Why the Rich Get Richer?
On the Balancedness of Random Partition Models

Changwoo J. Lee [1]   **Huiyan Sang** [1]

Abstract

Random partition models are widely used in Bayesian methods for various clustering tasks, such as mixture models, topic models, and community detection problems. While the number of clusters induced by random partition models has been studied extensively, another important model property regarding the balancedness of partition has been largely neglected. We formulate a framework to define and theoretically study the balancedness of exchangeable random partition models, by analyzing how a model assigns probabilities to partitions with different levels of balancedness. We demonstrate that the "rich-get-richer" characteristic of many existing popular random partition models is an inevitable consequence of two common assumptions: product-form exchangeability and projectivity. We propose a principled way to compare the balancedness of random partition models, which gives a better understanding of what model works better and what doesn't for different applications. We also introduce the "rich-get-poorer" random partition models and illustrate their application to entity resolution tasks.

1. Introduction

Bayesian formulation of clustering problems has gained considerable attention, especially when the number of clusters is unknown. By assigning a prior distribution on the partition space, Bayesian clustering methods give not only flexibility in the number of clusters but also an uncertainty quantification of latent clustering structure. Popular examples include random partition models induced by discrete random probability measures, such as Dirichlet process (DP; Ferguson, 1973), Pitman-Yor process (PYP; Perman et al., 1992;

[1]Department of Statistics, Texas A&M University, Texas, USA. Correspondence to: Changwoo J. Lee <c.lee@stat.tamu.edu>.

Pitman & Yor, 1997), and the mixture of finite mixtures (MFM; Miller & Harrison, 2018), to name a few. These random partition models reflect the prior belief on the clusters' characteristics, such as how the number of clusters K_n^+ grows as the number of datapoints n increases, and how datapoints are distributed among clusters. For example, it is well known that $K_n^+ \asymp \log n$ under the Chinese restaurant process (CRP), the distribution of partition induced by DP.

While the behavior of the number of clusters K_n^+ under various random partition models has been studied extensively (Gnedin & Pitman, 2005; Lijoi et al., 2007; Favaro et al., 2009; De Blasi et al., 2015; Camerlenghi et al., 2018; Lijoi et al., 2020; Frühwirth-Schnatter et al., 2021), the *balancedness* of random partition (i.e. probabilities of partitions with different levels of balancedness) has been largely uninvestigated. Nevertheless, there has been awareness of the crucial effect of the balancedness of random partitions on cluster inference. The most prominent example is the CRP prior that is known to yield the posterior estimate with few big and many small extra clusters (Petrone & Raftery, 1997), intuitively due to the prediction rule of CRP that assumes a new datapoint falls into an existing cluster with probability proportional to its size. This "rich-get-richer" characteristic of preferring unbalanced partitions is shared across many other random partition models, such as those induced by PYP, MFM, and Dirichlet-multinomial allocation (Green & Richardson, 2001), yet there is a less theoretical understanding of when and why this property arises.

There are many important questions surrounding the balancedness of random partition models that are not well studied in the literature. First, how to *define* a rigorous theoretical framework that is applicable to systematically studying the balancedness properties for a broad class of random partition models? The answer to this question would be crucial to understand the root causes of the "rich-get-richer" property: the product-form exchangeability and projectivity assumptions which we shall demonstrate later. Also, it provides deeper insight into the suitable application scenario of each model. The success of latent Dirichlet allocation (LDA; Blei et al., 2003) in topic modeling is partially attributed to the fact that per-document topic proportions are inherently unbalanced, and hence unbalanced random partition models are naturally suitable for topic modeling.

Second, how to *compare* and *quantify* the balancedness of different random partition models? The answer to this question helps to understand the cost and benefits of choosing one model over the other. For example, many practitioners choose DP-based models since it does not require the number of components to be fixed in advance, and indeed many methods incorporating DP emphasize this as a key feature (Teh et al., 2006; Kemp et al., 2006). However, this benefit comes with a cost: compared to the Dirichlet-multinomial allocation model, the random partition induced by DP is more unbalanced, which explains why the DP mixture model typically yields too many tiny clusters compared to the finite mixture or MFM (Green & Richardson, 2001; Miller & Harrison, 2018). Also, community detection results based on the Bayesian stochastic blockmodel (Nowicki & Snijders, 2001; Geng et al., 2019) are reported to be sensitive to the choice of the latent random partition model (Legramanti et al., 2022), and this may be partly explained by the balancedness of random partition prior; see Section 3.3 for the detailed discussion.

Finally, beyond the sea of models that implicitly favor unbalanced partitions, are there flexible models that favor balancedness or have no preference on the balancedness? There have been some efforts to follow a sequential construction of random partition and redesign prediction rules to change the "rich-get-richer" property (Jensen & Liu, 2008; Wallach et al., 2010; Lu et al., 2018; Poux-Médard et al., 2021), but these sequential construction methods (i.e. species sampling models) often no longer guarantee exchangeability of the resulting random partition model (Lee et al., 2013), a crucial assumption that the distribution of random partition is invariant of the permutation of indices. Indeed, there are many application scenarios for which such models are more suitable. In this paper, we illustrate with an entity resolution (ER) task which aims to find clusters of records that refer to the same individual in a noisy database where we do not expect a small number of clusters dominates the whole.

Scope and Contributions. First, we formulate the balancedness of exchangeable random partitions and study the related properties, which is crucial to check or build the model having the desired balancedness property. Especially, we focus on Gibbs partition models (Gnedin & Pitman, 2005), a very general class of exchangeable random partition models. Second, we provide the complete family of balance-neutral random partitions that is both exchangeable and projective, which can serve as a noninformative prior in terms of balancedness. Third, we define *B-sequence* to compare the strength of balancedness between Gibbs partitions and offer its intuitive explanation. Finally, we provide an in-depth example of a balance-seeking random partition model and demonstrate its application to the ER task.

2. Backgrounds

2.1. Notations

Let $[n] = \{1, \ldots, n\}$ denote the index set of datapoints. A partition of $[n]$ is a collection of unordered nonempty subsets (clusters) $\Pi_n = \{S_1, \ldots, S_k\}$ with $\cup_j S_j = [n]$ and $S_j \cap S_\ell = \emptyset$ for $j \neq \ell$. Let $\mathcal{P}_{[n]}^k$ denote the set of partitions of $[n]$ into k blocks, and $\mathcal{P}_{[n]} = \cup_{k=1}^n \mathcal{P}_{[n]}^k$ be the set of all partitions of $[n]$. We use notations $n_j = |S_j|$ and $\boldsymbol{n} = (n_1, \ldots, n_k)$ to denote the cluster sizes and $k = |\Pi_n|$ to denote the number of clusters in partition Π_n. Let $z_i \in \{1, \ldots, k\}$ be a cluster membership of i-th datapoint and $\mathbf{z} = (z_1, \ldots, z_n)$ be a cluster membership vector. Further, let $\mathcal{I}_n^k$ be a collection of integer partition of n into k parts, a nonincreasing $k-$tuple of positive integers $\boldsymbol{n} = (n_1, \ldots, n_k)$ whose sum is n. Similarly, define $\mathcal{I}_n = \cup_{k=1}^n \mathcal{I}_n^k$. For a single $\boldsymbol{n} \in \mathcal{I}_n^k$, there are $n!/(\prod_{i=1}^n (i!)^{m_i} m_i!)$ number of partitions in $\mathcal{P}_{[n]}^k$ whose nonincreasing cluster sizes correspond to $\boldsymbol{n}$, where $m_i = \sum_{j=1}^k 1(n_j = i)$ is the number of clusters of size i.

2.2. Comparison of Balancedness Between Partitions

Next we consider how to compare the balancedness of two *fixed* partitions with cluster sizes $\boldsymbol{n} = (n_1, \ldots, n_k)$ and $\boldsymbol{n}' = (n_1', \ldots, n_k')$. We restrict our attention to the case when both partitions have k clusters. A diversity index (Hill, 1973) is a measure to quantify the evenness of species distribution widely used in the ecology literature. To compare the balancedness of two partitions using the diversity index, we treat the proportion of cluster sizes $\{n_j/n\}_{j=1}^k$ as species distribution. Popular examples are the Shannon index (with base e) and the Gini-Simpson index:

$$H(\boldsymbol{n}) := -\sum_{j=1}^k (n_j/n) \log(n_j/n) \quad \text{(Shannon)} \quad (1)$$

$$G(\boldsymbol{n}) := 1 - \sum_{j=1}^k (n_j/n)^2 \quad \text{(Gini-Simpson)} \quad (2)$$

Other examples include the Rényi index and the Hill numbers (Daly et al., 2018).[1] A higher diversity index indicates a partition is more balanced (has clusters with similar sizes), and the index is maximized when all n_j/n are equal.

However, different diversity indices do not always give a consistent ordering (Hurlbert, 1971), and they may not provide meaningful comparisons for some pairs of partitions. For instance, consider partitions of $[10]$ with cluster sizes $\boldsymbol{n} = (6, 2, 2)$ and $\boldsymbol{n}' = (5, 4, 1)$. It is not clear which one is more balanced than the other because both can be obtained by moving one data from a larger cluster of the partition $(6, 3, 1)$ to a smaller cluster, and diversity indices are inconsistent: $H(\boldsymbol{n}) \approx 0.95 > 0.94 \approx H(\boldsymbol{n}')$ but $G(\boldsymbol{n}) \approx 0.56 < 0.58 \approx G(\boldsymbol{n}')$. Instead, without refer-

---

[1] We refrain from using the term *entropy* to avoid confusion since we don't consider $\boldsymbol{n}$ as a random variable in this context.

ring to indices, Patil & Taillie (1982) proposed intrinsic diversity ordering, a partial order relation to compare diversity between two species distribution. When it comes to the discrete space $\mathcal{I}_n^k$, intrinsic diversity ordering now corresponds to the reverse dominance order.

Definition 2.1 (Reverse dominance order). Let $\boldsymbol{n} = (n_1, \ldots, n_k)$ and $\boldsymbol{n}' = (n_1', \ldots, n_k')$ be two integer partitions of n into k parts. By Brylawski (1973), the following statements are equivalent definitions of partial order $\prec$ on $\mathcal{I}_n^k$ where $\boldsymbol{n} \prec \boldsymbol{n}'$ represents $\boldsymbol{n}'$ is more balanced than $\boldsymbol{n}$:

(a) (reverse dominance ordering) Write $\boldsymbol{n} \prec \boldsymbol{n}'$ if $\sum_{j=1}^{J} n_j \geq \sum_{j=1}^{J} n_j'$ for $J = 1, \ldots, k$ and $\boldsymbol{n} \neq \boldsymbol{n}'$.

(b) (one-step downshift) Write $\boldsymbol{n} \prec \boldsymbol{n}'$ if $\boldsymbol{n}$ leads to $\boldsymbol{n}'$ by a finite sequence of one-step downshifts: $(n_1, \ldots, n_k) \mapsto (n_1, \ldots, n_u - 1, \ldots, n_v + 1, \ldots, n_k)$ provided that $n_u - 1 \geq n_v + 1$, where $1 \leq u < v \leq k$.

(c) (covering relation) $\prec$ is induced by the following covering relation: $\boldsymbol{n}'$ covers $\boldsymbol{n}$ (i.e. no $\boldsymbol{n}''$ exist s.t. $\boldsymbol{n} \prec \boldsymbol{n}'' \prec \boldsymbol{n}'$) if $n_u' = n_u - 1, n_v' = n_v + 1$ for some $u < v$, $n_j = n_j'$ for all $j \in \{1, \ldots, k\} \backslash \{u, v\}$, and either $(*)$ $v = u + 1$ or $(**)$ $n_u' = n_v'$, or both.

See Figure 1 for an example of $\prec$. Although Definition 2.1 can be extended to the set $\mathcal{I}_n = \cup_{k=1}^{n} \mathcal{I}_n^k$, we shall only focus on the ordering between integer partitions with the same k to study the balancedness of random partition models.

2.3. Gibbs Partitions

Random partition of $[n]$ is a discrete probability distribution on the space of partitions $\mathcal{P}_{[n]}$. There are two common assumptions on random partition models: *finite exchangeability* and *projectivity*. A random partition Π_n on $[n]$ is called *finitely exchangeable* if probabilities remain unchanged under any permutation of the indices of datapoints, which is a natural assumption unless one wants to take account of data dependence structures. Next, a sequence of random partitions $(\Pi_n)_{n=1}^{\infty}$ satisfies *projectivity* if the restriction of Π_n to $[m]$ is Π_m for any $m < n$, which is also called the addition rule, self-consistency, or marginal invariance property. We call (Π_n) *infinitely exchangeable* if Π_n is finitely exchangeable for each n and satisfies projectivity.

In this paper, we always assume finite exchangeability but not necessarily projectivity, as advocated in Welling (2006). Under the finite exchangeability, we can express the probability mass function of Π_n using the *exchangeable partition probability function* (EPPF) (Pitman, 1995).

Definition 2.2 (EPPF). A symmetric function $p^{(n)}$ is an EPPF of an exchangeable random partition Π_n of $[n]$ if

$$\mathbb{P}(\Pi_n = \{S_1, \ldots, S_k\}) = p^{(n)}(|S_1|, \ldots, |S_k|)$$
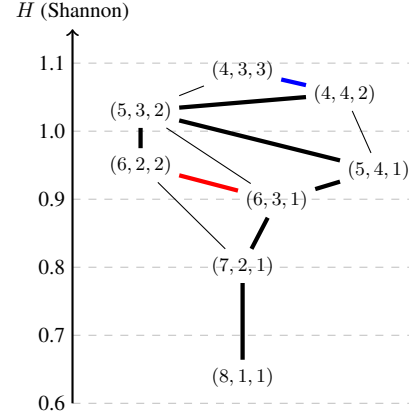

Figure 1. Diagram representing the partial order $\prec$ on $\mathcal{I}_{10}^3$. For example, $(8, 1, 1) \prec (7, 2, 1)$ which indicates $(7, 2, 1)$ is more balanced than $(8, 1, 1)$. The vertical position corresponds to the Shannon index H. Edges between two partitions represent the upper can be reached from the lower by a one-step downshift. Thick edges represent the covering relation in Definition 2.1(c). Especially, colored thick edges corresponds to $(**)$-covering with $n_u' = n_v' = 2$ (red) and $n_u' = n_v' = 3$ (blue); see Section 3.3 and Figure 4 for the detailed discussion on colored edges.

