Permuton-induced Chinese Restaurant Process

Anonymous Author(s)
Affiliation
Address
email

Abstract

This paper proposes a multi-dimensional extension of the Chinese restaurant process (CRP) and applies it to Bayesian nonparametric (BNP) relational data analysis. Specifically, we introduce a model in which the tables of CRP have random geometric arrangements on $[0, 1] \times [0, 1]$ induced by a *permuton*. Permuton is a probability measure on $[0, 1] \times [0, 1]$, which can be regarded as a geometric interpretation of the scaling limit of permutations. Moreover, several important classes of permutations are in one-to-one correspondence with various important classes of rectangular partitioning. As a result, our proposed model named permuton-induced CRP (PCRP) can be used as a unified stochastic process to represent various classes of rectangular partitioning. By analogy with the relationship between the stick-breaking process (SBP) and CRP for the infinite mixture model of a sequence, this model can be regarded as a multi-dimensional extension of CRP paired with the block-breaking process (BBP), which has been recently proposed as a multi-dimensional extension of SBP. While BBP always has an infinite number of redundant intermediate variables, PCRP can be composed of varying size intermediate variables in a data-driven manner depending on the size and quality of the observation data. Experiments show that PCRP can improve the prediction performance in relational data analysis by reducing the local optima and slow mixing problems compared with the conventional BNP models because the local transitions of PCRP in Markov chain Monte Carlo inference are more flexible than the previous models.

1 Introduction

The multi-dimensional extension of the Chinese restaurant process (CRP) [9, 51] for infinitely exchangeable random arrays is one of the most critical unsolved issues in the field of Bayesian nonparametrics and its applications to machine learning. The standard CRP is the stochastic process representing random clustering of a sequence using the metaphor of customers sitting at tables. Typically, it is often used when we need to cluster a sequence where the number of clusters is unknown in advance. The magic of CRP is that it requires only a finite number of parameters for finite data; nevertheless, the model itself has infinite representational power. Specifically, it is known that a mixture model using CRP is theoretically equivalent to the Dirichlet process mixture model (DPMM) [52, 28] represented by the stick-breaking process (SBP) [57] with an infinite number of parameters. The relationship between the two DPMM representations, CRP and SBP, can be explained by whether the intermediate random variables in de Finetti's representation theorem [29, 38] are marginalized out or not (Figure 1, (a) and (b)). CRP has contributed greatly to the development of the BNP machine learning field in the sense that it has made BNP not only theoretical but also practical. It enables models with infinite representational power to be accurately simulated on a computer with a finite number of parameters without any approximation (e.g., finite truncation). In fact, a variety of extensions of CRP have been studied, including the nested CRP [17, 18, 61], the Chinese restaurant franchise [62], the tree-structured CRP [6], the distance-dependent CRP [15, 16, 32]. However, in the development of CRP, the extension to BNP relational models has remained a historical conundrum.

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

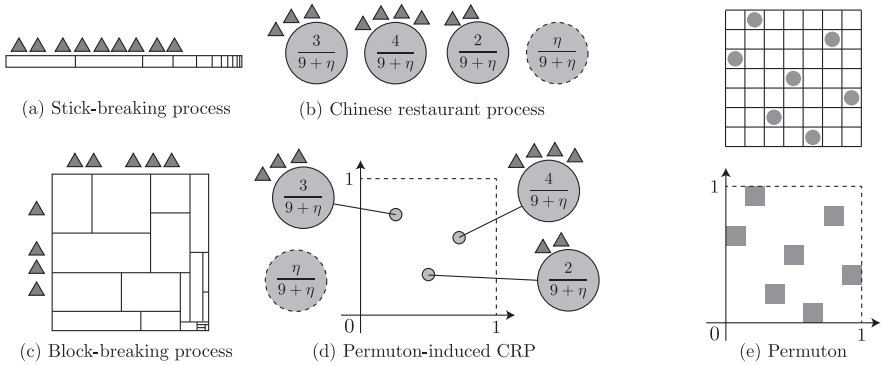


Figure 1: (a) SBP representation for sequence partitioning based on de Finetti's representation. (b) CRP representation for sequence partitioning. Each customer (triangle) sequentially choose a table (dark circle). (c) BBP representation for rectangular partitioning based on AHK representation theorem. (d) Our proposed PCRP. We recall that a certain class of rectangular partitioning has a bijection to a certain class of permutation. Therefore, if we could add to the order based on permutations for CRP tables, we could convert the table assignments of data by CRP into rectangular partitioning. To achieve this, we introduce a random geometric location on $[0, 1] \times [0, 1]$ into the CRP table drawn from permuton. (e) Illustration of permuton. Top shows an array representation of a permutation $\sigma = 5724163$. The horizontally i-th circle has the height $\sigma(i)$. Bottom shows the corresponding permuton to $\sigma = 5724163$, i.e., the probability measure on $[0, 1] \times [0, 1]$.

