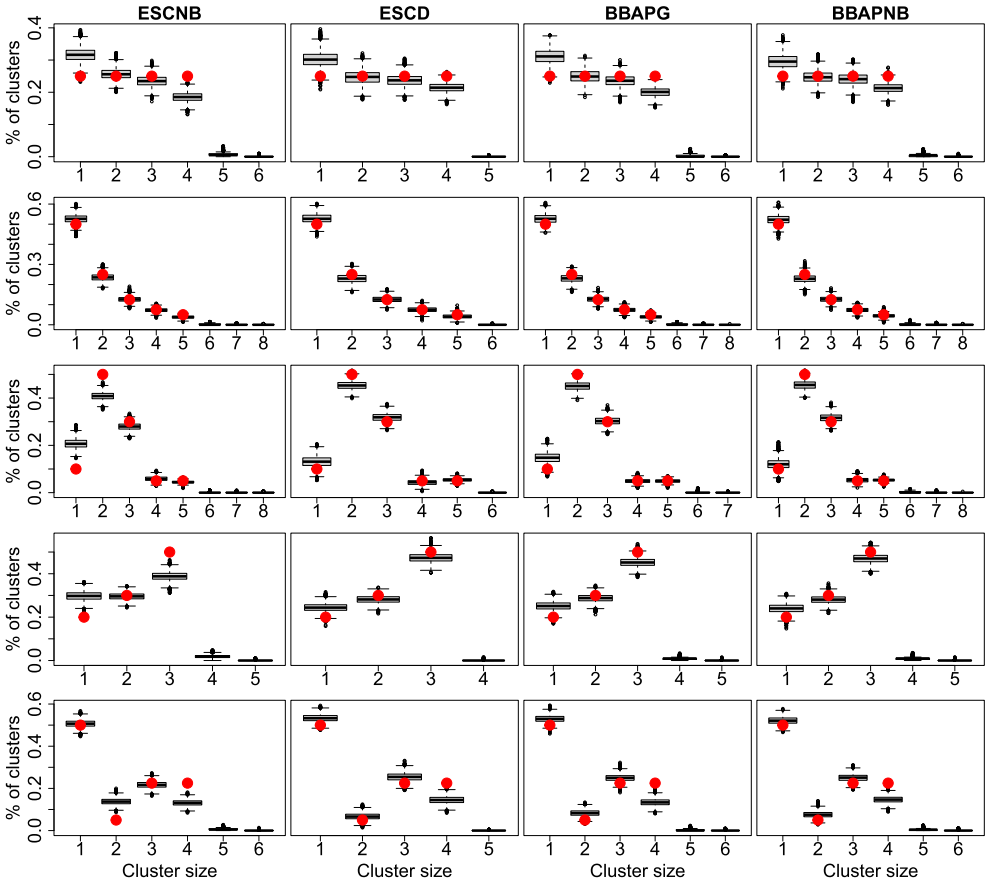

Fig. 4. Posterior distribution of the allelic partition (boxplots) for ESC and BBAP models, and true data partition (red dots) for five simulated scenarios (rows) with $K = 200$ clusters and distortion probability of the fields $\psi_\ell = 0.05$.

6. Applications

In this section, we illustrate the behavior and performance of the BBAP and ESC models using the following three official statistics data sets.

Durham: The North Carolina State Board of Elections (NCSBE) provides snapshots of demographic information of voters which are available to the public (https://ncsbe.gov). Using a snapshot from January of 2019, we consider a data set of 2,714 records of $K = 2,000$ unique registered voters from Durham county. Duplicate records in this data commonly arise from individuals registering to vote after moving from a different county (Kaplan et al., 2018; Wortman, 2019). Ground truth about the partition is available through the NC Voter ID provided by the NCSBE. In order to perform record linkage we employ the following six fields of information: age, sex, race, birth place, and first and last name initials.

SDS: The Social Diagnosis Survey (SDS) is a panel research project that studies indicators of quality of life in households in Poland (http://www.diagnoza.com/index-en.html). We consider a data set of $K = 2,000$ unique individual members of households that participated in the survey in at least one of the years 2011, 2013, and 2015. Duplicate records occur longitudinally across the three waves but not within a specific year for a total of 3,574 records in the data. The data is available in horizontal format providing ground truth for the partition. We use six fields of information for RL: sex, date of birth (day, month and year), province of residence, and education level.

SIPP: The Survey of Income and Program Participation (SIPP) is a longitudinal survey that collects information about the income and participation in federal, state, and local programs of individuals and households in the United States (U.S. Census Bureau, 2009). The data is publicly available through the Inter-university Consortium for Political and Social Research (ICPSR) (https://www.icpsr.umich.edu). We consider a data set of $K = 1,000$ unique individuals interviewed over five waves of the survey performed between 2005 and 2006. The data contains a total of 4,116 records from individuals that are only duplicated across waves (not within). We use five fields of information for RL: sex, year and month of birth, race, and state of residence.

In contrast to the Durham data, the SDS and SIPP datasets intrinsically provide prior information about the expected maximum cluster size in the partition due to their panel structure. Indeed, given the number of waves in each survey we