Since EPPF is symmetric of its arguments, the values of EPPF on integer partition $\boldsymbol{n} \in \mathcal{I}_n$ completely determine the probability mass function of random partition. Next, we introduce the *Gibbs (or Gibbs-type) partition* (Pitman, 2006), a wide class of exchangeable random partition model.

Definition 2.3 (Gibbs partition). Given two nonnegative sequences $\boldsymbol{V} = (V_{n,k})_{k \leq n}$ and $\boldsymbol{W} = (W_s)_{s \geq 1}$ with $V_{1,1} = W_1 = 1$, we say $\Pi_n \sim \mathsf{Gibbs}_{[n]}(\boldsymbol{V}, \boldsymbol{W})$ if its EPPF is

$$p^{(n)}(n_1, \ldots, n_k) = V_{n,k} \prod_{j=1}^{k} W_{n_j} \qquad (3)$$

That is, the probability mass function has a product form that ensures finite exchangeability as well as mathematical tractability. While some authors define Gibbs partition with projectivity assumption (see Proposition 2.4), our definition following Pitman (2006) does not assume projectivity. The family of Gibbs partition is broad (Lomelí et al., 2017, Fig.1), encompassing not only random partitions induced by DP, PYP, normalized generalized gamma process (Lijoi et al., 2007), and σ-stable Poisson-Kingman partitions (Pitman, 2003), but also those induced by parametric models such as Dirichlet-multinomial and MFM. To ensure Gibbs partition also satisfies projectivity (i.e., infinitely exchangeable), Gnedin & Pitman (2005); Lijoi et al. (2007) provided the following necessary and sufficient conditions.

Proposition 2.4. *A $\mathsf{Gibbs}_{[n]}(\boldsymbol{V}, \boldsymbol{W})$ partition for $n \in \mathbb{N}$, excluding partition of singletons and one-block partition, satisfies projectivity if and only if for some $\sigma \in [-\infty, 1)$, the following two conditions are both satisfied:*

(i) $W_s = \Gamma(s - \sigma)/\Gamma(1 - \sigma), \quad s = 1, 2, \dots$

(ii) $V_{n,k} = (n - \sigma k)V_{n+1,k} + V_{n+1,k+1}, \quad 1 \le k \le n$

When $\sigma = -\infty$, replace $W_s \equiv 1$ and $(n - \sigma k)$ by k in (ii).

The most important subclass of infinitely exchangeable Gibbs partitions is the *Ewens-Pitman two-parameter family* (Ewens, 1972; Pitman, 2006) with parameters (σ, θ) where

$$V_{n,k} = \frac{\prod_{i=1}^{k-1}(\theta + i\sigma)}{(\theta + 1)\cdots(\theta + n - 1)}, \quad W_s = \frac{\Gamma(s - \sigma)}{\Gamma(1 - \sigma)} \quad (4)$$

with exception of $V_{n,1} = 1/((\theta + 1)\cdots(\theta + n - 1))$ for $n \ge 2$ and $V_{1,1} = 1$. The range of (σ, θ) is either (a) $\sigma \in [0, 1)$, $\theta > -\sigma$ which corresponds to the random partition induced by DP and PYP, (b) $\sigma < 0$, $\theta = K|\sigma|$ for $K \in \mathbb{N}$ which corresponds to Dirichlet-multinomial allocation model, or (c) $\sigma = -\infty$, replacing $W_s \equiv 1$ and $V_{n,k} = K(K - 1)\cdots(K - k + 1)/K^n$ for some $K \in \mathbb{N}$ leads to the coupon-collector's partition. See Pitman (2006); De Blasi et al. (2015) for the detailed treatments.

3. Balancedness of Gibbs Partition and Related Properties

3.1. Why the Rich Get Richer?

First, we define the balancedness of a finitely exchangeable random partition using the ordering $\prec$ in Definition 2.1.

Definition 3.1. Let $p^{(n)}$ be an EPPF of a finitely exchangeable random partition Π_n on $[n]$. Then call Π_n and $p^{(n)}$

- *balance-averse if $\boldsymbol{n} \prec \boldsymbol{n}' \implies p^{(n)}(\boldsymbol{n}) \ge p^{(n)}(\boldsymbol{n}')$*

- *balance-seeking if $\boldsymbol{n} \prec \boldsymbol{n}' \implies p^{(n)}(\boldsymbol{n}) \le p^{(n)}(\boldsymbol{n}')$*

for any fixed $k \le n$ and $\boldsymbol{n}, \boldsymbol{n}' \in \mathcal{I}_n^k$. Also, call Π_n *balance-neutral* if Π_n is both balance-averse and balance-seeking.

The definition has a straightforward interpretation: Π_n is balance-averse (seeking) when it assigns a lower (higher) probability to a more balanced partition in terms of $\prec$.

Now we present the necessary and sufficient conditions which characterize the balancedness of Gibbs partitions, where the proof is deferred to Appendix A.

Theorem 3.2. *Let $p^{(n)}$ be an EPPF of $\mathsf{Gibbs}_{[n]}(\boldsymbol{V}, \boldsymbol{W})$. Then for any $n = 1, 2, \dots, p^{(n)}$ is*

- *balance-averse if and only if $\boldsymbol{W}$ is log-convex,*

- *balance-seeking if and only if $\boldsymbol{W}$ is log-concave,*

where the sequence $\boldsymbol{W} = (W_s)_{s=1}^{\infty}$ is called log-convex if $W_s^2 \le W_{s-1}W_{s+1}$ for $s \ge 2$, and is called log-concave if $W_s^2 \ge W_{s-1}W_{s+1}$ for $s \ge 2$ and has no internal zeros.

Thus, to check the balancedness of a Gibbs partition, it is not necessary to verify the conditions in Definition 3.1, and one only needs to analyze the log-convexity of $\boldsymbol{W}$.

Gibbs partition admits a simple form of reallocation rule (prediction rule if projectivity holds) thanks to the product-form EPPF. The following corollary draws an explicit connection between the "rich-get-richer" metaphor and the balancedness of Gibbs partition based on Definition 3.1; see Appendix A for the details.

Corollary 3.3. *Let $\Pi_{n+1} \sim \mathsf{Gibbs}_{[n+1]}(\boldsymbol{V}, \boldsymbol{W})$. Given the cluster memberships of the first n datapoints $\mathbf{z}_{1:n}$ with k clusters, the reallocation rule for the next datapoint is*

$$\mathbb{P}(z_{n+1} = j|\mathbf{z}_{1:n}) \propto \begin{cases} f(n_j) & \text{if } j = 1, \dots, k \\ g(n, k) & \text{if } j = k + 1 \end{cases} \quad (5)$$

Then f is an increasing (decreasing) function over $\mathbb{N}$ if and only if Π_{n+1} is balance-averse (seeking) for any $n \in \mathbb{N}$.

In addition to the finite exchangeability, many existing random partition model also assumes projectivity (aforementioned examples of Gibbs partitions in Section 2.3 are all infinitely exchangeable). By Proposition 2.4, the following corollary states that infinitely exchangeable Gibbs partitions are always balance-averse.

Corollary 3.4. *Let $p^{(n)}$ be an EPPF of infinitely exchangeable Gibbs partition. Then $p^{(n)}$ is always balance-averse; i.e., for two integer partitions $\boldsymbol{n}, \boldsymbol{n}' \in \mathcal{I}_n^k$ with $k \le n$,*

$$\boldsymbol{n} \prec \boldsymbol{n}' \implies p^{(n)}(\boldsymbol{n}) \ge p^{(n)}(\boldsymbol{n}')$$

with equality holds only if $\sigma = -\infty$.

To see this, $W_s = \Gamma(s-\sigma)/\Gamma(1-\sigma), s = 1, 2, \dots$ is strictly log-convex for any $\sigma \in (-\infty, 1)$ and $W_s \equiv 1$ for $\sigma = -\infty$. Figure 2 depicts the behavior of $\log p^{(n)}$ for the three different Ewens-Pitman two-parameter models in (4). The x-axis is the Shannon index H, which satisfies $\boldsymbol{n} \prec \boldsymbol{n}' \implies H(\boldsymbol{n}) < H(\boldsymbol{n}')$. All scatterplots show a decreasing pattern for each k as H increases (i.e., partitions become more balanced), since two-parameter models are all infinitely exchangeable and thus balance-averse.

Corollary 3.4 answers why the "rich-get-richer" property is shared across many existing random partition models. At the same time, it also implies that one should sacrifice projectivity to get a more flexible class of exchangeable random partition models with the balance-seeking property. The cost of sacrificing projectivity depends on the application. When the cluster membership of the future datapoint is a primary interest, lack of projectivity leads to an undesirable consequence that the implied joint distribution on the current dataset changes as we get more data. However, it has little impact if the primary goal is to make an inference on the given dataset such as entity resolution, and opens up a wide range of possible models with different balancedness.
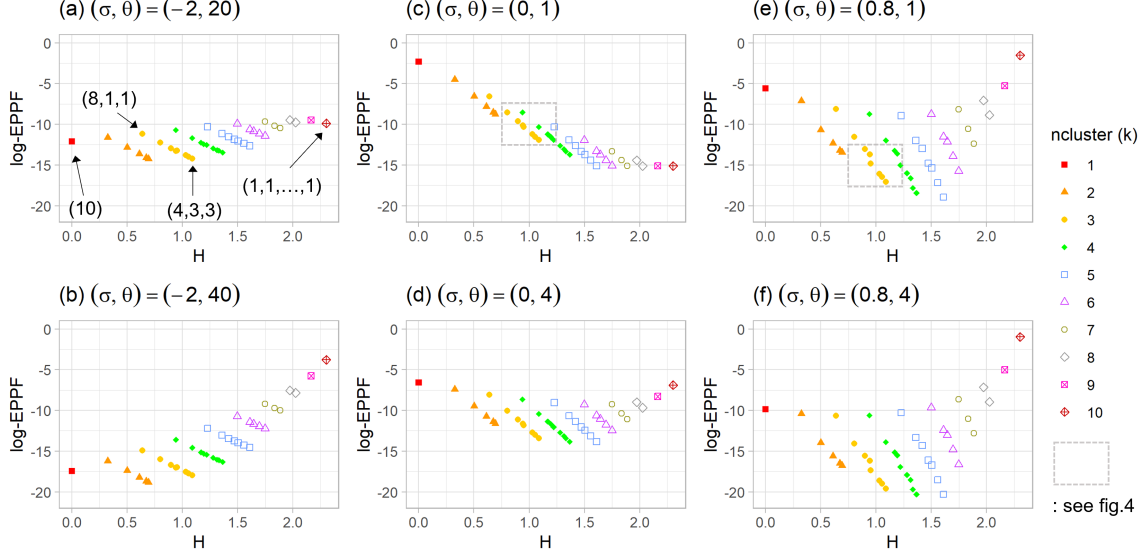
Figure 2. log-EPPF plots against Shannon index H of Ewens-Pitman two-parameter models when $n = 10$, colored by the number of clusters. (a,b) Random partitions induced by Dirichlet-multinomial with symmetric Dirichlet parameter $|\sigma| = 2$ and number of components $K = 10, 20$ respectively. (c,d) Chinese restaurant process with concentration parameters 1 and 4 respectively. (e,f) Random partitions induced by Pitman-Yor process with common discount parameter $\sigma = 0.8$ and concentration parameters 1 and 4 respectively.

Theorem 3.2 can be extended to the Gibbs partitions with additional hierarchical structure on V or W.

Theorem 3.5. *We say Π_n is a mixture of Gibbs partition if it is finitely exchangeable and its EPPF $p^{(n)}$ has a form*

$$p^{(n)}(n_1, \ldots, n_k) = \int V_{n,k}(\vartheta) \prod_{j=1}^{k} W_{n_j}(\vartheta) \nu(d\vartheta) \quad (6)$$

where ϑ is a mixing parameter which may be either discrete or continuous. Then for any $n = 1, 2, \ldots, p^{(n)}$ is

- *balance-averse if $(W_s(\vartheta))$ is log-convex for each ϑ,*

- *balance-seeking if $(W_s(\vartheta))$ is log-concave for each ϑ.*

It provides sufficient conditions for (6) being balance-averse or balance-seeking; see Appendix A for the proof. Examples include random partitions induced by DP with prior on the concentration parameter θ, PYP with prior on the discount parameter σ (Lijoi et al., 2007), and generalized MFM recently proposed by Frühwirth-Schnatter et al. (2021).

3.2. Balance-neutral Random Partition Model

There have been some efforts to develop random partitions that do not exhibit the "rich-get-richer" property (Jensen & Liu, 2008; Wallach et al., 2010) or control its rate (Lu et al., 2018; Poux-Médard et al., 2021) but at the cost of sacrificing exchangeability. Here we show that there does exist an exchangeable, projective random partition model that is balance-neutral, and Theorem 3.6 provides its only possible form of EPPF; see Appendix A for the proof.

Theorem 3.6. *Let $p^{(n)}$ be an EPPF of balance-neutral infinitely exchangeable random partition. Then there exist some mixing distribution q on the number of components $K \in \mathbb{N}$ so that*

$$p^{(n)}(n_1, \ldots, n_k) = \sum_{K=k}^{\infty} q(K) \frac{K(K-1)\cdots(K-k+1)}{K^n} \quad (7)$$

Figure 3 (left) gives an example when q is Poisson(3) shifted by 1, showing the flat EPPF pattern for each k. It can be also viewed as the limiting case of the random partition induced by MFM when $|\sigma| \to \infty$. The prediction rule is

$$\mathbb{P}(z_{n+1} = j | \mathbf{z}_{1:n}) \propto \begin{cases} 1 & \text{if } j = 1, \ldots, k \\ \frac{V_{n+1,k+1}(q)}{V_{n+1,k}(q)} & \text{if } j = k+1 \end{cases} \quad (8)$$

where $V_{n,k}(q)$ is the RHS of (7) and these can be precomputed in advance; see Miller & Harrison (2018) for details. This balance-neutral random partition model can serve as a noninformative prior choice in terms of balancedness, as it assigns uniform probabilities for each $\Pi_n \in \mathcal{P}_{[n]}^k$ while q controls the creation of new clusters.