As a special case of the BNP relational models, the infinite relational model (IRM) [39] consisting of the product of CRPs was proposed. The IRM can express random rectangular partitioning of a matrix; however, its representational capability is restricted to a minimal subset of rectangular partitioning called *regular grid* (Figure 2, left). Therefore, in order to obtain a model with higher representational capability, a variety of BNP models for relational data have been rapidly developed in recent years [39, 56, 55, 21, 54, 58, 46, 37, 20, 7, 24, 27, 43, 44, 31, 22, 26, 25]. Recent excellent surveys can be found in [23, 50]. In particular, the Mondrian process (MP) [56, 55] and the rectangular tiling process (RTP) [48] have been proposed as new BNP models to expand the expressive classes of rectangular partitioning. However, both of them require an infinite number of parameters as intermediate random variables in the Aldous-Hoover-Kallenberg (AHK) representation theorem [8, 35, 38] (an extension of de Finetti's theorem to multi-dimensional arrays), similar to the SBP representation for DPMM. As a result, Bayesian inference to those models always had to take care of an infinite number of parameters, which compromise the flexibility of local movements for the Markov chain Monte Carlo (MCMC) inference and make it highly affected by local optima and slow mixing problems. As a natural question, we wonder if it is possible to marginalize out the infinite number of intermediate random variables of these models, as in the CRP representation of DPMM. Unfortunately, an approach to achieve this has remained unsolved for a long time.

This paper deals with the multi-dimensional extension of CRP to represent wider classes of rectangular partitioning than regular grid partitioning. As an important stepping stone for this, the block-breaking process (BBP) [47] has been proposed very recently as a multi-dimensional extension of SBP (Figure 1, (c)). A special class of permutations called *Baxter permutation* [14], which has a one-to-one correspondence with a particular class of rectangular partitioning, is used to extend the stick-breaking procedures of SBP for a sequence to the block-breaking procedures for a matrix (Figure 3). Although BBP can indeed be used as an infinitely exchangeable relational model, it requires an infinite number of redundant parameters, similar to SBP, making Bayesian inference highly affected by local optima and slow mixing problems. Therefore, a CRP-like representation that avoids the infinite number of redundant parameters of BBP, analogous to the relationship between SBP and CRP for DPMM, has been desired. Our contributions are summarized as follows:

Construction of new stochastic process - Section 3.1 proposes a stochastic process such that each table of the CRP has a random coordinates on $[0, 1] \times [0, 1]$. Specifically, we suppose that the coordinates of each table is independent and identically distributed (i.i.d.) random variable drawn from a *permuton* [36] on $[0, 1] \times [0, 1]$. Permuton is a probabilistic measure on $[0, 1] \times [0, 1]$ and can also be regarded as a prior model for random permutations. By exploiting the correspondence between permutations and rectangular partitions [63, 53], PCRP can be used as a generative probabilistic

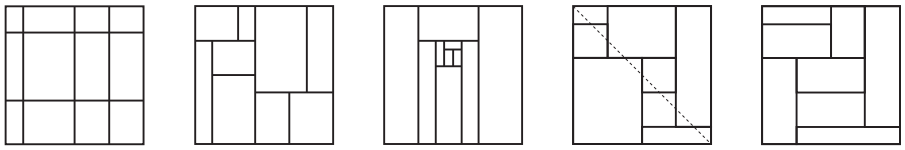


Figure 2: Five classes of rectangular partitioning. The details of each class will also be described in the main body of Section 2.1. **Left:** Regular grid partitioning. The rows and columns are partitioned into clusters. Each block is characterized by the product of the row and column clusters. **Second from left:** Hierarchical partitioning. **Third from left:** Hierarchical partitioning with one place deeper in each layer. **Fourth from left:** Diagonal rectangulation. **Right:** Generic rectangulation.

model of rectangular partitioning of a matrix. While BBP [47] (i.e., a multi-dimensional extension of SBP) always has an infinite number of redundant intermediate variables, our PCRP can be composed of varying size (finite) intermediate variables in a data-driven manner depending on the size and quality of the observation data. As a result, PCRP can lead to very flexible local movements in Bayesian inference, reducing the problems of local optima and slow mixing.