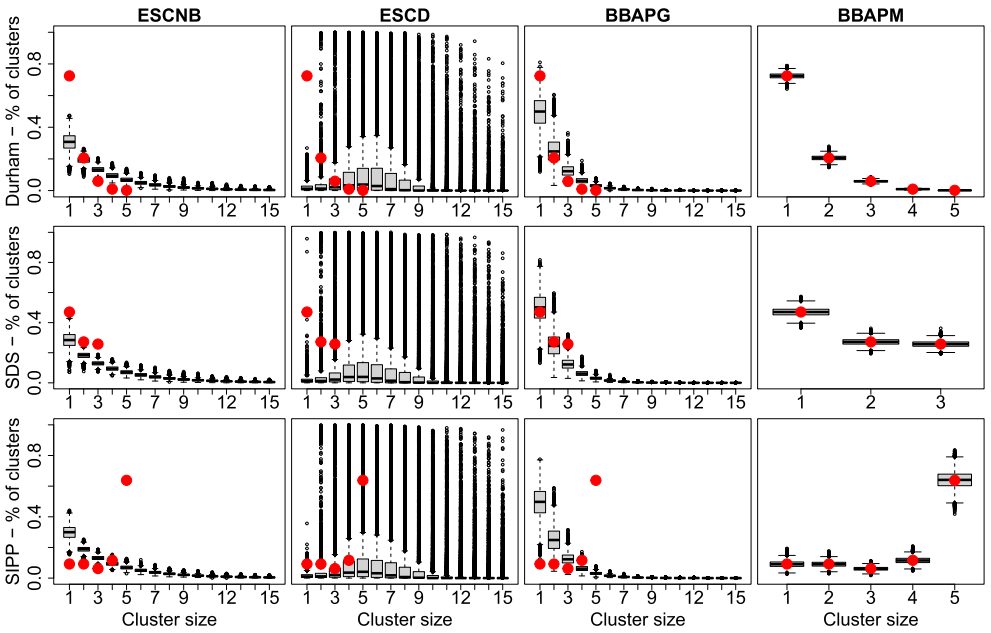

Fig. 5. Prior distribution of the allelic partition (boxplots) and true data partition (red dots) for ESC and BBAP models for Durham, SDS and SIPP data sets.

expect the size of the largest clusters to be three and five for SDS and SIPP, respectively. Although these illustrations do not necessarily reflect the conditions of real data applications where ground truth might not be available, we use these data sets to display the adaptability of the BBAP model to prior knowledge at different scales. For this purpose, we consider two different calibrations of the BBAP model for all datasets. First, the default Geometric specification with $\boldsymbol{\pi} = \text{Geom}(0.5)$ using $M^* = 15$ – denoted as BBAPG. Second, an informed specification where $\boldsymbol{\pi}$ reflects the true data partition and M^* is fixed at the true maximum cluster size – denoted as BBAPM. To perform the elicitation of the hyperparameters, we use coefficients of variation of 25% and 5% for BBAPG and BBAPM, respectively.

For the ESC models, similar to the simulation studies, we set $\alpha = 1$, $a \sim \text{Gamma}(1, 1)$, and $q \sim \text{Beta}(2, 2)$ and work with a truncated version of the ESC models with $M^* = 100$. Finally, we assume a Beta prior distribution with mean 0.01 and standard deviation of 0.01 for the distortion probabilities of the fields, ψ_ℓ, for all the models (see Section 4.3.1). Fig. 5 displays samples from all the prior distributions against the true allelic partition for each dataset (ESC results are shown up to $M^* = 15$ for visibility). Durham data displays the more traditional geometric-like behavior of the true allelic partition, while the SDS and SIPP partitions are less conventional. Evidently, the prior belief for the SIPP data is extremely misspecified under all the non-informed prior models i.e. excluding the BBAPM calibration. Consistent with what we observed in section 5, ESCNB and BBAPG behave similarly (BBAPG has a faster rate of decay), while the ESCD prior behavior is quite different and in this case favors configurations in which the most frequent cluster size is between 5 and 6. On the other hand, the BBAPM calibration is designed to match the true allelic partition quite closely.

All results presented below are based on 20,000 samples from the combination of two chains of 10,000 iterations, obtained after a burn-in period of 10,000 samples for each chain. Traceplots used for convergence diagnostics for the BBAPG model are included in Appendix D.

6.1. Results

Fig. 6 shows the posterior distribution of the number of clusters (i.e., the number of unique individuals in the dataset) under each prior and dataset. Note that the models fail to capture the true number of clusters by consistently overestimating it in all cases. However, BBAPG seems to have a slightly more accurate performance in the Durham and SDS datasets. Interestingly, it is the ESCNB prior that provides the most accurate estimate of the number of unique individuals for the SIPP dataset. This seems to be due to an overestimation in the number of clusters of size 5 (see Fig. 7 and the explanation below).

Fig. 7 displays the posterior distribution over allelic partitions for each prior and data set, and compares them against the truth. In addition, Table 2 displays the posterior average JS distance, FNR and FDR. From Table 2, we observe that the FNR values for the Durham data are the highest for all the datasets (above 13%), compared to values below 5.2% for the SDS and SIPP applications. On the other hand, the FDR values are below 4.4% for all models and data sets. The largest JS distances are observed for the SIPP dataset, while the lowest ones are seen in SDS.

All priors perform similarly for the Durham dataset, which has the more traditional allelic partition distribution. In spite of the very similar performance, BBAPG seems to have a slight edge over ESCNB and ESCD in terms of the mean JS

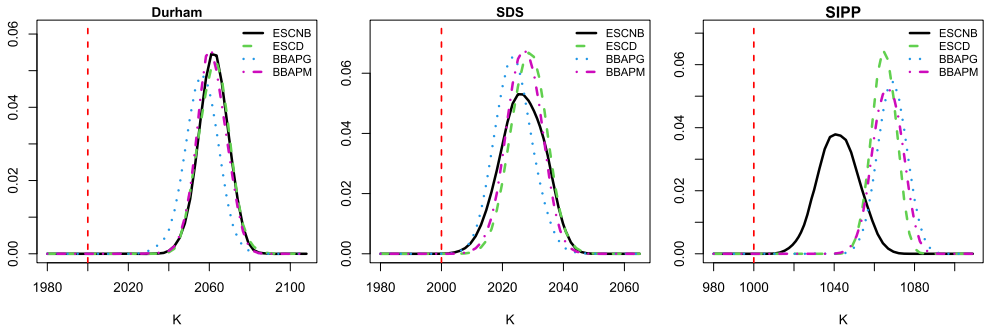

Fig. 6. Posterior distribution of the number of clusters (K) for ESC and BBAP models for Durham, SDS and SIPP data sets. The vertical line represents the true number of clusters in each application.