3.3. Comparing the Strength of Balancedness with B-sequence

Theorem 3.2 indicates that the log-convexity/concavity of sequence $(W_s)_{s=1}^{\infty}$ determines the balancedness of Gibbs partition. Then a natural question arises: how to compare the balancedness between two random partitions? Which Gibbs partition is more balanced over the other?
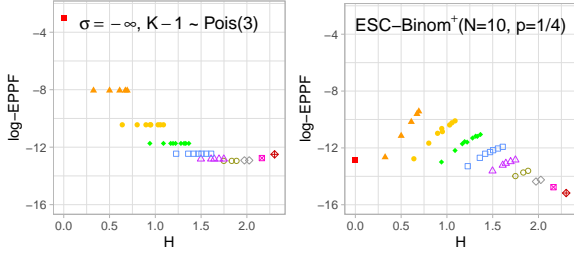
Figure 3. Two examples of log-EPPF plots, see Figure 2 for the legend. (Left) Balance-neutral, when $q = 1 + \text{Poisson}(3)$. (Right) Balance-seeking, when ESC model with $\boldsymbol{\mu} = \text{Binom}^+(10, 1/4)$.
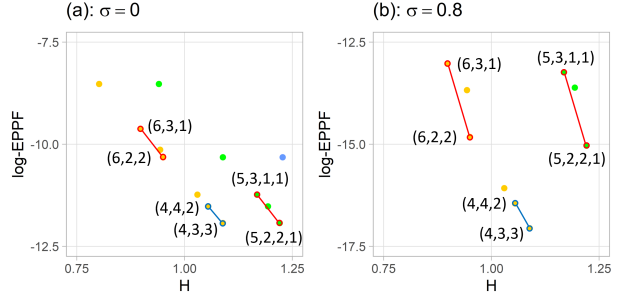

Figure 4. Graphical interpretation of B-sequence via (10), where $n = 10$ and $\boldsymbol{W}$ in (4) is parameterized by σ. Subplots (a), (b) are zoomed portion of Figure 2(c), (e) respectively. Red segments correspond to $(\ast\ast)$-covering relation with $s = 2$, and blue with $s = 3$ (see also Figure 1). Approximately, the slope of red segments are $nB_2(\sigma = 0)$ (left) and $nB_2(\sigma = 0.8)$ (right), and blue segments are $nB_3(\sigma = 0)$ (left) and $nB_3(\sigma = 0.8)$ (right). As σ increases, slopes become steeper and the EPPF becomes more unbalanced.

Assuming (W_s) has no internal zeros, we can use the concept of *relative log-concavity ordering* $\leq_{\text{lc}}$ (Whitt, 1985; Yu, 2010) which is a preorder satisfying reflexivity and transitivity. $\boldsymbol{W}$ is called log-concave relative to $\boldsymbol{W}'$, written as $\boldsymbol{W} \leq_{\text{lc}} \boldsymbol{W}'$, if $\text{supp}(\boldsymbol{W}) \subseteq \text{supp}(\boldsymbol{W}')$ and $(\log(W_s/W_s'))$ is concave in $\text{supp}(\boldsymbol{W})$, the support of $\boldsymbol{W}$. While the comparison of $\boldsymbol{W}$ in ordering $\leq_{\text{lc}}$ itself can be used to compare the balancedness, there is also a need to define an intuitive measure that quantifies the balancedness so that we propose *B-sequence* as follows.

Definition 3.7 (B-sequence). $(B_s(\boldsymbol{W}))_{s \geq 2}$ is a B-sequence of $\text{Gibbs}_{[n]}(\boldsymbol{V}, \boldsymbol{W})$ for $n \in \mathbb{N}$, a sequence of extended real numbers which only depends on $\boldsymbol{W}$, defined as

$$B_s(\boldsymbol{W}) = -s(\log W_{s+1} - 2\log W_s + \log W_{s-1}) \quad (9)$$

with the provision that $B_s(\boldsymbol{W}) = +\infty$ if $W_{s+1} = 0$.

This definition is closely related to the slope of log-EPPF against the Shannon index H. Recall the covering relation from Definition 2.1(c) corresponding to the case $(\ast\ast)$. If $\boldsymbol{n}, \boldsymbol{n}' \in \mathcal{I}_n^k$ and $\boldsymbol{n}'$ is a $(\ast\ast)$-cover of $\boldsymbol{n}$, then there exists some $s \geq 2$ such that $s = n_u' = n_u - 1 = n_v' = n_v + 1$ with $u < v$. We prove that $nB_s(\boldsymbol{W})$ is an approximation of the slope along this covering relation as shown in Figure 4:

$$\frac{\log p^{(n)}(\boldsymbol{n}') - \log p^{(n)}(\boldsymbol{n})}{H(\boldsymbol{n}') - H(\boldsymbol{n})} \approx nB_s(\boldsymbol{W}), \quad (10)$$

where we defer the detailed derivation of approximation to Appendix B. That is, B-sequence represents the strength of balancedness by measuring the derivative of log-EPPF against Shannon index H, the part which is invariant of n.

Here $B_s(\boldsymbol{W}) \equiv 0$ is equivalent to the Gibbs partition being balance-neutral which serves as the origin, and $B_s(\boldsymbol{W}) \geq 0$ (≤ 0) for all s is equivalent to the Gibbs partition being balance-seeking (balance-averse) respectively. The part of (9) inside the parenthesis is the discrete second derivative of the sequence $(\log W_s)$ and naturally relates to log-convexity. Indeed, we have $\boldsymbol{W} \leq_{\text{lc}} \boldsymbol{W}'$ if and only if $B_s(\boldsymbol{W}) \geq B_s(\boldsymbol{W}')$ for all $s \geq 2$. See Appendix B for the proof.

Examples. Consider Ewens-Pitman two-parameter family (4) where $\boldsymbol{W}$ is parameterized by σ. Then, for $\sigma < \sigma' < 1$, the corresponding B-sequences $(B_s(\sigma))$ and $(B_s(\sigma'))$ satisfy $B_s(\sigma') < B_s(\sigma) < 0$ for all $s \geq 2$. It implies that Dirichlet-multinomial or MFM ($\sigma < 0$) based random partition model is more balanced than CRP ($\sigma = 0$) induced by DP, and CRP is more balanced than the one induced by PYP ($\sigma \in (0, 1)$). The slope interpretation in (10) well matches with the pattern shown in Figure 2; the negative slope becomes steeper as σ increases. This result is also consistent with the comparisons in Green & Richardson (2001); Miller & Harrison (2018) on the mixture modeling tasks, where CRP typically yields small extraneous clusters compared to Dirichlet-multinomial or MFM.

The effect of balancedness can be also found in topic model applications. Although the difference between LDA (Blei et al., 2003) and hierarchical DP (HDP; Teh et al., 2006) is often emphasized in terms of the limited/unlimited number of topics, it is almost always neglected that the prior on the per-document topic distribution of HDP is more unbalanced than those of LDA. As an example of the effect of balancedness, Shi et al. (2019) studied how stopwords (e.g. "a", "the", …) affect the performance of topic modeling algorithms. As the proportion of stopwords increases, the performance of HDP deteriorates significantly compared to LDA since stopwords make the per-document topic proportion more balanced. It also suggests the importance of the hyperparameter choice which affects the balancedness of the model (Asuncion et al., 2009).

For the community detection problems, Legramanti et al. (2022) utilized infinitely exchangeable Gibbs partitions as a prior on the partition of nodes. In their simulation studies with true clusters being fairly balanced, a random partition prior with $\sigma = -1$ (Gnedin, 2010) performed the best com-

pared to other more unbalanced priors with $\sigma > -1$. Also see Schmidt & Morup (2013) for a real network data application of the CRP prior (Kemp et al., 2006), which yields many tiny community estimates. To sum up, the balancedness of random partition models can help understand these findings and choose random partition models with desired balancedness to tackle these problems.

Remark. The comparison of balancedness between some special cases of random partitions has been discussed before (Miller & Harrison, 2018; Frühwirth-Schnatter et al., 2021) but from a different perspective. They considered the aggregated probabilities $\sum_{\Pi_n \in \mathcal{A}} \mathbb{P}(\Pi_n = \{S_1, \ldots, S_k\})$ with $\mathcal{A} = \{\Pi_n \in \mathcal{P}_{[n]}^k : \Pi_n$ has cluster sizes $(n_1, \ldots, n_k)\}$ (or equivalently, the distribution of *labeled cluster sizes*), instead of the probability mass of each partition $\mathbb{P}(\Pi_n = \{S_1, \ldots, S_k\}) = p^{(n)}(|S_1|, \ldots, |S_k|)$, the EPPF. Comparison of aggregated probabilities has a somewhat unnatural interpretation when it comes to explaining the prior effect of random partition models on the posterior estimate. The probability mass of individual partition (EPPF) is what contributes to the posterior, not the aggregated probability. While the distribution of functionals over the labeled cluster sizes, such as mean and variance of the Shannon index (Greve et al., 2022), could be indicative of comparison regarding balancedness, comparison of EPPF provides a more direct interpretation of the prior effect on the posterior estimate which is often a point of interest.

4. Balance-seeking Random Partition Models

4.1. Entity Resolution and ESC Model

Entity resolution (ER) is the process of matching records that describe the same individual when there are no unique identifiers available (such as due to privacy reasons). Assuming each record corresponds to one individual, ER can be understood as a clustering task where each cluster corresponds to each individual; see Binette & Steorts (2022) for the most recent review of ER literature.

Databases have many noisy records, where reasons include address changes, name changes, measurement/transcription errors among many others. To deal with such distortions with uncertainty quantification, Steorts et al. (2016) developed a full Bayesian hierarchical model, and we briefly describe the model here. Assume n records $(\boldsymbol{x}_i)_{i=1}^n$ contain L categorical fields with entries $(x_{i\ell})$. Let D_ℓ be the number of categories in the ℓth field and let the corresponding probabilities $\boldsymbol{\theta}_\ell = (\theta_{\ell 1}, \ldots, \theta_{\ell D_\ell})$. Also let the cluster indicators be $(z_i)_{i=1}^n$, where the total number of individuals K^+ is unknown. Assuming the latent entities $(y_{k\ell})_{k=1}^{K^+}$ are drawn from Categorical$(\boldsymbol{\theta}_\ell)$, the generative process of entries is $x_{i\ell} = y_{z_i\ell}$ with probability $1 - \beta_\ell$ (non-distorted), and $x_{i\ell} \sim$ Categorical$(\boldsymbol{\theta}_\ell)$ with probability β_ℓ (distorted).

Table 1. An example database with categorical entities $(x_{i\ell})$. Here z_i are assumed to be unknown. Reds are 'distorted' entities, modeled with distortion probabilities (β_ℓ) and latent entities $(y_{k\ell})$.

i	SEX	SURNAME	STATE		z_i
1	F	SMITH	CA	$\cdots$	1
2	F	JOHNSON	CA	$\cdots$	1
3	F	JOHNSON	CA	$\cdots$	1
4	M	WILLIAMS	TX	$\cdots$	2
5	M	WILLIAMS	TX	$\cdots$	2
6	M	WILLIAMS	FL	$\cdots$	2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

See Table 1 for an example. With this Bayesian hierarchical model, by assigning a prior on the partition represented by cluster indicators (z_i), one can obtain a posterior estimate of z_i with uncertainty quantification of record matches.

However, many traditional random partition models assume infinite exchangeability and thus are not suitable for ER tasks, since it leads to the behavior that the number of datapoints in each cluster grows linearly with n, by Kingman's paintbox representation theorem (Kingman, 1978). To overcome this problem, Miller et al. (2015); Zanella et al. (2016) formalized the *microclustering property*: $M_n/n \xrightarrow{p} 0$ where M_n is the maximum size of clusters so that cluster sizes grow sublinearly with n. Recently Betancourt et al. (2020) proposed the exchangeable sequences of clusters (ESC) model, which provides a very general framework of random partition models possessing the microclustering property under mild assumptions. It sacrifices projectivity, and the key idea is to consider a discrete probability measure $\boldsymbol{\mu} = (\mu_s)_{s=1}^\infty$ with $\mu_1 > 0$ that represents the distribution of the cluster sizes $n_j = |S_j|$, $j = 1, \ldots, k$. Conditional on $\boldsymbol{\mu}$, ESC is a Gibbs partition model where its EPPF is:

$$p^{(n)}(n_1, \ldots, n_k|\boldsymbol{\mu}) = \frac{1}{\mathbb{P}(E_n|\boldsymbol{\mu})} \frac{k!}{n!} \prod_{j=1}^k n_j! \mu_{n_j} \quad (11)$$

where $E_n := \left\{ \text{there exists } k \in \mathbb{N} \text{ such that } \sum_{j=1}^k n_j = n \right\}$. When $\sum_{s=1}^\infty s\mu_s < \infty$, a random partition with EPPF of form (11) has the microclustering property.

4.2. Balance-seeking ESC Models

The Balancedness of random partition models also plays an important role in ER tasks, since it is not plausible to assume that a small number of clusters dominate the whole database. For example, assume there are $n = 1000$ number of records of $K^+ = 100$ individuals. Then a balance-averse random partition model assigns a higher probability to a partition with cluster sizes $(901, 1, \ldots, 1)$ (length 100) compared to a partition with cluster sizes $(10, 10, \ldots, 10)$, suggesting the balance-seeking model would be better suited. The mi-

croclustering property has a similar rationale by limiting the growth rate of the maximum cluster size. While the micro-clustering property is based on the asymptotic behavior of cluster sizes, balancedness focuses on the non-asymptotic point of view and they complement each other.

We expand the current knowledge of ESC models by providing subclasses of ESC models with different balancedness properties. By Theorem 3.2 and (9), $\boldsymbol{\mu}$ being the zero-truncated Poisson distribution in (11) is the choice such that the B-sequence becomes identically zero, leading to the following theorem; see Appendix A for the proof.

Theorem 4.1. *Let $p^{(n)}$ be an EPPF of ESC model with fixed measure $\boldsymbol{\mu}$. Then for any $n = 1, 2, 3 \cdots$,*

- $p^{(n)}$ *is balance-averse if and only if $\boldsymbol{\mu} \geq_{\mathrm{lc}} \mathsf{Poisson}^+$,*

- $p^{(n)}$ *is balance-seeking if and only if $\boldsymbol{\mu} \leq_{\mathrm{lc}} \mathsf{Poisson}^+$,*

- $p^{(n)}$ *is balance-neutral if and only if $\boldsymbol{\mu} = \mathsf{Poisson}^+$.*

where $\mathsf{Poisson}^+$ indicates the p.m.f. of zero-truncated Poisson distribution with arbitrary parameter λ.

In Table 2, we provide various examples of $\boldsymbol{\mu}$ that lead to balance-averse, neutral, and balance-seeking ESC models. See Figure 3 (right) for an example of log-EPPF plot of balance-seeking model when $\boldsymbol{\mu} = \mathsf{Binomial}^+(10, 1/4)$, showing the increasing pattern for each k. Theorem 4.1 implies that a balance-seeking ESC model always possesses the microclustering property since all $\boldsymbol{\mu}$ which are relative log-concave to Poisson have finite moments of all orders (Johnson et al., 2013). Especially, $\boldsymbol{\mu} = \mathsf{Binomial}^+$ with fixed number of trial N has the *bounded microclustering property* (Betancourt et al., 2022), the size of the largest cluster is upper bounded.

With the Bayesian hierarchical model for ER task outlined in Section 4.1, one can choose a family of $\boldsymbol{\mu}$ to construct an ESC partition prior and carry out the posterior inference with Markov chain Monte Carlo (MCMC) methods. This includes a posterior sampling of cluster indices z_i as well as θ_μ (parameters of $\boldsymbol{\mu}$) from their conditional distributions. For example, the full conditional distribution of z_i is

$$\mathbb{P}(z_i = j|-) \propto p(\boldsymbol{x}|\mathbf{z}_{-i}, z_i = j, -)\mathbb{P}(z_i = j|\mathbf{z}_{-i}, \theta_\mu)$$

where the latter part $p_j^{-i} := \mathbb{P}(z_i = j|\mathbf{z}_{-i}, \theta_\mu)$ is

$$p_j^{-i} \propto \begin{cases} \frac{(n_j^{-i}+1)\mu_{n_j^{-i}+1}}{\mu_{n_j^{-i}}} \overset{let}{=} f(n_j^{-i}; \theta_\mu) & \text{if } j = 1, \ldots, k^{-i}, \\ \left(k^{-i} + 1\right)\mu_1 & \text{if } j = k^{-i} + 1 \end{cases}$$

where k^{-i} and n_j^{-i} are the number and sizes of clusters in $\mathbf{z}_{-i}$. In Appendix C, we provide details of EPPFs, reallocation rules and posterior inference algorithms of five different examples of $\boldsymbol{\mu}$ in Table 2.

Table 2. Examples of $\boldsymbol{\mu}$ in ESC model and their supports, grouped by balancedness property. Superscript $+$ indicates zero-truncated distribution. CMP stands for a Conway-Maxwell-Poisson distribution (Shmueli et al., 2005) with parameters λ and ν.

	$\boldsymbol{\mu}$	Support
Balance-averse ($B_s < 0$)	NegativeBinomial$^+$	$\mathbb{N}$
	Geometric	$\mathbb{N}$
	Logarithmic	$\mathbb{N}$
Neutral	Poisson$^+$	$\mathbb{N}$
Balance-seeking ($B_s > 0$)	Binomial$^+$	finite
	HyperGeometric$^+$	finite, must contain 1
	CMP$^+$ with $\nu > 1$	$\mathbb{N}$

Notably when $\boldsymbol{\mu} = \mathsf{Binomial}^+(N, p)$, the corresponding ESC model is balance-seeking and the reallocation probability to an existing cluster is proportional to $f(n_j^{-i}; \theta_\mu) = N - n_j^{-i}$ for $n_j^{-i} \leq N$, a linear function of n_j^{-i} with a negative slope showing the "rich-get-poorer" characteristic.

When n is very large, the standard Gibbs sampling algorithm which reallocates z_i one by one may suffer from slow mixing of the Markov chain. We utilize the *chaperones algorithm* (Miller et al., 2015; Zanella et al., 2016) which focuses on reallocations that have higher probabilities, but adopting other strategies (e.g. split-merge sampler) is also possible. We remark that the change of balancedness of Gibbs partition will not affect the computational cost as long as $\boldsymbol{V}, \boldsymbol{W}$ are readily available. More details on posterior inference algorithms can be found in Appendix C.

4.3. Real Data Application of Balance-seeking Models

In this section, we demonstrate the effectiveness of balance-seeking random partition for the ER task using the Survey of Income and Program Participation data (U.S. Census Bureau, 2009). We use the same dataset (SIPP1000) that Betancourt et al. (2020) used to benchmark the performance; the database with $n = 4116$ (number of records) and $K^+ = 1000$ (number of entities) was collected from the five waves of the longitudinal survey performed between 2005-2006. The main task is to recover the identifiers only using the $L = 5$ categorical fields (sex, birth year, birth month, race, and state of residence), assuming the unique identifiers are unknown and compare the results with the true partition.

With the SIPP1000 dataset and the hierarchical model described in Section 4.1, we compare the performance of four different ESC models where $\boldsymbol{\mu}$ is: (i) Binomial$^+$ with a fixed number of trials $N = 5$ (balance-seeking), (ii) Poisson$^+$ (balance-neutral), (iii) NegativeBinomial$^+$ (balance-averse), and (iv) Dirichlet (neither balance-seeking nor averse; Betancourt et al., 2020). Hyperpriors and MCMC specification details are described in Appendix D.
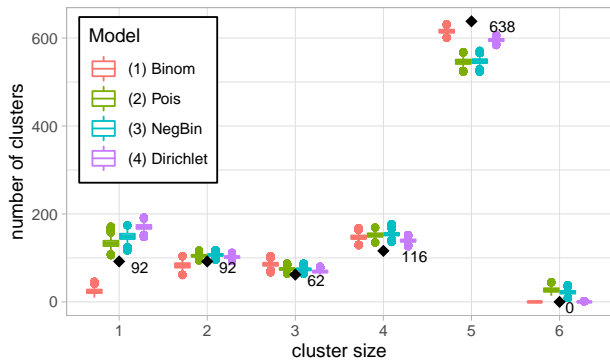
Figure 5. Posterior distributions (boxplots) of the number of clusters of size i for each models, m_i, for the SIPP1000 data. True m_i are shown as black diamonds and annotated by their values.

Table 3. SIPP1000 data: posterior mean and standard deviation of the number of entities ($K_{true}^+ = 1000$); FNRs and FDRs in %.[3]

$\boldsymbol{\mu}$ (balance-)	$\mathbb{E}(K^+|\boldsymbol{x})$	SD	FNR	FDR
Binom$^+$ (seeking)	951.5	6.2	**4.3**	4.6
Poisson$^+$ (neutral)	1043.8	8.8	4.9	3.3
NegBin$^+$ (averse)	1058.6	8.8	4.9	3.0
Dirichlet (n/a)	1076.4	5.3	4.8	**1.5**

The result in Figure 5 shows that the balance-seeking random partition prior penalizes the number of small-size clusters such as singletons, while the balance-averse prior does the opposite. This is evident since the balance-seeking model assigns a less prior probability to the partition with many singleton clusters where singletons make a partition unbalanced. Consequently, in Table 3, the balance-seeking model gives a more conservative estimate of the number of clusters with a lower false negative rate (FNR), but at the cost of yielding a higher false discovery rate (FDR).

The results indicate that balance-seeking random partition models are more effective in false negative control. Indeed, in many practical applications of ER, FNR control carries more weight than FDR control (Randall et al., 2013). For example, if the matched record is used to inform patients of their conditions, it is much more important to reduce FNR (fail to inform) than FDR (inform to the wrong patient). In Appendix D, we present additional simulation results with various different cluster size distributions and different distortion probabilities (β_ℓ). The results show that the balance-seeking model generally achieves a lower FNR rate compared to balance-neutral or balance-averse models.

---

[3]FNR and FDR are calculated based on the true partition and the point estimate obtained with Dahl (2006)'s method. Let CP, MP, WP be the number of correct, missed, and wrong pairs respectively, then FNR = MP/(MP+CP) and FDR = WP/(WP+CP).

5. Concluding Remarks

In this paper, we rigorously define and study the balancedness property of exchangeable random partition models, especially focusing on Gibbs partition models. By characterizing the "rich-get-richer" property as a balance-averseness, Corollary 3.4 provides an insight that two common assumptions (product-form exchangeability and projectivity) of random partition models lead to the "rich-get-richer" property. Although we conjecture that *every* infinitely exchangeable random partitions are balance-averse, we remark that our result is very general because almost all existing random partition models are either Gibbs or a mixture of Gibbs due to the simple reallocation rule (5). There is emerging literature beyond the Gibbs partition framework (Favaro et al., 2011; Camerlenghi et al., 2018) and analyzing the balancedness of such models is an interesting future direction to pursue.

Another interesting future direction is to study the balancedness of non-exchangeable random partition models. Examples include random partitions with covariate information (Park & Dunson, 2010; Müller et al., 2011), spatial or pairwise information (Blei & Frazier, 2011; Page & Quintana, 2016; Dahl et al., 2017; Xie & Xu, 2020), baseline partition serves as a "center" (Smith & Allenby, 2019; Paganin et al., 2021), temporal contiguity (Barry & Hartigan, 1992; 1993; Monteiro et al., 2011) and spatial contiguity (Teixeira et al., 2019; Luo et al., 2021a; Lee et al., 2021; Luo et al., 2021b). When random partition is no longer exchangeable, probability mass cannot be represented as an EPPF, and the extension of Definition 3.1 is highly nontrivial.

To conclude, we characterize the complete family of balance-neutral random partitions, which can be used as a noninformative prior choice in terms of balancedness. We also propose *B-sequence* to represent and compare the balancedness in a more principled and intuitive way. We offer various flexible balance-seeking random partition models, which provide better modeling solutions for applications where traditional balance-averse models are not suitable and prediction is not the main interest (due to the lack of projectivity). These models regularize the emergence of small-size clusters, which is useful in tasks such as entity resolution. We hope this article can serve as a useful guide for researchers to better understand the different behaviors of random partition models and practitioners to decide on a suitable model according to specific applications.

Acknowledgements

References

Ahuja, J. On the distribution of sum of independent positive binomial variables. *Canadian Mathematical Bulletin*, 13 (1):151–152, 1970.

Ahuja, J. Distribution of the sum of independent decapitated negative binomial variables. *The Annals of Mathematical Statistics*, 42(1):383–384, 1971.

Asuncion, A., Welling, M., Smyth, P., and Teh, Y. On smoothing and inference for topic models. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence, UAI 2009*, 2009.

Barry, D. and Hartigan, J. A. Product partition models for change point problems. *The Annals of Statistics*, pp. 260–279, 1992.

Barry, D. and Hartigan, J. A. A Bayesian analysis for change point problems. *Journal of the American Statistical Association*, 88(421):309–319, 1993.

Berger, J. O., Bernardo, J. M., and Sun, D. Objective priors for discrete parameter spaces. *Journal of the American Statistical Association*, 107(498):636–648, 2012.

Betancourt, B., Zanella, G., and Steorts, R. C. Random partition models for microclustering tasks. *Journal of the American Statistical Association*, pp. 1–13, 2020.

Betancourt, B., Sosa, J., and Rodríguez, A. A prior for record linkage based on allelic partitions. *Computational Statistics & Data Analysis*, 172:107474, 2022.

Binette, O. and Steorts, R. C. (Almost) all of entity resolution. *Science Advances*, 8(12):eabi8021, 2022.

Blei, D. M. and Frazier, P. I. Distance dependent chinese restaurant processes. *Journal of Machine Learning Research*, 12(8), 2011.

Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent Dirichlet allocation. *the Journal of Machine Learning Research*, 3: 993–1022, 2003.

Brylawski, T. The lattice of integer partitions. *Discrete mathematics*, 6(3):201–219, 1973.

Camerlenghi, F., Lijoi, A., and Prünster, I. Bayesian non-parametric inference beyond the Gibbs-type framework. *Scandinavian Journal of Statistics*, 45(4):1062–1091, 2018.

Dahl, D. B. Model-based clustering for expression data via a Dirichlet process mixture model. *Bayesian inference for gene expression and proteomics*, 4:201–218, 2006.

Dahl, D. B., Day, R., and Tsai, J. W. Random partition distribution indexed by pairwise information. *Journal of the American Statistical Association*, 112(518):721–732, 2017.

Daly, A. J., Baetens, J. M., and De Baets, B. Ecological diversity: measuring the unmeasurable. *Mathematics*, 6 (7):119, 2018.

De Blasi, P., Favaro, S., Lijoi, A., Mena, R. H., Prunster, I., and Ruggiero, M. Are Gibbs-type priors the most natural generalization of the Dirichlet process? *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 37(02):212–229, 2015.

Ewens, W. J. The sampling theory of selectively neutral alleles. *Theoretical population biology*, 3(1):87–112, 1972.

Favaro, S., Lijoi, A., Mena, R. H., and Prünster, I. Bayesian non-parametric inference for species variety with a two-parameter Poisson–Dirichlet process prior. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(5):993–1008, 2009.

Favaro, S., Prünster, I., and Walker, S. G. On a class of random probability measures with general predictive structure. *Scandinavian Journal of Statistics*, 38(2):359–376, 2011.

Ferguson, T. S. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, pp. 209–230, 1973.

Frühwirth-Schnatter, S., Malsiner-Walli, G., and Grün, B. Generalized mixtures of finite mixtures and telescoping sampling. *Bayesian Analysis*, 16(4):1279–1307, 2021.

Geng, J., Bhattacharya, A., and Pati, D. Probabilistic community detection with unknown number of communities. *Journal of the American Statistical Association*, 114(526): 893–905, 2019.

Gnedin, A. A species sampling model with finitely many types. *Electronic Communications in Probability*, 15: 79–88, 2010.

Gnedin, A. and Pitman, J. Exchangeable Gibbs partitions and Stirling triangles. *Zapiski Nauchnykh Seminarov POMI*, 325:83–102, 2005.

Green, P. J. and Richardson, S. Modelling heterogeneity with and without the Dirichlet process. *Scandinavian Journal of Statistics*, 28(2):355–375, 2001.

Greve, J., Grün, B., Malsiner-Walli, G., and Frühwirth-Schnatter, S. Spying on the prior of the number of data clusters and the partition distribution in Bayesian cluster analysis. *Australian & New Zealand Journal of Statistics*, in press, 2022.

Hill, M. O. Diversity and evenness: a unifying notation and its consequences. *Ecology*, 54(2):427–432, 1973.

Hurlbert, S. H. The nonconcept of species diversity: a critique and alternative parameters. *Ecology*, 52(4):577–586, 1971.

Jensen, S. T. and Liu, J. S. Bayesian clustering of transcription factor binding motifs. *Journal of the American Statistical Association*, 103(481):188–200, 2008.

Johnson, N. L., Kemp, A. W., and Kotz, S. *Univariate discrete distributions*, volume 444. John Wiley & Sons, 2005.

Johnson, O., Kontoyiannis, I., and Madiman, M. Log-concavity, ultra-log-concavity, and a maximum entropy property of discrete compound Poisson measures. *Discrete Applied Mathematics*, 161(9):1232–1250, 2013.