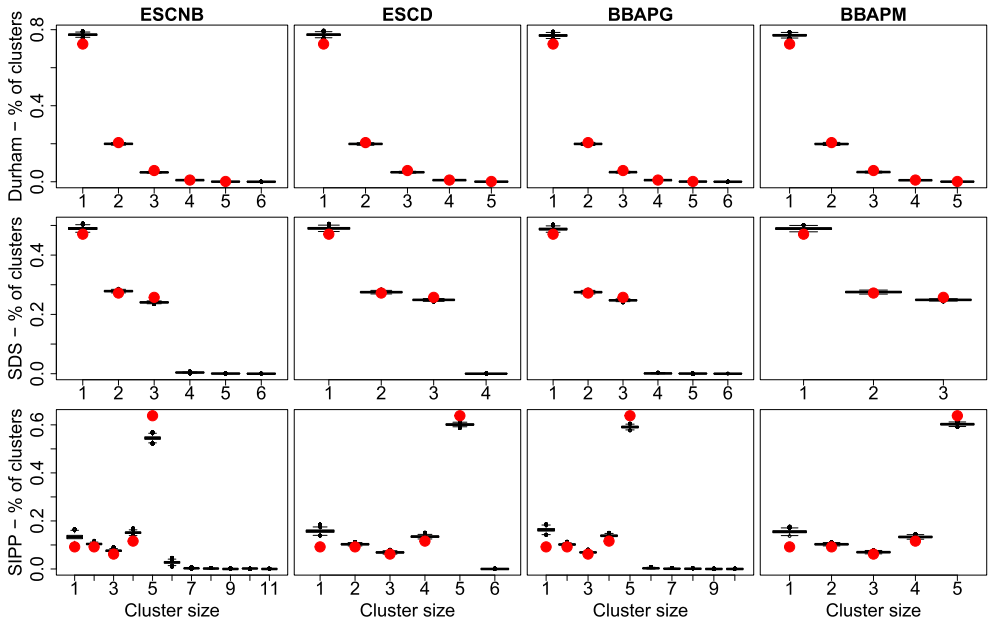

Fig. 7. Posterior distribution of allelic partition (boxplots) and true data partition (red dots) for ESC and BBAP models for Durham, SDS and SIPP data sets.

distance and FNR, at the price of a slightly higher FDR. The reason seems to be that BBAPG is more aggressive in terms of encouraging the creation of non-singleton clusters (see Fig. 5). Note that this result is consistent with our previous observation that BBAPG seems to perform slightly better in terms of estimating the number of unique individuals in the sample for this dataset. BBAPM (the "informed" prior) has a very similar performance to BBAPG in the Durham dataset. On the other hand, in the SDS and SIPP datasets, ESCNB tends to underperform across all three metrics. Among the other two "uninformed" models, ESCD seems to have the best performance in terms of the JS distance and FDR, but the behavior in terms of the FNR is very similar to that of BBAPG. Finally, the behavior of BBAPM is very similar to that ESCD in these two datasets, although BBAPM seems to exhibit a slightly better FNR and FDR than ESCD for the SIPP dataset. Consistently with the simulation study, we find that overall our proposed prior performs better than ESCNB and displays competitive results compared to ESCD. See Appendix D for posterior estimates of the distortion probabilities of the fields under the BBAPG model.

6.2. Point estimation

As in other clustering applications, finding a unique optimal partition of the data is of interest for RL problems. In many cases, in addition to estimation of the number of unique entities, RL is also a required preprocessing step for subsequent statistical analysis with the linked data (Gutman et al., 2013; Sadinle, 2014, 2018; Hof et al., 2017; Kaplan et al., 2018). In the microlustering context, summarizing the information provided by a sample of partitions into an optimal partition is specially challenging. The large number of small clusters expected in the posterior samples of the partition and the high-

Table 2
Posterior average Jensen-Shannon (JS) distance, FNR and FDR (in percentages) for ESC and BBAP models for the Durham, SDS and SIPP data sets.

	Durham			SDS			SIPP		
Prior	JS	FNR	FDR	JS	FNR	FDR	JS	FNR	FDR
ESCNB	0.025	13.7	3.5	0.042	4.1	2.7	0.129	5.2	4.4
ESCD	0.028	13.9	3.4	0.011	3.8	1.7	0.067	4.8	1.8
BBAPG	0.023	13.0	4.1	0.025	3.7	2.2	0.084	4.9	2.1
BBAPM	0.024	13.3	3.8	0.011	3.8	1.7	0.066	4.6	1.7

Table 3
Estimated number of clusters (K), JS distance, FNR and FDR for point estimates of the partitions obtained with a greedy EPL algorithm for Binder's (B), Normalized Information Distance (NID), and Variation of Information (VI) loss functions for Durham, SDS and SIPP. True number of clusters is K=2,000 for Durham and SDS, and K=1,000 for SIPP.