Kemp, C., Tenenbaum, J. B., Griffiths, T. L., Yamada, T., and Ueda, N. Learning systems of concepts with an infinite relational model. In *AAAI*, volume 3, pp. 5, 2006.

Kingman, J. F. The representation of partition structures. *Journal of the London Mathematical Society*, 2(2):374–380, 1978.

Klugman, S. A., Panjer, H. H., and Willmot, G. E. *Loss models: from data to decisions*, volume 715. John Wiley & Sons, 2012.

Lee, C. J., Luo, Z. T., and Sang, H. T-LoHo: A Bayesian regularization model for structured sparsity and smoothness on graphs. *Advances in Neural Information Processing Systems*, 34, 2021.

Lee, J., Quintana, F. A., Müller, P., and Trippa, L. Defining predictive probability functions for species sampling models. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 28(2):209, 2013.

Legramanti, S., Rigon, T., Durante, D., and Dunson, D. B. Extended stochastic block models with application to criminal networks. *Annals of Applied Statistics*, in press, 2022.

Lijoi, A., Mena, R. H., and Prünster, I. Controlling the reinforcement in Bayesian non-parametric mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4):715–740, 2007.

Lijoi, A., Prünster, I., and Rigon, T. The Pitman–Yor multinomial process for mixture modelling. *Biometrika*, 107 (4):891–906, 2020.

Lomelí, M., Favaro, S., and Teh, Y. W. A marginal sampler for σ-stable Poisson–Kingman mixture models. *Journal of Computational and Graphical Statistics*, 26(1):44–53, 2017.

Lu, J., Li, M., and Dunson, D. Reducing over-clustering via the powered chinese restaurant process. *arXiv preprint arXiv:1802.05392*, 2018.

Luo, Z., Sang, H., and Mallick, B. A Bayesian contiguous partitioning method for learning clustered latent variables. *Journal of Machine Learning Research*, 22, 2021a.

Luo, Z. T., Sang, H., and Mallick, B. BAST: Bayesian additive regression spanning trees for complex constrained domain. *Advances in Neural Information Processing Systems*, 34, 2021b.

Miller, J., Betancourt, B., Zaidi, A., Wallach, H., and Steorts, R. C. Microclustering: When the cluster sizes grow sublinearly with the size of the data set. *NIPS Bayesian Nonparametrics: The Next Generation Workshop Series*, 2015.

Miller, J. W. and Harrison, M. T. Mixture models with a prior on the number of components. *Journal of the American Statistical Association*, 113(521):340–356, 2018.

Monteiro, J. V., Assunçao, R. M., and Loschi, R. H. Product partition models with correlated parameters. *Bayesian Analysis*, 6(4):691–726, 2011.

Müller, P., Quintana, F., and Rosner, G. L. A product partition model with regression on covariates. *Journal of Computational and Graphical Statistics*, 20(1):260–278, 2011.

Neal, R. M. Slice sampling. *The Annals of Statistics*, 31(3): 705–767, 2003.

Nowicki, K. and Snijders, T. A. B. Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, 96(455):1077–1087, 2001.

Paganin, S., Herring, A. H., Olshan, A. F., and Dunson, D. B. Centered partition processes: Informative priors for clustering (with discussion). *Bayesian Analysis*, 16 (1):301–370, 2021.

Page, G. L. and Quintana, F. A. Spatial product partition models. *Bayesian Analysis*, 11(1):265–298, 2016.

Park, J.-H. and Dunson, D. B. Bayesian generalized product partition model. *Statistica Sinica*, pp. 1203–1226, 2010.

Patil, G. and Taillie, C. Diversity as a concept and its measurement. *Journal of the American Statistical Association*, 77(379):548–561, 1982.

Patil, G. and Wani, J. On certain structural properties of the logarithmic series distribution and the first type stirling distribution. *Sankhyā: The Indian Journal of Statistics, Series A*, pp. 271–280, 1965.

Perman, M., Pitman, J., and Yor, M. Size-biased sampling of Poisson point processes and excursions. *Probability Theory and Related Fields*, 92(1):21–39, 1992.

Petrone, S. and Raftery, A. E. A note on the Dirichlet process prior in Bayesian nonparametric inference with partial exchangeability. *Statistics & Probability Letters*, 36(1):69–83, 1997.

Pitman, J. Exchangeable and partially exchangeable random partitions. *Probability theory and related fields*, 102(2): 145–158, 1995.

Pitman, J. Poisson-Kingman partitions. *Institute of Mathematical Statistics Lecture Notes - Monograph Series*, pp. 1–34, 2003.

Pitman, J. *Combinatorial stochastic processes: Ecole d'eté de probabilités de saint-flour xxxii-2002*. Springer, 2006.

Pitman, J. and Yor, M. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, pp. 855–900, 1997.

Poux-Médard, G., Velcin, J., and Loudcher, S. Powered Dirichlet process for controlling the importance of" rich-get-richer" prior assumptions in Bayesian clustering. *arXiv preprint arXiv:2104.12485*, 2021.

Randall, S. M., Ferrante, A. M., Boyd, J. H., and Semmens, J. B. The effect of data cleaning on record linkage quality. *BMC medical informatics and decision making*, 13(1): 1–10, 2013.

Schmidt, M. N. and Morup, M. Nonparametric Bayesian modeling of complex networks: An introduction. *IEEE Signal Processing Magazine*, 30(3):110–128, 2013.

Shi, H., Gerlach, M., Diersen, I., Downey, D., and Amaral, L. A new evaluation framework for topic modeling algorithms based on synthetic corpora. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 816–826. PMLR, 2019.

Shmueli, G., Minka, T. P., Kadane, J. B., Borle, S., and Boatwright, P. A useful distribution for fitting discrete data: revival of the Conway–Maxwell–Poisson distribution. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(1):127–142, 2005.

Smith, A. N. and Allenby, G. M. Demand models with random partitions. *Journal of the American Statistical Association*, 2019.

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

Steorts, R. C., Hall, R., and Fienberg, S. E. A Bayesian approach to graphical record linkage and deduplication. *Journal of the American Statistical Association*, 111(516): 1660–1672, 2016.

Steorts, R. C., Betancourt, B., and Zanella, G. *microclustr: Entity Resolution with Random Partition Priors for Microclustering*, 2020. URL https://CRAN.R-project.org/package=microclustr. R package version 0.1.0.

Sundt, B. and Jewell, W. S. Further results on recursive evaluation of compound distributions. *ASTIN Bulletin: The Journal of the IAA*, 12(1):27–39, 1981.

Tate, R. F. and Goen, R. Minimum variance unbiased estimation for the truncated Poisson distribution. *The Annals of Mathematical Statistics*, pp. 755–765, 1958.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

Teixeira, L. V., Assunção, R. M., and Loschi, R. H. Bayesian space-time partitioning by sampling and pruning spanning trees. *Journal of Machine Learning Research*, 20: 85–1, 2019.

U.S. Census Bureau. Survey of income and program participation (SIPP) 2004 panel, 2009. URL https://www.icpsr.umich.edu/web/pages/.

Wallach, H., Jensen, S., Dicker, L., and Heller, K. An alternative prior process for nonparametric Bayesian clustering. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 892–899. JMLR Workshop and Conference Proceedings, 2010.

Welling, M. Flexible priors for infinite mixture models. In *Proceedings of the Workshop on Learning with Nonparametric Bayesian Methods, 23rd ICML*, 2006.

Whitt, W. Uniform conditional variability ordering of probability distributions. *Journal of Applied Probability*, 22 (3):619–633, 1985.

Xie, F. and Xu, Y. Bayesian repulsive Gaussian mixture model. *Journal of the American Statistical Association*, 115(529):187–203, 2020.

Yu, Y. Relative log-concavity and a pair of triangle inequalities. *Bernoulli*, 16(2):459–470, 2010.

Zanella, G., Betancourt, B., Miller, J. W., Wallach, H., Zaidi, A., and Steorts, R. C. Flexible models for microclustering with application to entity resolution. *Advances in Neural Information Processing Systems*, 29, 2016.

In Appendix A, we provide the proofs of all theorems in this paper: Theorem 3.2, Corollary 3.3, Theorem 3.5, Theorem 3.6, and Theorem 4.1. In Appendix B, we provide a detailed analysis of the B-sequence on its graphical interpretation and properties regarding balancedness. In Appendix C, we provide more details on the subclasses of the ESC model discussed in Section 4.2 and their corresponding inference algorithms using MCMC methods. Finally, in Appendix D, we provide details on the SIPP1000 data analysis results in Section 4.3 and also present the simulation study results. In the supplementary material, we provide an interactive demo regarding the balancedness of the Ewens-Pitman two-parameter family and R code to run the balance-seeking ESC model described in Section 4.2 and Appendix C.

A. Proofs of Theorems

A.1. Proof of Theorem 3.2

(Log-convexity of $\boldsymbol{W}$ implies balance-averseness) Let $\boldsymbol{n}, \boldsymbol{n}' \in \mathcal{I}_n^k$ be two distinct integer partitions of n into k parts. To show that $\boldsymbol{n} \prec \boldsymbol{n}' \implies p^{(n)}(\boldsymbol{n}) \geq p^{(n)}(\boldsymbol{n}')$, by the transitivity of $\prec$ and $\leq$, it is sufficient to show for the case when $\boldsymbol{n}'$ can be reached by a single one-step downshift from $\boldsymbol{n}$. Let $\boldsymbol{n} = (n_1, \cdots, n_k)$ and its one-step downshift $\boldsymbol{n}' = (n_1', \cdots, n_k') = (n_1, \cdots n_u - 1, \cdots, n_v + 1, \cdots n_k)$ for some $1 \leq u < v \leq k$ such that $n_u - 1 \geq n_v + 1$. Then the log-convexity of $\boldsymbol{W}$, i.e. convexity of the sequence $(\log W_s)_{s=1}^\infty$ implies

$$\log W_{n_v+1} - \log W_{n_v} \leq \log W_{n_u} - \log W_{n_u-1} \qquad \because n_u - 1 \geq n_v + 1$$
$$\implies \log W_{n_u} + \log W_{n_v} \geq \log W_{n_u-1} + \log W_{n_v+1}$$
$$\implies W_{n_u} W_{n_v} \geq W_{n_u-1} W_{n_v+1} = W_{n_u'} W_{n_v'}$$
$$\implies p^{(n)}(\boldsymbol{n}) = V_{n,k} \prod_{j=1}^k W_{n_j} \geq V_{n,k} \prod_{j=1}^k W_{n_j'} = p^{(n)}(\boldsymbol{n}')$$

Note that $W_s > 0$ for all s, since $W_1 = 1$ and $W_s^2 \leq W_{s-1} W_{s+1}$ for all s. To see this, if $W_s = 0$ for some $s = s^*$, then it implies $W_s = 0$ for all $s \leq s^*$, which contradicts to $W_1 = 1$.

(Balance-averseness implies log-convexity of $\boldsymbol{W}$) Let $\boldsymbol{n} = (n_1, \cdots, n_k)$ and $\boldsymbol{n}' = (n_1, \cdots, n_u - 1, \cdots, n_v + 1, \cdots, n_k)$ with $n_u - 1 = n_v + 1 \overset{let}{=} n^*$. Then $p^{(n)}(\boldsymbol{n}) \geq p^{(n)}(\boldsymbol{n}')$ by balance-averseness, and this implies $W_{n^*}^2 \leq W_{n^*+1} W_{n^*-1}$. Since $n^* \in \{2, 3, \ldots\}$ can be chosen arbitrarily as balance-averseness holds for any k and n such that $k \leq n$, thus $\boldsymbol{W}$ is log-convex.

The equivalency between balance-seeking and log-concavity follows with similar arguments.

A.2. Proof of Corollary 3.3

Let $\Pi_{n+1} \sim \text{Gibbs}_{[n+1]}(\boldsymbol{V}, \boldsymbol{W})$ so that the EPPF is $p^{(n+1)}(n_1, \ldots, n_k) = V_{n+1,k} \prod_{l=1}^k W_{n_l}$. Let the cluster membership of first n datapoints with k clusters be $\mathbf{z}_{1:n} \in \{1, \ldots, k\}^n$ and $n_l = \sum_{i=1}^n 1(z_i = l)$ for $l = 1, \ldots, k$. We first derive the expression (5), using $W_1 = 1$ by definition,

$$\mathbb{P}(z_{n+1} = j | \mathbf{z}_{1:n}) \propto \mathbb{P}(z_{n+1} = j, \mathbf{z}_{1:n})$$
$$= \begin{cases} V_{n+1,k} W_{n_j+1} \prod_{l \neq j} W_{n_l} & \text{if } j = 1, \ldots, k \\ V_{n+1,k+1} \prod_{l=1}^k W_{n_l} & \text{if } j = k+1 \end{cases}$$
$$\propto \begin{cases} W_{n_j+1}/W_{n_j} \overset{let}{=} f(n_j) & \text{if } j = 1, \ldots, k \\ V_{n+1,k+1}/V_{n+1,k} \overset{let}{=} g(n,k) & \text{if } j = k+1 \end{cases}$$

For any $n \in \mathbb{N}$, reallocation probability to the existing cluster $f(n_j) = W_{n_j+1}/W_{n_j}$ is an increasing (decreasing) function of $n_j = 1, 2, \ldots$ if and only if $\boldsymbol{W} = (W_s)_{s=1}^\infty$ is log-convex (log-concave). Combined with Theorem 3.2, we have the Corollary 3.3.

A.3. Proof of Theorem 3.5

Let $p^{(n)}$ be an EPPF of mixture of Gibbs partition of the form in (6), i.e.,

$$p^{(n)}(n_1, \ldots, n_k) = \int_\Theta V_{n,k}(\vartheta) \prod_{j=1}^k W_{n_j}(\vartheta) \nu(d\vartheta)$$

Assume that $(W_s(\vartheta))_{s=1}^\infty$ is log-convex for each $\vartheta \in \Theta$, where Θ is the domain of ϑ that is either discrete or continuous. Fix ϑ at an arbitrary value, and let $\boldsymbol{n}, \boldsymbol{n}' \in \mathcal{I}_n^k$. To show $\boldsymbol{n} \prec \boldsymbol{n}' \implies p^{(n)}(\boldsymbol{n}) \geq p^{(n)}(\boldsymbol{n}')$, by the transitivity of $\prec$ and $\leq$, it is again sufficient to show for the case when $\boldsymbol{n}'$ can be reached by a single one-step downshift from $\boldsymbol{n}$. Let $\boldsymbol{n} = (n_1, \cdots, n_k)$ and $\boldsymbol{n}' = (n_1', \cdots, n_k') = (n_1, \cdots n_u - 1, \cdots, n_v + 1, \cdots n_k)$ for some $1 \leq u < v \leq k$ such that $n_u - 1 \geq n_v + 1$. Then the convexity of sequence $(\log W_s(\vartheta))_{s=1}^\infty$ implies

$$\log W_{n_v+1}(\vartheta) - \log W_{n_v}(\vartheta) \leq \log W_{n_u}(\vartheta) - \log W_{n_u-1}(\vartheta) \qquad \because n_u - 1 \geq n_v + 1$$
$$\implies \log W_{n_u}(\vartheta) + \log W_{n_v}(\vartheta) \geq \log W_{n_u-1}(\vartheta) + \log W_{n_v+1}(\vartheta)$$
$$\implies W_{n_u}(\vartheta) W_{n_v}(\vartheta) \geq W_{n_u-1}(\vartheta) W_{n_v+1}(\vartheta) = W_{n_u'}(\vartheta) W_{n_v'}(\vartheta)$$
$$\implies V_{n,k}(\vartheta) \prod_{j=1}^k W_{n_j}(\vartheta) \geq V_{n,k}(\vartheta) \prod_{j=1}^k W_{n_j'}(\vartheta) \geq 0,$$

where the last line holds for arbitrary $\vartheta \in \Theta$. Since integration preserves inequality of functions, we have $p^{(n)}(\boldsymbol{n}) \geq p^{(n)}(\boldsymbol{n}')$. The statement that log-concave $(W_s(\vartheta))_{s=1}^\infty$ implies a balance-seeking $p^{(n)}$ follows with similar arguments.

A.4. Proof of Theorem 3.6

Let p be an EPPF of balance-neutral, infinitely exchangeable random partition. Note that any $\boldsymbol{n} \in \mathcal{I}_n^k$ is comparable with the most unbalanced one $\boldsymbol{n}^\star = (n-k+1, 1, \cdots, 1)$ and the most balanced one $\boldsymbol{n}^{\star\star} = (\underbrace{\lceil n/k \rceil, \cdots, \lceil n/k \rceil}_{(n \bmod k) \text{ times}}, \lfloor n/k \rfloor, \cdots, \lfloor n/k \rfloor)$,

and $\boldsymbol{n}^\star \prec \boldsymbol{n}^{\star\star} \implies p(\boldsymbol{n}^\star) = p(\boldsymbol{n}^{\star\star})$. We have $p(\boldsymbol{n}) = p(\boldsymbol{n}')$ for all $\boldsymbol{n}, \boldsymbol{n}' \in \mathcal{I}_n^k$. Thus, EPPF p only depends on n and k and hence can be written as $p(n_1, \ldots, n_k) \stackrel{let}{=} v(n, k)$. In other words, p is a Gibbs partition with $V_{n,k} = v(n, k)$ and $W_s \equiv 1$. Gnedin & Pitman (2005) showed that *any* infinitely exchangeable Gibbs partition with $\sigma < 0$, including $\sigma = -\infty$, can be expressed as a mixture of Ewens-Pitman two parameter family $(\sigma, |\sigma| K)$ with some mixing distribution q on the number of components $K \in \mathbb{N}$, thus the theorem follows.

A.5. Proof of Theorem 4.1

By Theorem 3.2, to prove the equivalency of balance-averseness and $\boldsymbol{\mu} \geq_{\text{lc}} \text{Poisson}^+$, it is sufficient to prove that $\boldsymbol{\mu} \geq_{\text{lc}} \text{Poisson}^+$ if and only if the corresponding $\boldsymbol{W}$ is log-convex. Rewrite the EPPF of ESC model as:

$$p^{(n)}(n_1, \ldots, n_k | \boldsymbol{\mu}) = \frac{\mu_1^k}{\mathbb{P}(E_n | \boldsymbol{\mu})} \frac{k!}{n!} \prod_{j=1}^k n_j! \mu_{n_j} / \mu_1 = V_{n,k} \prod_{j=1}^k W_{n_j} \tag{12}$$

The log-convexity of $(W_s)_{s=1}^\infty$ is equivalent to the convexity of $(\log(s! \mu_s))_{s=1}^\infty$, or concavity of $(-\log(s!\mu_s) + as + b)_{s=1}^\infty$ for arbitrary constants a and b. The log-convexity of $(s!\mu_s)_{s=1}^\infty$, in other words $s\mu_s^2 \leq (s+1)\mu_{s-1}\mu_{s+1}$, implies $\mu_s > 0$ for all $s = 1, 2, \ldots$ (full support), since assuming $\mu_s = 0$ for some $s = s^*$ leads to $\mu_s = 0$ for $s \leq s^*$ that contradicts with $\mu_1 > 0$. By choosing $a = \log \lambda$ and $b = \log(e^{-\lambda}/(1 - e^{-\lambda}))$ for some $\lambda > 0$, we can see that concavity of $(-\log(s!\mu_s) + as + b)_{s=1}^\infty$ is equivalent to $\boldsymbol{\mu} \geq_{\text{lc}} \text{Poisson}^+(\lambda)$, by the definition of relative log-concavity and $\mu_s > 0$ for all s (full support). Next, to show that $\boldsymbol{\mu} \leq_{\text{lc}} \text{Poisson}^+$ if and only if the corresponding $\boldsymbol{W}$ is log-concave, we can apply the similar argument as above by flipping the inequalities. Now we prove the balance-neutral case, which amounts to showing that $\boldsymbol{\mu} = \text{Poisson}^+$ is the only discrete probability measure such that $\boldsymbol{W} = (s!\mu_s/\mu_1)_{s=1}^\infty$ is both log-convex and log-concave. In other words, if $\log s!\mu_s$ is a linear sequence of s, then $\boldsymbol{\mu} = \text{Poisson}^+$. Let $\log s!\mu_s = as + b$ for some constants a and b, then $\mu_s = (e^{as}e^b)/s!$ for $s = 1, 2, \cdots$. Solving $\sum_{s=1}^\infty (e^{as}e^b)/s! = 1$ leads to $b = -\log(\exp(e^a) - 1)$. Now substituting $\lambda = e^a > 0$, we have $\mu_s = \frac{\lambda^s}{s!} \frac{1}{e^\lambda - 1}$, a probability mass function of $\text{Poisson}^+(\lambda)$.

B. Details on B-sequence

B.1. Slope Interpretation of B-sequence

This subsection provides the derivation of Equation (10). Let $\boldsymbol{n} = (n_1, \ldots, n_k), \boldsymbol{n}' = (n_1', \ldots, n_k') \in \mathcal{I}_n^k$ where $\boldsymbol{n}'$ is a $(**)-$cover of $\boldsymbol{n}$. Then there exists some $s \geq 2$ such that

$$\boldsymbol{n} = (n_1, \cdots n_u, \cdots, n_v, \cdots, n_k) \mapsto (n_1, \cdots, \underbrace{n_u - 1}_{=s}, \cdots, \underbrace{n_v + 1}_{=s}, \cdots, n_k) = \boldsymbol{n}'$$

where $s = n_u' = n_u - 1 = n_v' = n_v + 1$ with $u < v$. By Equation (3), since $\boldsymbol{n}$ and $\boldsymbol{n}'$ both have k clusters,

$$\log p^{(n)}(\boldsymbol{n}') - \log p^{(n)}(\boldsymbol{n}) = \log W_{n_u'} + \log W_{n_v'} - \log W_{n_u} - \log W_{n_v} = -\log W_{s+1} + 2\log W_s - \log W_{s-1}$$

Also, for the Shannon index $H(\boldsymbol{n}) = -\sum_{j=1}^{k} (n_j/n) \log(n_j/n)$,

$$H(\boldsymbol{n}') - H(\boldsymbol{n}) = \left(-\frac{n_u'}{n}\log\frac{n_u'}{n} - \frac{n_v'}{n}\log\frac{n_v'}{n}\right) - \left(-\frac{n_u}{n}\log\frac{n_u}{n} - \frac{n_v}{n}\log\frac{n_v}{n}\right) \tag{13}$$

$$= \frac{1}{n}\left[(s+1)\log\frac{s+1}{n} - 2s\log\frac{s}{n} + (s-1)\log\frac{s-1}{n}\right] \tag{14}$$

$$\approx^4 \frac{1}{n}\left[\frac{d^2}{dx^2}\left(x\log\frac{x}{n}\right)\right]_{x=s} = \frac{1}{n}\frac{1}{s} \tag{15}$$

Therefore

$$\frac{\log p^{(n)}(\boldsymbol{n}') - \log p^{(n)}(\boldsymbol{n})}{H(\boldsymbol{n}') - H(\boldsymbol{n})} \approx n \times -s(\log W_{s+1} - 2\log W_s + \log W_{s-1}) = nB_s(\boldsymbol{W}) \tag{16}$$

and see Figure 4 for the grapical illustration.

B.2. B-sequence and Balancedness

First, we prove that the ordering based on the B-sequence comparison is equivalent to the log-concavity ordering.

Theorem B.1. $\boldsymbol{W} \leq_{\mathrm{lc}} \boldsymbol{W}'$ *if and only if* $B_s(\boldsymbol{W}) \geq B_s(\boldsymbol{W}')$ *for all* $s \geq 2$.

Proof. Below, we prove the cases when $\boldsymbol{W}$ have a full support or a finite support, respectively. Assume that $\boldsymbol{W}$ have a full support; i.e. $W_s > 0$ for all $s = 1, 2 \ldots$. Then $\boldsymbol{W} \leq_{\mathrm{lc}} \boldsymbol{W}'$ implies $\boldsymbol{W}'$ also have a full support by its definition. Note that $B_s(\boldsymbol{W}) \geq B_s(\boldsymbol{W}'), \forall s$ also implies $B_s(\boldsymbol{W}') < \infty$ for all $s = 2, 3, \ldots$, and thus $\boldsymbol{W}'$ have a full support. Therefore, either assuming $\boldsymbol{W} \leq_{\mathrm{lc}} \boldsymbol{W}'$ or $B_s(\boldsymbol{W}) \geq B_s(\boldsymbol{W}')$ for $s = 2, 3, \ldots$, we have $\mathrm{supp}(\boldsymbol{W}) \subseteq \mathrm{supp}(\boldsymbol{W}')$. Then,

$$\boldsymbol{W} \leq_{\mathrm{lc}} \boldsymbol{W}' \iff (\log(W_s/W_s')) \text{ is concave}$$

$$\iff \left(\frac{W_s}{W_s'}\right)^2 \geq \frac{W_{s+1}W_{s-1}}{W_{s+1}'W_{s-1}'}, \quad \forall s = 2, 3, \ldots$$

$$\iff \frac{W_{s+1}'W_{s-1}'}{(W_s')^2} \geq \frac{W_{s+1}W_{s-1}}{(W_s)^2}, \quad \forall s = 2, 3, \ldots$$

$$\iff \log W_{s-1}' - 2\log W_s' + \log W_{s+1}' \geq \log W_{s-1} - 2\log W_s + \log W_{s+1}, \quad \forall s = 2, 3, \ldots$$

$$\iff B_s(\boldsymbol{W}) \geq B_s(\boldsymbol{W}'), \quad \forall s = 2, 3, \ldots$$

Next, assume that $\boldsymbol{W}$ has a finite support $\{1, 2, \ldots, m\}$, which only happens when $\boldsymbol{W}$ is log-concave. Then by $\boldsymbol{W} \leq_{\mathrm{lc}} \boldsymbol{W}'$, $\boldsymbol{W}'$ must have either a finite or a full support, and if it has a finite support $\{1, 2, \ldots, m'\}$, then $m' \geq m$. On the other hand, $B_s(\boldsymbol{W}) \geq B_s(\boldsymbol{W}')$ for $s = 2, 3, \ldots$ implies that the support of $\boldsymbol{W}$ must be contained in the support of $\boldsymbol{W}'$, since

---

[4]The approximation error is about $0.023/n$ for $s = 2$, $0.007/n$ for $s = 3$, $0.003/n$ for $s = 4$, $0.001/n$ for $s = 5$, and less than $0.001/n$ for $s \geq 6$.

otherwise there is an s such that $B_s(\boldsymbol{W}') = +\infty$ which contradicts with $B_s(\boldsymbol{W}) < \infty$. Thus, either assuming $\boldsymbol{W} \leq_{\mathrm{lc}} \boldsymbol{W}'$ or $B_s(\boldsymbol{W}) \geq B_s(\boldsymbol{W}')$ for $s = 2, 3, \ldots$, we have $\mathrm{supp}(\boldsymbol{W}) \subseteq \mathrm{supp}(\boldsymbol{W}')$. Then,

$$
\begin{aligned}
\boldsymbol{W} \leq_{\mathrm{lc}} \boldsymbol{W}' &\iff (\log(W_s/W_s')) \text{ is concave in } \mathrm{supp}(\boldsymbol{W}) \\
&\iff \left(\frac{W_s}{W_s'}\right)^2 \geq \frac{W_{s+1}W_{s-1}}{W_{s+1}'W_{s-1}'}, \quad \forall s = 2, 3, \ldots, m-1 \\
&\iff \frac{W_{s+1}'W_{s-1}'}{(W_s')^2} \geq \frac{W_{s+1}W_{s-1}}{(W_s)^2}, \quad \forall s = 2, 3, \ldots, m-1 \\
&\iff \log W_{s-1}' - 2\log W_s' + \log W_{s+1}' \geq \log W_{s-1} - 2\log W_s + \log W_{s+1}, \quad \forall s = 2, 3, \ldots, m-1 \\
&\iff B_s(\boldsymbol{W}) \geq B_s(\boldsymbol{W}'), \quad \forall s = 2, 3, \ldots, m-1
\end{aligned}
$$

which proves the claim. $\qquad\square$

By Theorem B.1 and Theorem 3.2, we have the following corollary:

Corollary B.2. *For any $n = 1, 2, \ldots$, $\mathsf{Gibbs}_{[n]}(\boldsymbol{V}, \boldsymbol{W})$ is*

- *balance-averse if and only if $B_s(\boldsymbol{W}) \leq 0$ for all $s \geq 2$*

- *balance-neutral if and only if $B_s(\boldsymbol{W}) = 0$ for all $s \geq 2$*

- *balance-seeking if and only if $B_s(\boldsymbol{W}) \geq 0$ for all $s \geq 2$*

To see this, $B_s(\boldsymbol{W}) \equiv 0$ if and only if $\boldsymbol{W}$ has a full support and $W_s^2 = W_{s+1}W_{s-1}$ for all $s \geq 2$, which is equivalent to (W_s) being log-linear (both log-convex and log-concave). Similar argument follows for the balance-averse and the balance-seeking cases.