		Durham				SDS				SIPP			
	Prior	K	JS	FNR	FDR	K	JS	FNR	FDR	K	JS	FNR	FDR
B	ESCNB	2063	0.025	12.4	1.9	2024	0.042	4.0	2.7	1053	0.104	4.3	2.8
	ESCD	2060	0.024	12.1	2.0	2026	0.010	3.4	1.3	1066	0.068	4.6	1.6
	BBAPG	2057	0.023	11.5	2.2	2019	0.025	3.0	1.8	1072	0.083	4.7	1.6
	BBAPM	2062	0.025	11.7	1.5	2028	0.011	3.4	1.2	1070	0.076	4.4	1.3
NID	ESCNB	2063	0.028	12.7	3.1	2027	0.045	4.2	3.1	1064	0.137	5.0	6.2
	ESCD	2062	0.025	12.2	1.7	2033	0.027	3.7	1.9	1070	0.089	4.7	2.6
	BBAPG	2057	0.025	11.7	3.3	2023	0.034	3.4	2.6	1081	0.109	4.6	2.9
	BBAPM	2062	0.026	11.8	2.0	2031	0.032	3.8	2.0	1077	0.100	4.7	2.8
VI	ESCNB	2031	0.057	12.4	16.0	1908	0.159	4.0	25.2	971	0.200	4.4	20.4
	ESCD	2023	0.062	12.1	18.7	1912	0.154	3.4	24.4	982	0.191	4.3	19.3
	BBAPG	2016	0.067	11.7	20.7	1892	0.163	3.0	26	982	0.200	4.2	20.2
	BBAPM	2030	0.056	11.7	15.6	1932	0.139	3.5	21.6	975	0.194	4.1	20.0

dimensionality of the space in real data scenarios compels the use of scalable algorithms to find the optimal partition. Decision theoretical approaches for optimal Bayesian estimation that rely on loss functions functions for the space of partitions include those of Lau and Green (2007), Wade and Ghahramani (2018) and Rastelli and Friel (2018). Here, we utilize the approach of Rastelli and Friel (2018) who proposes a scalable greedy algorithm to minimize the expected posterior loss (EPL) under different loss functions. Table 3 displays the estimated number of clusters (K), the JS distance and the error classification rates for the point estimates of the partition obtained with the greedy EPL algorithm for Binder's (B), the Normalized Information Distance (NID), and the Variation of Information (VI) loss functions. Results are based on the last 2,000 iterations of the posterior samples using the `R` package `GreedyEPL` (Rastelli, 2017).

Based on the results for all three data sets and prior distributions, we find that using the VI loss in the microclustering context of RL is not advisable. The greedy EPL with VI loss consistently underestimates the number of clusters compared to B and NID. Even though this behavior leads VI to an estimated number of clusters that is closer to the truth for Durham and SIPP, the overclustering also results in overinflated JS distance and FDR values (above 15% for all datasets and priors). As discussed in Rastelli and Friel (2018), the greedy EPL with VI loss is the most suitable for conventional clustering applications where the number of clusters is not too large. For microclustering, however, we observe that greedy EPL with Binder's loss yields the best performance in terms of lower JS distance and error rates as it tends to slightly overestimate the number of clusters compared to VI. The performance of greedy EPL with NID comes as a close second to Binder's loss for all data sets and priors.

7. Discussion

We have developed a new prior specification for the linkage structure in record linkage problems based on allelic partitions and introduced a set of novel microclustering conditions. Our main contribution is proposing a prior distribution that satisfies the bounded microclustering property introduced in Definition 3, is computationally tractable and permits easy incorporation of prior information. As discussed in Section 3, the ESC models only satisfy the weak microclustering property (Definition 1) which allows the size of the largest cluster to grow unboundedly both a priori and a posteriori. Our BBAP specification overcomes this limitation and in practical terms, consistently outperforms the ESCNB model and shows competitive results compared to the ESCD (with any calibration). In summary, we proposed a prior distribution for random partitions that achieves desirable theoretical guarantees for the record linkage setting and performs competitively in terms of posterior inference compared to state-of-the-art priors in this context. We also want to point out that exploring the

behavior of the models a priori is considerably more difficult for the ESC models. Generating prior samples from the ESC models requires the implementation of a rejection sampler (Betancourt et al., 2020, section 3.3.3), while samples from the BBAP model are easily obtained by sampling directly from Beta and Binomial distributions. Given that the two models perform similarly in practical terms, we leave it to the practitioners to choose between the ESCD and BBAP models depending on the desired theoretical guarantees for large n.

Our work opens up several doors for future research. Scalability is still the main challenging aspect of big data applications of RL involving Bayesian models. Real world data sets, such as the NCSBE voter registration data discussed in Section 6, can contain millions of records leading to a high-dimensional space of partitions. A crucial aspect of future work involves the development of computational algorithms for efficient posterior inference in the microclustering setting using, for example, Metropolis-Hastings (MH) schemes with better properties (Zanella, 2019) or fast computation techniques in the domain of variational approaches (Broderick and Steorts, 2014; Blei et al., 2017). The computational limitations also extend to the implementation of scalable algorithms for optimal Bayesian estimation of the partitions in such high-dimensional discrete spaces.

Acknowledgements

Funding: This work was supported by the National Science Foundation grant numbers 2051911, 1738053, 2023495 and 2027846.

Appendix A. BBAP calibration

In order to calibrate the BBAP model for a specific value of n, we need M^* and a vector of probabilities of the cluster sizes $\boldsymbol{\pi} = (\pi_2, \ldots, \pi_{M^*})$ representing prior knowledge about the partition. Here, $\boldsymbol{\pi}$ provides information about the duplicate percentage expected in the data (π_1 is simply $1 - \sum_{i=2}^{M^*} \pi_i$). When prior information is not available, we propose to use a default calibration with a decaying distribution on the cluster sizes. For example, choosing $\boldsymbol{\pi} \propto \mathsf{Geom}(p)$ for M^* implies that:

$$\pi_i = \frac{g(i|p)}{\sum_{j=1}^{M^*} g(j|p)} \text{ for } i = 2, \ldots, M^*,$$

where $g(i|p)$ is the zero-truncated Geometric density function with parameter p. In order to calibrate the BBAP model to this prior information, we assume that the expected proportion of clusters of size i is π_i such that the expected number of clusters of size i is $\mathsf{E}[r_i] = \pi_i K_{\boldsymbol{\pi}}$, where $K_{\boldsymbol{\pi}}$ represents the expected number of clusters under $\boldsymbol{\pi}$. Since we are calibrating the prior conditional on n and $\sum_{i=1}^{n} i\, r_i = n$, we can roughly estimate the expected number of clusters as $K_{\boldsymbol{\pi}} = \frac{n}{\sum_{i=1}^{M^*} i\, \pi_i}$.