C. Details on ESC Model and Posterior Inference

C.1. Examples of ESC Model with Different Balancedness

In this section, we provide details of ESC models with different balancedness properties. We consider 5 different cases, where $\boldsymbol{\mu} = (\mu_s)_{s=1}^{\infty}$ is a probability mass function of 1) shifted binomial, 2) zero-truncated binomial, 3) zero-truncated Poisson, 4) zero-truncated (extended) negative binomial, and 5) logarithmic distribution. We assume the number of trials N is fixed in zero-truncated binomial, while shifted binomial is not. Recall that EPPF of ESC model with fixed $\boldsymbol{\mu}$ and $\mu_1 > 0$ is written as follows (Betancourt et al., 2020):

$$
p^{(n)}(n_1, \ldots, n_k | \boldsymbol{\mu}) = \frac{1}{\mathbb{P}(E_n | \boldsymbol{\mu})} \frac{k!}{n!} \prod_{j=1}^{k} n_j! \mu_{n_j} \tag{17}
$$

where the event E_n is $E_n = \left\{ \text{there exists } k \in \mathbb{N} \text{ such that } \sum_{j=1}^{k} n_j = n \right\}$. We can calculate the normalizing constant $\mathbb{P}(E_n | \boldsymbol{\mu})$ with the k-fold convolution of random variables following the distribution $\boldsymbol{\mu}$. That is, if $X_1, \ldots, X_k \overset{iid}{\sim} \boldsymbol{\mu}$ and $S_k := X_1 + \cdots + X_k$, then $\mathbb{P}(E_n | \boldsymbol{\mu}) = \sum_{k=1}^{n} \mathbb{P}(S_k = n)$, which is always positive if $\mu_1 > 0$. We provide five examples of $\boldsymbol{\mu}$ in the ESC model in Table 4, with their complete form of EPPFs including the normalizing constant. For the derivation of k-fold convolution formulas, we refer the readers to Ahuja (1970); Tate & Goen (1958); Ahuja (1971); Patil & Wani (1965). Regarding the notations, $S_1(n, k)$ and $S_2(n, k)$ are the Stirling number of the first and second kind respectively, and the binomial coefficient with real argument is defined as $\binom{a}{b} = \frac{a(a-1)\cdots(a-b+1)}{b!}$ for $a \in \mathbb{R}$ and $b \in \mathbb{N}$. When $r = 1$ in $\mathsf{NegBin}^+(r, p)$, it becomes $\mathsf{Geometric}(p)$ distribution and the normalizing constant $\mathbb{P}(E_n | \boldsymbol{\mu})$ is simply p. We note that $\mathsf{NegBin}^+(r, p)$ with $r > -1, r \neq 0$ is also called Engen's extended negative binomial distribution since it admits negative r values.

Table 4. Examples of $\boldsymbol{\mu}$ in the ESC model. ESC models with shifted binomial and zero-truncated binomial are balance-seeking, zero-truncated Poisson is balance-neutral, and zero-truncated negative binomial and logarithmic are balance-averse.

$\boldsymbol{\mu}$	Support	μ_s	$\mathbb{P}(E_n \mid \boldsymbol{\mu})$
$1 + \text{Bin}(N,p)$; $N \in \mathbb{N}$ $p \in (0,1)$	$\{1, \cdots, N+1\}$	$\dfrac{N! p^{s-1}(1-p)^{N-s+1}}{(s-1)!(N-s+1)!}$	$\displaystyle\sum_{k=\lceil n/(N+1)\rceil}^{n} \binom{Nk}{n-k} p^{n-k}(1-p)^{Nk-n+k}$
$\text{Bin}^+(N,p)$; $N \in \mathbb{N}$ (fixed) $p \in (0,1)$	$\{1, \cdots, N\}$	$\binom{N}{s}\dfrac{p^s(1-p)^{N-s}}{1-(1-p)^N}$	$\displaystyle\sum_{k=\lceil n/N\rceil}^{n}\sum_{i=1}^{k}\left\{ \binom{k}{i}\binom{Ni}{n}\dfrac{(-1)^{k-i}p^n(1-p)^{Nk-n}}{(1-(1-p)^N)^k} \right\}$
$\text{Poisson}^+(\lambda)$; $\lambda > 0$	$\mathbb{N}$	$\dfrac{\lambda^s e^{-\lambda}}{(1-e^{-\lambda})s!}$	$\displaystyle\sum_{k=1}^{n}\dfrac{k!\lambda^n}{(e^\lambda-1)^k n!}S_2(n,k)$
$\text{NegBin}^+(r,p)$; $r > -1, r \neq 0$ $p \in (0,1)$	$\mathbb{N}$	$\binom{s+r-1}{s}\dfrac{(1-p)^r p^s}{1-(1-p)^r}$	$\displaystyle\sum_{k=1}^{n}\sum_{i=1}^{k}\left\{ \binom{k}{i}\binom{n+ri-1}{n}\dfrac{(-1)^{k-i}p^n}{((1-p)^{-r}-1)^k} \right\}$
$\text{Logarithmic}(p)$; $p \in (0,1)$	$\mathbb{N}$	$\dfrac{-1}{\log(1-p)}\dfrac{p^s}{s}$	$\displaystyle\sum_{k=1}^{n}\dfrac{k!p^n}{n!(-\log(1-p))^k}\lvert S_1(n,k)\rvert$

C.2. Posterior Inference Algorithms

We provide posterior inference algorithms for the ESC models presented in the previous subsection. Let $\boldsymbol{x}$ be data, θ_μ be the parameter(s) of $\boldsymbol{\mu}$, $\boldsymbol{\beta} = (\beta_\ell)$ be the distortion probabilities, and $\mathbf{z} = (z_1, \ldots, z_n)$ be the cluster membership vector. Categorical distribution probabilities $\boldsymbol{\theta}_\ell$ are fixed and assumed to be the empirical distribution of the data, following Steorts (2015). Each iteration of MCMC sampler consists of three steps: first update $\Pi_n \mid \theta_\mu, \boldsymbol{\beta}, \boldsymbol{x}$, next update $\theta_\mu \mid \Pi_n, \boldsymbol{x}$, and then update $\boldsymbol{\beta} \mid \Pi_n, \theta_\mu, \boldsymbol{x}$. We provide details of the first two which are affected by the modification of random partition prior; the likelihood function and updating scheme of $\boldsymbol{\beta}$ are similar to Steorts (2015); Betancourt et al. (2020) and hence omitted.

1. Update $\Pi_n \mid \theta_\mu, \boldsymbol{\beta}, \boldsymbol{x}$ using the following full conditional distribution:

$$\mathbb{P}(z_i = j \mid -) \propto p(\boldsymbol{x} \mid \mathbf{z}_{-i}, z_i = j, \boldsymbol{\beta})\mathbb{P}(z_i = j \mid \mathbf{z}_{-i}, \theta_\mu),$$

where the prior reallocation probability $\mathbb{P}(z_i = j \mid \mathbf{z}_{-i}, \theta_\mu)$ is

- (Shifted binomial) $\quad \mathbb{P}(z_i = j \mid \mathbf{z}_{-i}, N, p) \propto \begin{cases} \dfrac{n_j^{-i}+1}{n_j^{-i}}(N - n_j^{-i} + 1) & \text{if } j = 1, \ldots, k^{-i}, \\ \left(k^{-i}+1\right)(1-p)^{N+1}/p & \text{if } j = k^{-i}+1 \end{cases}$

- (Zero-truncated binomial) $\quad \mathbb{P}(z_i = j \mid \mathbf{z}_{-i}, N, p) \propto \begin{cases} (N - n_j^{-i}) & \text{if } j = 1, \ldots, k^{-i}, \\ \left(k^{-i}+1\right)\dfrac{N(1-p)^N}{1-(1-p)^N} & \text{if } j = k^{-i}+1 \end{cases}$

- (Zero-truncated Poisson) $\quad \mathbb{P}(z_i = j \mid \mathbf{z}_{-i}, \lambda) \propto \begin{cases} 1 & \text{if } j = 1, \ldots, k^{-i}, \\ \left(k^{-i}+1\right)\dfrac{e^{-\lambda}}{1-e^{-\lambda}} & \text{if } j = k^{-i}+1 \end{cases}$

- (Zero-truncated negative binomial) $\quad \mathbb{P}(z_i = j \mid \mathbf{z}_{-i}, r, p) \propto \begin{cases} n_j^{-i} + r & \text{if } j = 1, \ldots, k^{-i}, \\ \left(k^{-i}+1\right)\dfrac{r(1-p)^r}{1-(1-p)^r} & \text{if } j = k^{-i}+1 \end{cases}$

- (Logarithmic) $\quad \mathbb{P}(z_i = j \mid \mathbf{z}_{-i}, p) \propto \begin{cases} n_j^{-i} & \text{if } j = 1, \ldots, k^{-i}, \\ \left(k^{-i}+1\right)\dfrac{-1}{\log(1-p)} & \text{if } j = k^{-i}+1 \end{cases}$

where k^{-i} and n_j^{-i} are the number and sizes of clusters in $\mathbf{z}_{-i}$. Notice that in the last four examples, the reallocation probability to an existing cluster is a linear function of its size, and these are the only possible cases of $\boldsymbol{\mu}$ that have such a property (Johnson et al., 2005). These are called Sundt and Jewell family (Sundt & Jewell, 1981) or (a,b,1) zero-truncated family (Klugman et al., 2012) in actuarial applications.

2. Update $\theta_\mu | \Pi_n, \boldsymbol{x}$, for example using slice sampler (Neal, 2003),

- (Shifted binomial) Here $\theta_\mu = (N, p)$ and let prior $\pi(N, p) \propto p^{-0.5}(1-p)^{-0.5}/N$ following Berger et al. (2012). Then unlike the zero-truncated binomial, we can use the following joint distribution sampler: first update N from $[N|\Pi_n, \boldsymbol{x}]$,

$$[N|\Pi_n, \boldsymbol{x}] \propto \frac{\mathcal{B}(n - k + \frac{1}{2}, Nk - n + k + \frac{1}{2})}{N} \prod_{j=1}^{k} \frac{n_j N!}{(N - n_j + 1)!}, \quad N = \max n_j - 1, \max n_j, \cdots$$

  where $\mathcal{B}$ is a beta function. Given this N, update $p|N, \Pi_n, \boldsymbol{x} \sim \text{Beta}(n - k + \frac{1}{2}, Nk - n + k + \frac{1}{2})$.

- (Zero-truncated binomial with fixed N) Here $\theta_\mu = p$ and let prior $p \sim \text{Beta}(a_p, b_p)$. Then sample p from

$$[p|\Pi_n, \boldsymbol{x}] \propto p^{a_p - 1}(1-p)^{b_p - 1} \frac{p^n (1-p)^{Nk-n}}{(1 - (1-p)^N)^k} \tag{18}$$

- (Zero-truncated Poisson) Here $\theta_\mu = \lambda$ and let prior $\lambda \sim \text{Gamma}(a_\lambda, b_\lambda)$. Then sample λ from

$$[\lambda|\Pi_n, \boldsymbol{x}] \propto \lambda^{a_\lambda - 1} \exp(-b_\lambda \lambda) \lambda^n \left(\frac{e^{-\lambda}}{1 - e^{-\lambda}}\right)^k \tag{19}$$

- (Zero-truncated negative binomial) Here $\theta_\mu = (r, p)$ and let prior $(r, p) \sim \pi(r, p)$. Then sample (r, p) from

$$[r, p|\Pi_n, \boldsymbol{x}] \propto \pi(r, p) p^n \left(\frac{(1-p)^r}{1 - (1-p)^r}\right)^k \prod_{j=1}^{k} [(n_j + r - 1) \cdots (r)] \tag{20}$$

- (Logarithmic) Here $\theta_\mu = p$ and let prior $p \sim \text{Beta}(a_p, b_p)$. Then sample p from

$$[p|\Pi_n, \boldsymbol{x}] \propto p^{a_p - 1}(1-p)^{b_p - 1} p^n \left(\frac{-1}{\log(1-p)}\right)^k \tag{21}$$

The software for the described posterior inference algorithms will be available in R package `microclustr` (Steorts et al., 2020).

D. Details on Real and Synthetic Data Studies

D.1. SIPP1000 Data Analysis Details

In this section, we provide more details on SIPP1000 data analysis described in 4.3. SIPP1000 dataset has $n = 4116$ records, and it contains 5 categorical fields (sex, birth year, birth month, race, and state of residence) of $K^+ = 1000$ individuals. We fit ESC models with four different choices of $\boldsymbol{\mu}$: i) Binomial$^+$ with fixed number of trials $N = 5$ (balance-seeking), ii) Poisson$^+$ (balance-neutral), iii) NegativeBinomial$^+$ (balance-averse), and iv) Dirichlet (neither balance-seeking nor averse; Betancourt et al., 2020). We fix the number of trials as $N = 5$ in the binomial case since dataset is constructed from five waves of longitudinal study, so there is no cluster with size greater than 5 a priori. The hyperprior specifications are i) $p \sim \text{Beta}(0.5, 0.5)$ for the binomial success probability, ii) $\lambda \sim \text{Gamma}(1, 1)$ for Poisson$^+$, iii) $(r, p) \sim \text{Gamma}(1, 1) \times \text{Beta}(2, 2)$ for NegBin$^+$, and iv) $\alpha = 1$, $\boldsymbol{\mu}^{(0)}|r, p = \text{NegBin}^+(r, p)$, and $(r, p) \sim \text{Gamma}(1, 1) \times \text{Beta}(2, 2)$ for Dirichlet (see Betancourt et al. (2020) for details). Prior for distortion parameters (β_ℓ) are assumed to be independent beta distributions with mean 0.005 and standard deviation 0.01.

We collect 15000 posterior samples after 5000 burn-in iterations, where we update cluster indicators (z_i) for each individual within each one (global) MCMC iteration. All computations were performed on an Intel E5-2690 v3 CPU with 128GB of
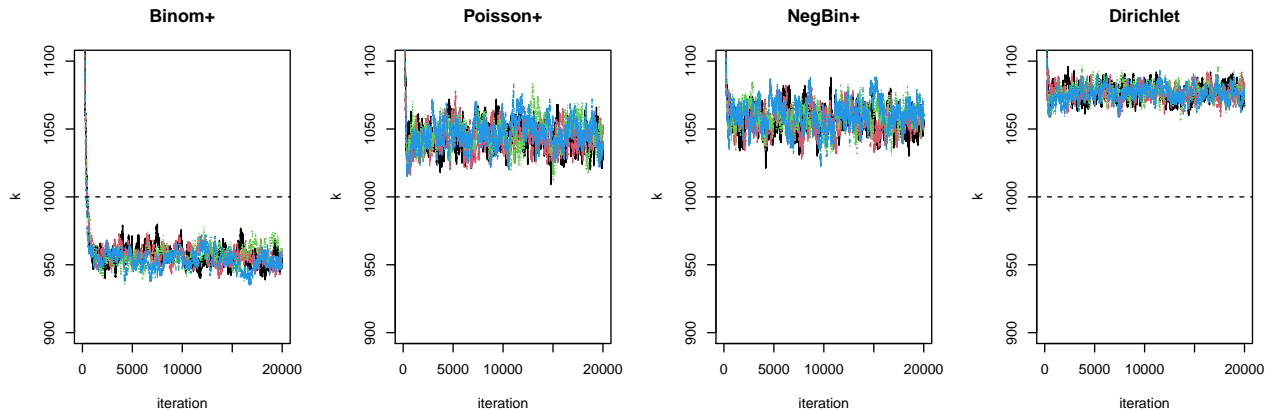
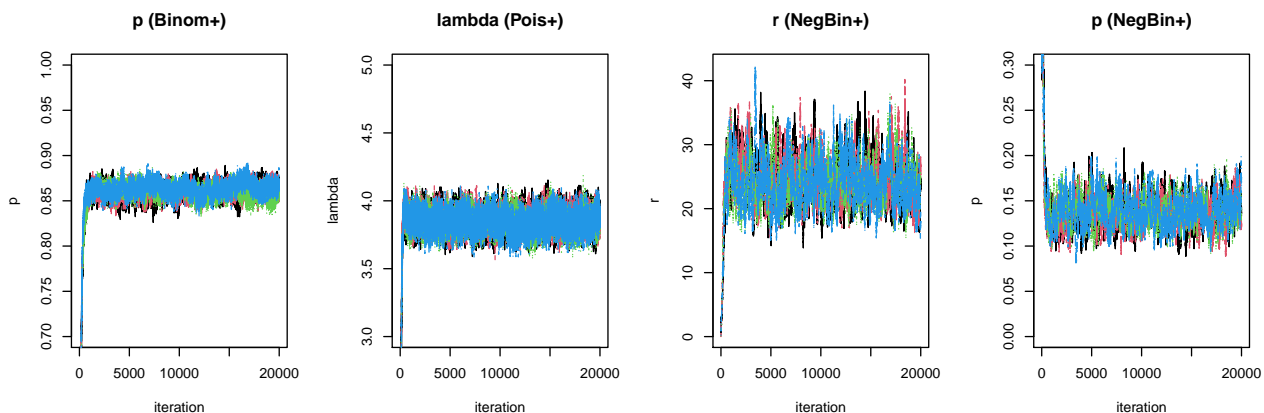*Figure 6.* SIPP1000 dataset; trace plots of the number of clusters (entities) K^+ where true is 1000.

Figure 7. SIPP1000 dataset; trace plots of random partition model-specific parameters θ_μ.

memory. In Figure 6 and Figure 7, present trace plots for the number of entities K^+ as well as parameters θ_μ, where four chains all become stationary after 5000 burn-in iterations.

In addition to Table 3, we also report the summary of posterior distributions of distortion probabilities (β_ℓ) of SIPP1000 dataset. The distortion probability of the month of birth, β_3, is the highest for all models. Balance-seeking model (Binom$^+$) provides higher estimated distortion probabilities compared to other models, even though the prior for (β_ℓ) are all the same. This is because the balance-seeking model avoids the generation of singleton clusters and prefers to treat the difference as 'distorted entry'. This leads to the reduced false negative rate (FNR) by reducing the number of missed pairs (i.e. record pairs that are linked under the truth but not linked under the estimate), but at the cost of increased FDR.

Table 5. SIPP1000 dataset; Summary of posterior distributions (mean, standard deviation) of distortion probabilities (β_l) for different $\boldsymbol{\mu}$. $\beta_1, \ldots, \beta_5$ correspond to distortion probabilities of sex, year of birth, month of birth, race, and state of residence, respectively.

$\boldsymbol{\mu}$ (balance-)	$\mathbb{E}(\beta_1\|\boldsymbol{x})$	SD	$\mathbb{E}(\beta_2\|\boldsymbol{x})$	SD	$\mathbb{E}(\beta_3\|\boldsymbol{x})$	SD	$\mathbb{E}(\beta_4\|\boldsymbol{x})$	SD	$\mathbb{E}(\beta_5\|\boldsymbol{x})$	SD
Binom$^+$ (seeking)	0.0012	0.0014	0.0151	0.0027	0.0492	0.0039	0.0074	0.0029	0.0140	0.0025
Poisson$^+$ (neutral)	0.0003	0.0005	0.0032	0.0018	0.0401	0.0035	0.0008	0.0012	0.0039	0.0019
NegBin$^+$ (averse)	0.0003	0.0006	0.0026	0.0019	0.0391	0.0036	0.0007	0.0010	0.0025	0.0019
Dirichlet (n/a)	0.0002	0.0004	0.0002	0.0004	0.0377	0.0034	0.0016	0.0016	0.0009	0.0008

D.2. Simulation Studies

We conduct additional simulations to study the effect of balancedness on entity resolution tasks. We consider three different scenarios of cluster size distribution: 1) binomial, 2) Poisson, and 3) negative binomial, where Table 6 shows the cluster size distributions for each scenario. We simulate data with the generating scheme described in Section 4.1, where $L = 5$, $D_\ell = 10$, $\boldsymbol{\theta}_\ell = (0.1, \ldots, 0.1)$, and three different choices of distortion probabilities $\beta \in \{0.01, 0.05, 0.1\}$.

Table 6. The true number of clusters $m_s = \sum_j 1(n_j = s)$ of each size $s = 1, \ldots, 14$ for three different scenarios. Each entries correspond to 100 times the probability mass of the distribution shown in the first column, rounded to the nearest integer.

Scenario	m_1	m_2	m_3	m_4	m_5	m_6	m_7	m_8	m_9	m_{10}	m_{11}	m_{12}	m_{13}	m_{14}	K
1. Binom$^+(10, 0.5)$	1	4	12	21	25	21	12	4	1	0	0	0	0	0	101
2. Poisson$^+(5)$	3	8	14	18	18	15	11	7	4	2	1	0	0	0	101
3. NegBin$^+(5, 0.5)$	8	12	14	14	13	11	8	6	5	3	2	1	1	1	99

We fit the ESC model with four different $\boldsymbol{\mu}$: i) $1 +$ Binom (shifted binomial), ii) Poisson$^+$, iii) NegBin$^+$, and iv) Dirichlet. We use shifted binomial instead of zero-truncated binomial, assuming the situation when the maximum size of clusters is not known a priori. For shifted binomial distribution, we put $p(N, p) \propto 1/N \times p^{-0.5}(1 - p)^{-0.5}$ prior on (N, p) following Berger et al. (2012), and other hyperparameter settings for other three models and MCMC specification details are same as the previous subsection in the SIPP1000 data analysis. Here distortion probabilities (β_ℓ) are assumed to be unknown.

Table 7. Simulation study results for three different scenarios and three different true distortion probabilities. Presented are the posterior mean and standard deviation of the number of clusters, and the point estimates of FNRs and FDRs in % by Dahl (2006)'s method.

β	$\boldsymbol{\mu}$ (balance-)	Scenario 1 (true $K^+ = 101$)			Scenario 2 (true $K^+ = 101$)			Scenario 3 (true $K^+ = 99$)		
		$\mathbb{E}(K^+\|\boldsymbol{x})$ (SD)	FNR	FDR	$\mathbb{E}(K^+\|\boldsymbol{x})$ (SD)	FNR	FDR	$\mathbb{E}(K^+\|\boldsymbol{x})$ (SD)	FNR	FDR
0.01	$1 +$ Binom (seeking)	100.8 (0.6)	**0.0**	**0.0**	100.6 (1.1)	**0.9**	0.8	98.9 (1.3)	0.6	0.1
	Poisson$^+$ (neutral)	101.3 (0.7)	0.4	0.5	101.5 (1.1)	1.1	**0.6**	100.1 (1.1)	0.6	0.1
	NegBin$^+$ (averse)	101.6 (0.9)	0.4	0.5	102.2 (1.2)	1.1	**0.6**	101.3 (1.4)	0.6	0.1
	Dirichlet (n/a)	101.7 (0.9)	0.7	**0.0**	102.1 (1.5)	1.5	0.8	103.6 (2.0)	0.6	0.1
0.05	$1 +$ Binom (seeking)	99.7 (1.2)	**2.9**	3.2	100.0 (1.5)	**2.4**	2.9	95.2 (1.5)	**3.0**	3.5
	Poisson$^+$ (neutral)	103.8 (2.1)	4.2	**2.4**	102.4 (2.0)	2.9	2.6	96.4 (1.7)	3.6	4.4
	NegBin$^+$ (averse)	106.4 (2.2)	6.7	3.6	105.2 (2.4)	3.2	2.6	99.4 (2.3)	3.1	**2.8**
	Dirichlet (n/a)	102.1 (2.6)	4.8	4.1	103.6 (3.3)	3.8	**2.0**	100.6 (3.0)	4.9	4.0
0.1	$1 +$ Binom (seeking)	105.2 (1.9)	**8.1**	4.1	98.5 (2.2)	**7.0**	9.4	94.0 (2.2)	8.1	7.5
	Poisson$^+$ (neutral)	111.1 (2.9)	9.2	4.1	102.9 (2.9)	8.9	10.1	98.7 (3.0)	**7.3**	8.1
	NegBin$^+$ (averse)	115.6 (3.6)	10.5	**3.9**	109.5 (4.0)	10.7	8.8	109.1 (4.3)	10.9	**5.4**
	Dirichlet (n/a)	109.9 (5.0)	8.8	4.1	113.9 (5.5)	12.9	**6.5**	112.0 (5.0)	12.7	6.2

Table 7 shows that the balance-seeking random partition model generally achieves a lower FNR compared to others, even under Scenarios 2 and 3 where the model is misspecified. Figure 8 shows the posterior distributions of m_i, the number of clusters of size i. It can be clearly seen that the balance-seeking model regularizes the generation of singleton clusters, suggesting its usefulness for ER tasks when a single record per individual is less likely to happen or when false negative control is more important.
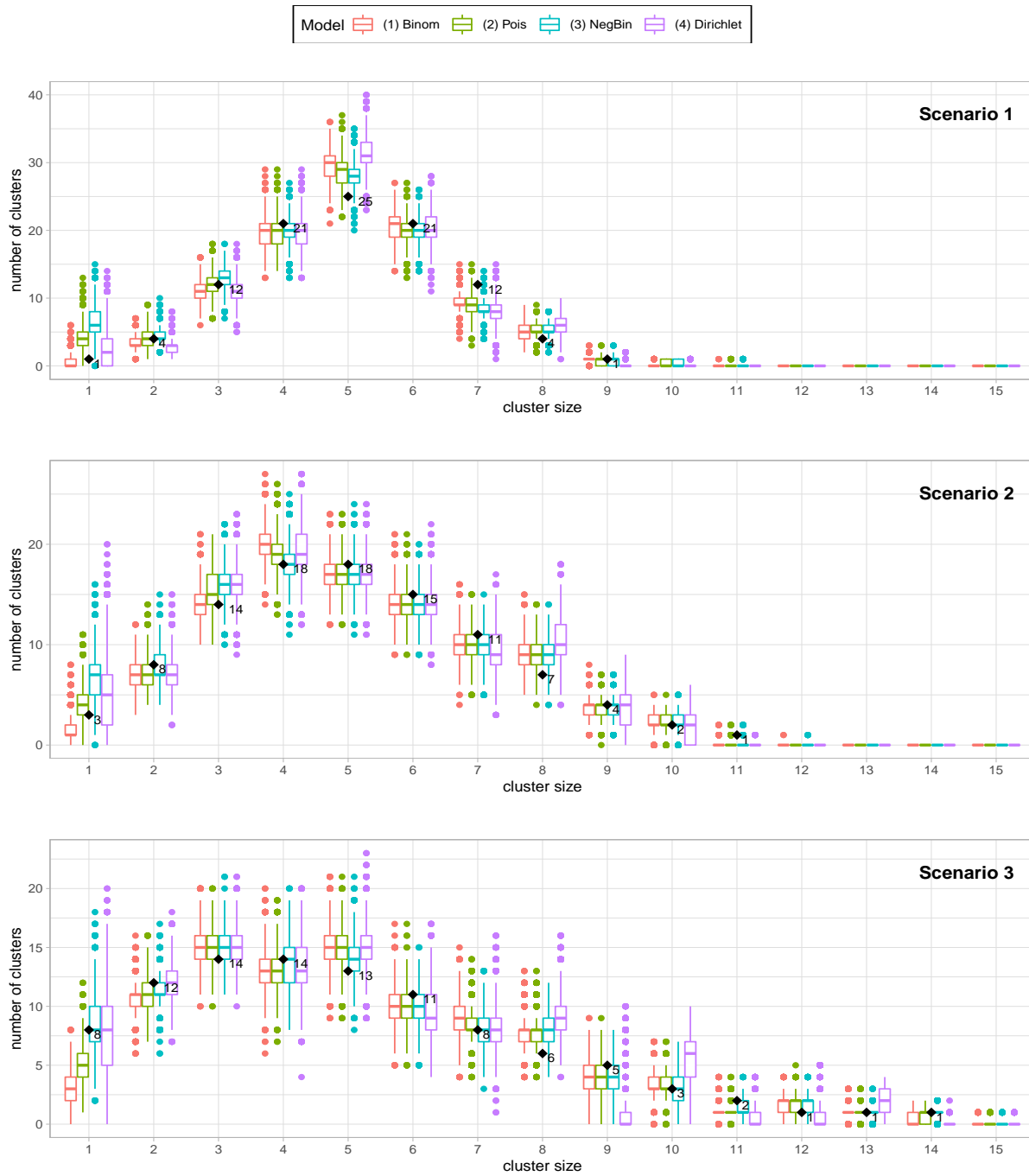
Figure 8. Posterior boxplots of the number of clusters of size i for each models, m_i. (Top) Scenario 1 with $\beta_{true} = 0.05$. (Middle) Scenario 2 with $\beta_{true} = 0.05$. (Bottom) Scenario 3 with $\beta_{true} = 0.05$. True m_i are shown as black diamonds and annotated by their values.