Now, according to the BBAP definition, we have that $r_{M^*} \sim \mathsf{Bin}(\lfloor n/M^* \rfloor, \theta_{M^*})$. This implies that the expected number of cluster of size M^* is $\mathsf{E}[r_{M^*}] = \lfloor n/M^* \rfloor \theta_{M^*} = \pi_{M^*} K_{\boldsymbol{\pi}}$. From these expression we estimate θ_{M^*} as $\pi_{M^*} K_{\boldsymbol{\pi}} / \lfloor n/M^* \rfloor$. Finally, we find values a_{M^*} and b_{M^*} of the hyper-parameters of the Beta distribution such that the mean is centered at $\pi_{M^*} K_{\boldsymbol{\pi}} / \lfloor n/M^* \rfloor$ with the desired coefficient of variation e.g. $\gamma = 0.25$. This same process is repeated for $i = M^* - 1, \ldots, 2$ using the recursive formulas of the BBAP model definition. This specific approach for calibration is practical and involves minimal computational cost but is not unique or optimal in a formal sense.

We chose Geometric and Negative Binomial default calibrations of the prior for comparison purposes with the ESC models, in addition to the decaying behavior on the tail of these distributions which represents a small number of cluster of larger sizes. Practitioners can freely specify any other type of behavior through the $\boldsymbol{\pi}$ values. Naturally, if a practitioner believes that clusters of large sizes (compared to n) are expected in their specific application, the microclustering setting does not apply and traditional models (Dirichlet/Pitman-Yor process) could be used instead.

Appendix B. Comparison with Sadinle (2017) and sensitivity analysis for M^*

In order to illustrate and compare the performance of our BBAP model and the proposal of Sadinle (2017) for $J = M^* = 2$ files with no duplicates within, we simulate data with $K = 200$ clusters (as in our other simulated scenarios in Section 5) where 150 clusters have size one and 50 clusters have size two. This results in $n = 250$ records generated with five fields of information with ten categories each and a distortion level $\psi_\ell = 0.01$. To implement Sadinle's approach, we split the data into two files of equal sizes (=125 records) such that duplicates are only found across files. In this particular simulated scenario, we observed error rates of 5.6% and 1.8% for the BBAP model and 1.7% and 6.8% for Sadinle (2017), respectively for FNR and FDR. We also simulated an additional data set with $K = 400$ with 300 clusters of size one and 100 clusters of size two for a total of $n = 500$ records. We maintained five fields of information with ten categories each and a distortion level $\psi_\ell = 0.01$. In this case, we obtained error rates of 8.2% and 1.4% for BBAP, and 4.4% and 5.4% for Sadinle (2017). Clearly, there is a tradeoff between FNR and FDR between the two approaches but it is no clear that one model will consistently outperform the other based on these simulations. Our disadvantage, however, might increase with the number of records in the data and the level of noise since we are not reducing the space of possible links a priori.

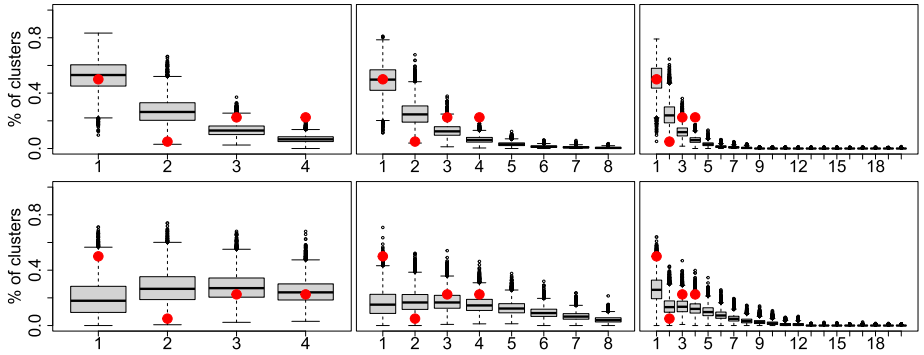

Fig. B.8. Prior distribution of the allelic partition (boxplots) for BBAPG (top) and BBAPNB (bottom) models for different values of $M^* = 4, 8, 20$ for simulated scenario (red dots) with $n = 435$, $K = 200$, and true maximum cluster size of 4.

Table B.4
Posterior average FNR and FDR (in percentages) for BBAP models for simulated scenario with distortion level $\psi_\ell = 0.01$, $n = 435$, $K = 200$, and true max cluster size of 4.

		$M^* = M_n$		$M^* = 2M_n$		$M^* = 5M_n$	
Scenario	Prior	FNR	FDR	FNR	FDR	FNR	FDR
1	BBAPG	3.3	2.5	2.8	3.4	2.8	3.4
	BBAPNB	2.7	2.6	2.3	3.6	2.3	3.6
2	BBAPG	4.9	3.6	4.5	4.4	4.5	4.5
	BBAPNB	4.5	3.8	3.8	4.9	3.8	4.9
3	BBAPG	1.6	0.4	1.7	0.4	1.6	0.4
	BBAPNB	1.2	0.5	1.2	0.6	1.2	0.5
4	BBAPG	2.9	0.6	3.0	1.2	3.0	1.2
	BBAPNB	2.6	0.6	2.6	1.1	2.7	1.2
5	BBAPG	3.8	0.7	3.7	1.6	3.7	1.6
	BBAPNB	3.0	0.7	2.9	1.8	2.9	1.8

For the multifile case, in order to illustrate the effect of the choice of M^* combined with the prior calibrations, we generate prior samples from the BBAP model for scenario 5 in Fig. 2. Note that samples from the prior only depend on $n = 435$ and the choice of M^*. Fig. B.8, shows prior samples of the BBAP model (boxplots) calibrated with a Geometric distribution with parameter $p = 0.5$ and a Negative Binomial distribution with parameters $r = 4$ and $p = 0.5$ (both zero-truncated) for values of $M^* = 4, 8, 20$. These values of M^* correspond to 1, 2, and 5 times the true maximum cluster size of the partition, $M_n = 4$, specified in scenario 5 (red dots).

In the case of a Geometric calibration with $p = 0.5$ (top of Fig. B.8), the BBAP assigns a very small probability to clusters of sizes greater than eight. Therefore, even when we choose $M^* = 5M_n = 20$ the prior probabilities of clusters of sizes greater than eight remain small. A similar behavior is observed for the Negative Binomial calibration where the probabilities of clusters of sizes greater than twelve are very small. This same pattern also occurs for the other simulated scenarios. Table B.4 displays the posterior error rates for five simulated data sets generated with a distortion level $\psi_\ell = 0.01$ and the respective true partitions in Fig. 2. For the two BBAP calibrations in scenario 3, we observe that the rates remain stable for the different values of M^*. In scenarios 1, 4, and 5, the FNRs remained stable while the FDRs increased about 1% when M^* changed from the true maximum cluster size to twice its size. For scenario 2, we observe a tradeoff between FNR and FDR with slightly higher rates for larger M^* values. Overall, for $M^* = 5M_n$ the error rates remained unchanged compared to the results for $M^* = 2M_n$. In conclusion, posterior results are affected by the choice of M^* in combination with the prior calibration but the results are relatively robust to variations of both.

Appendix C. Performance metrics

Given the ground truth about the linkage structure, there are four possible ways of how predictions about pairs of records can be classified: correct links (true positives, TP), correct non-links (true negatives, TN), incorrect links (false positives, FP), and incorrect non-links (false negatives, FN). In order to summarize the classification error of the microclustering approaches, we utilize the false negative rate (FNR) and false discovery rate (FDR) given by

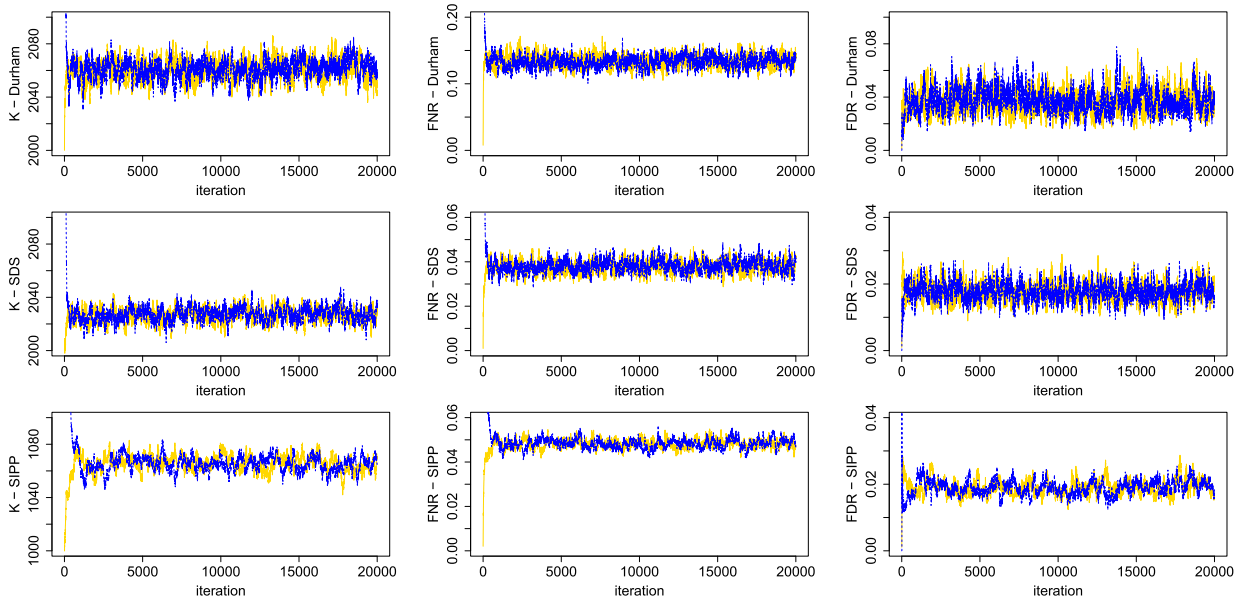

Fig. D.9. Trace plots of number of clusters (K), false negative rate (FNR) and false discovery rate (FDR) for two chains of 20,000 iterations of the BBAPG model for Durham, SDS and SIPP data sets, respectively.

Table D.5
Posterior estimates of the distortion probabilities for Durham, SDS and SIPP data sets using the BBAPG prior.

	$\mathsf{E}[\psi_1]$	$\mathsf{E}[\psi_2]$	$\mathsf{E}[\psi_3]$	$\mathsf{E}[\psi_4]$	$\mathsf{E}[\psi_5]$	$\mathsf{E}[\psi_6]$
Durham	0.00302	0.00334	**0.04826**	0.00484	0.00320	0.00817
SDS	0.00108	0.00054	0.00088	0.01217	0.00592	**0.09154**
SIPP	**0.03819**	0.00087	0.00057	0.00279	0.00156	–

$$FNR = \frac{FN}{FN + TP} \quad \text{and} \quad FDR = \frac{FP}{FP + TP},$$

such that the goal is to achieve values as close to zero as possible.

Moreover, to compare the posterior distribution of the partition (P) with the distribution associated with the true linkage structure (Q), we use the Jensen-Shannon distance metric (Lin, 1991) based on a symmetrization of the Kullback–Leibler divergence, D_{KL}, given by

$$JS(P \mid\mid Q) = \tfrac{1}{2} D_{KL}(P \mid\mid O) + \tfrac{1}{2} D_{KL}(Q \mid\mid O),$$

where $O = (P + Q)/2$. Values closer to zero indicate that the distribution of the partition induced by a specific microclustering prior is closer to the true data partition compared to larger values of JS.

Appendix D. Convergence diagnostics and distortion probabilities

Fig. D.9 displays the traceplots for K, FNR and FDR for two chains of the BBAPG model for the Durham, SDS and SIPP data sets discussed in Section 6. No issues of convergence are observed in either case. However, the mixing of the chains for the SIPP data is slower compared to the Durham and SDS data sets.

For the three applications, we assumed independent Beta prior distributions with mean 0.01 and standard deviation 0.01 for the distortion probabilities of the fields, ψ_ℓ, and obtained posterior estimates of ψ_ℓ that show that the chosen prior specification is not dominating the data. Table D.5 displays the posterior estimates of the distortion probabilities for Durham, SDS and SIPP data sets under the BBAPG prior. We observe (in bold) that in spite of our prior specification favoring small values of ψ_ℓ, one field in each of the applications shows considerably large values of distortion. The fields correspond to race (4.8%), education level (9.2%) and sex (3.8%), respectively for Durham, SDS and SIPP.

Appendix E. Additional simulation results

The following plots (Fig. E.10) display the posterior distributions of the allelic partitions for the five simulated scenarios generated with a distortion probability value of 1% for the fields (see Section 5).

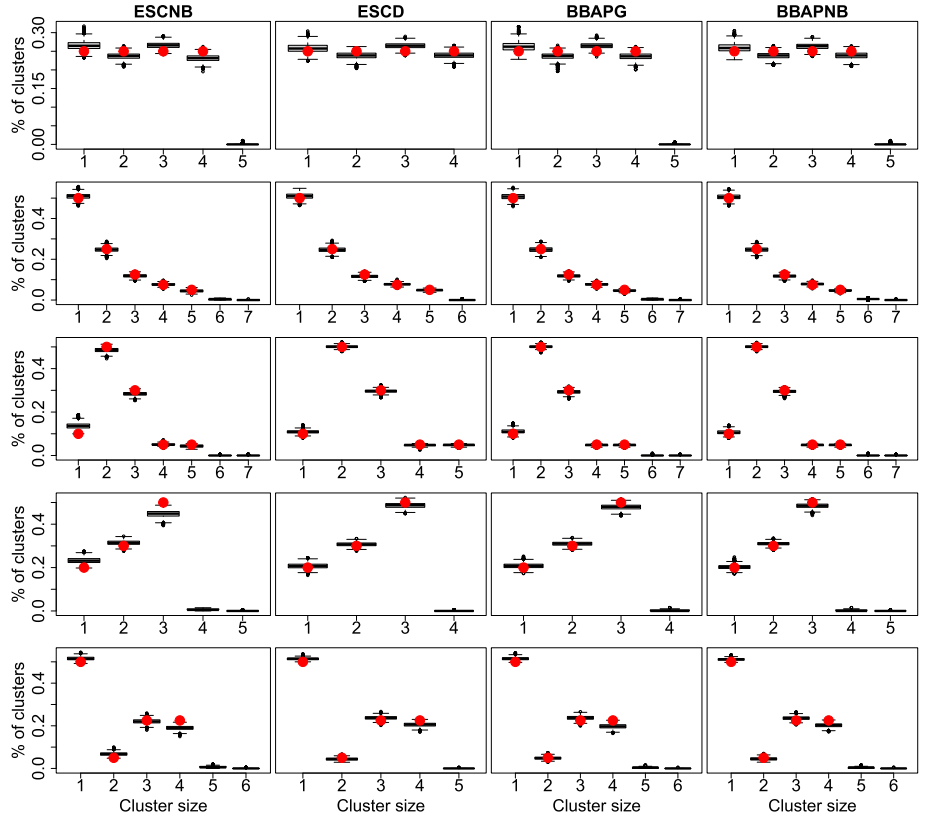

Fig. E.10. Posterior distribution of the allelic partition (boxplots) for ESC and BBAP models, and true data partition (red dots) for five simulated scenarios with distortion probability $\psi_\ell = 0.01$.

References

Aleshin-Guendel, S., Sadinle, M., 2021. Multifile partitioning for record linkage and duplicate detection. arXiv:2110.03839.

Belin, T.R., Rubin, D.B., 1995. A method for calibrating false-match rates in record linkage. J. Am. Stat. Assoc. 90 (430), 694–707.

Betancourt, B., Zanella, G., Steorts, R.C., 2020. Random partition models for microclustering tasks. arXiv preprint. arXiv:2004.02008.

Bhattacharya, I., Getoor, L., 2006. A latent Dirichlet model for unsupervised entity resolution. In: SDM, vol. 5. SIAM, p. 59.

Blei, D.M., Kucukelbir, A., McAuliffe, J.D., 2017. Variational inference: a review for statisticians. J. Am. Stat. Assoc. 112 (518), 859–877. https://doi.org/10.1080/01621459.2017.1285773.

Bloem-Reddy, B., Foster, A., Mathieu, E., Teh, Y.W., 2018. Sampling and inference for beta neutral-to-the-left models of sparse networks. arXiv:e-prints. arXiv:1807.03113.

Broderick, T., Steorts, R.C., 2014. Variational bayes for merging noisy databases. arXiv preprint. arXiv:1410.4792.

Casella, G., Moreno, E., Girón, F.J., et al., 2014. Cluster analysis, model selection, and prior distributions on models. Bayesian Anal. 9 (3), 613–658.

Christen, P., 2012. Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Springer Science & Business Media.

Crane, H., et al., 2016. The ubiquitous Ewens sampling formula. Stat. Sci. 31 (1), 1–19.

Di Benedetto, G., Caron, F., Teh, Y.W., 2017. Non-exchangeable random partition models for microclustering. arXiv preprint. arXiv:1711.07287.

Fellegi, I.P., Sunter, A.B., 1969. A theory for record linkage. J. Am. Stat. Assoc. 64 (328), 1183–1210.

Fienberg, S.E., Makov, U.E., Sanil, A.P., 1997. A bayesian approach to data disclosure: optimal intruder behavior for continuous data. J. Off. Stat. 13, 75–79.

Gutman, R., Afendulis, C.C., Zaslavsky, A.M., 2013. A bayesian procedure for file linking to analyze end-of-life medical costs. J. Am. Stat. Assoc. 108 (501), 34–47.

Hof, M.H., Ravelli, A.C., To, A.H.Z., 2017. A probabilistic record linkage model for survival data. J. Am. Stat. Assoc. 112 (520), 1504–1515.

Jain, S., Neal, R.M., 2004. A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model. J. Comput. Graph. Stat. 13 (1), 158–182.

Kaplan, A., Betancourt, B., Steorts, R.C., 2018. Posterior prototyping: bridging the gap between bayesian record linkage and regression. arXiv preprint. arXiv:1810.01538.

Klami, A., Jitta, A., 2016. Probabilistic size-constrained microclustering. In: Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence, pp. 329–338.

Kolchin, V.F., 1971. A problem of the allocation of particles in cells and cycles of random permutations. Theory Probab. Appl. 16 (1), 74–90.

Larsen, M.D., Rubin, D.B., 2001. Iterative automated record linkage using mixture models. J. Am. Stat. Assoc. 96 (453), 32–41.
Lau, J.W., Green, P.J., 2007. Bayesian model-based clustering procedures. J. Comput. Graph. Stat. 16 (3), 526–558.
Lin, J., 1991. Divergence measures based on the Shannon entropy. IEEE Trans. Inf. Theory 37 (1), 145–151.
McCullagh, P., Yang, J., 2006. Stochastic classification models. In: Proceedings of the International Congress of Mathematicians Madrid. August 22–30, 2006, pp. 669–686.
Miller, J., Betancourt, B., Zaidi, A., Wallach, H., Steorts, R.C., 2015. Microclustering: when the cluster sizes grow sublinearly with the size of the data set. arXiv preprint. arXiv:1512.00792.
Miller, J.W., Harrison, M.T., 2018. Mixture models with a prior on the number of components. J. Am. Stat. Assoc. 113 (521), 340–356.
Müller, P., Rodriguez, A., 2013. Nonparametric Bayesian Inference. Institute of Mathematical Statistics.
Rastelli, R., 2017. GreedyEPL, r package version 1.0. https://cran.r-project.org/web/packages/GreedyEPL/index.html.
Rastelli, R., Friel, N., 2018. Optimal bayesian estimators for latent variable cluster models. Stat. Comput. 28 (6), 1169–1186. https://doi.org/10.1007/s11222-017-9786-y.
Sadinle, M., 2014. Detecting duplicates in a homicide registry using a bayesian partitioning approach. Ann. Appl. Stat. 8 (4), 2404–2434.
Sadinle, M., 2017. Bayesian estimation of bipartite matchings for record linkage. J. Am. Stat. Assoc. 112 (518), 600–612. https://doi.org/10.1080/01621459.2016.1148612.
Sadinle, M., 2018. Bayesian propagation of record linkage uncertainty into population size estimation of human rights violations. Ann. Appl. Stat. 12 (2), 1013–1038. https://doi.org/10.1214/18-AOAS1178.
Sadinle, M., Fienberg, S.E., 2013. A generalized Fellegi–Sunter framework for multiple record linkage with application to homicide record systems. J. Am. Stat. Assoc. 108 (502), 385–397. https://doi.org/10.1080/01621459.2012.757231.
Steorts, R.C., et al., 2015. Entity resolution with empirically motivated priors. Bayesian Anal. 10 (4), 849–875.
Steorts, R.C., Hall, R., Fienberg, S.E., 2016. A bayesian approach to graphical record linkage and deduplication. J. Am. Stat. Assoc. 111 (516), 1660–1672.
Tancredi, A., Liseo, B., 2011. A hierarchical bayesian approach to record linkage and population size problems. Ann. Appl. Stat. 5 (2B), 1553–1585.
U.S. Census Bureau, 2009. Survey of income and program participation (SIPP) 2004 panel. https://doi.org/10.3886/ICPSR04517.v1.
Vatsalan, D., Sehili, Z., Christen, P., Rahm, E., 2017. Big data applications. In: Privacy-Preserving Record Linkage for Big Data: Current Approaches and Research Challenges. Springer, Cham, pp. 851–895.
Wade, S., Ghahramani, Z., 2018. Bayesian cluster analysis: point estimation and credible balls. Bayesian Anal. 13 (2), 559–626.
Winkler, W.E., 2014. Matching and record linkage. Wiley Interdiscip. Rev.: Comput. Stat. 6 (5), 313–325. https://doi.org/10.1002/wics.1317.
Wortman, J.P.H., 2019. Record linkage methods with applications to causal inference and election voting data. Ph.D. thesis. Duke University. https://hdl.handle.net/10161/18657.
Zanella, G., 2019. Informed proposals for local mcmc in discrete spaces. J. Am. Stat. Assoc. 115 (530), 825–865.
Zanella, G., Betancourt, B., Miller, J.W., Wallach, H., Zaidi, A., Steorts, R.C., 2016. Flexible models for microclustering with application to entity resolution. In: Advances in Neural Information Processing Systems, pp. 1417–1425.