Unified framework for various classes of rectangular partitioning - Conventionally, the stochastic processes for each class of rectangular partitioning have been constructed individually. However, according to the design of the permuton, PCRP provides a unified framework for various classes of rectangular partitioning. Section 3.2 shows an approach to tune the permuton of PCRP to adapt it to various classes of rectangular partitioning.

New representation of BNP model - As mentioned earlier, a multi-dimensional extension of CRP is a historical conundrum, and thus PCRP does not provide a perfect solution to this issue. In Section 4, the theoretical imperfection of PCRP is clarified. Specifically, we intuitively explain that projectivity and exchangeability of PCRP change when viewed from two different perspectives: rectangular partitioning of a matrix and table assignment of data. Thus, we propose a new representation of the BNP model as a way to overcome this imperfection, that applies PCRP as a bridging state to existing infinitely exchangeable relational models, inspired by the MCMC inference via bridging [42].

2 Preliminaries

2.1 Permutation, permutation class, and its relation to rectangular partitioning

The permutation of a set S is defined as a bijection from S to itself. Specifically, this paper always focus on permutations of natural numbers $S := \{1, 2, \ldots, n\}$ $(n \in \mathbb{N})$. For convenience, we use one-line notation for a permutation $\sigma = (\sigma_1 \sigma_2 \ldots \sigma_n)$. We suppose that τ and σ are permutations of size k and n, respectively. An *occurrences* of pattern τ in σ is a subsequence $\sigma_{i_1}, \ldots, \sigma_{i_k}$ that is order-isomorphic to τ, that is, $\sigma_{i_s} < \sigma_{i_t} \Leftrightarrow \tau_s < \tau_t$, and is denoted by $\tau < \sigma$.

In the previous studies, several papers focuses on a subset of permutations as a *permutation class*. To identify a permutation class, the set of patterns which are forbidden to be occurred in the class is called the *antichain*. For any permutation class $\mathcal{C}$, there is a unique antichain B such that $\mathcal{C} = \mathrm{Av}(B) := \{\sigma : \sigma \not\geq \pi \text{ for all } \pi \in B\}$. In other words, we can identify a certain permutation class by specifying the antichain, for example, $\mathrm{Av}(312, 213)$.

Furthermore, we introduce the notion of *dashed patterns* (also called *vincular* patterns) to represent various antichains. A dashed pattern is a permutation τ where each pair of consecutive numbers may or may not have a dash between them, e.g., 312 and 12-3. No dash means that consecutive numbers of τ must appear adjacently. On the other hand, dash means that consecutive numbers may appear arbitrarily far apart. For example, a permutation $\sigma = \mathbf{25}1\mathbf{6}43$ contains a dashed pattern 12-3 and 1-2-3 but avoids 1-23. Using the dashed patterns, we introduce some important classes of permutations, which are particularly closely related to particular classes of rectangular partitioning [10, 49, 60].

2-clumped permutation: $\mathrm{Av}(\text{3-51-24, 3-51-42, 24-51-3, 42-51-3})$. The 2-clumped permutations have a bijection to *generic rectangulations* (Figure 2, right) [53], that no restrictions are required.

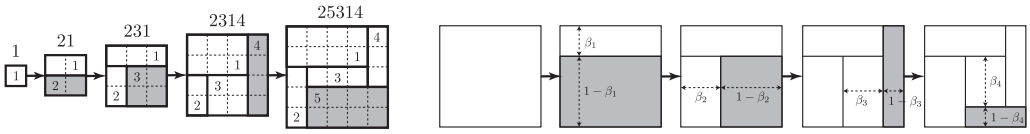


Figure 3: Illustration of BBP [47], which can be regarded as a multi-dimensional extension of SBP. First BBP uses a Markov process on Baxter permutations, and transforms the evolution of Baxter permutations to the evolution of floorplan partitioning (i.e., rectangular partitioning ignoring the size of each rectangle block) by using the one-to-one correspondence between Baxter permutations and floorplan partitioning [34] (**Left**). Next BBP uses the block-breaking procedures to sequentially assign the random size to each rooms floorplan partitioning (**Right**).

Baxter permutation: $\mathrm{Av}(2\text{-}41\text{-}3,\ 3\text{-}14\text{-}2)$ and **twisted Baxter permutation**: $\mathrm{Av}(2\text{-}41\text{-}3,\ 3\text{-}41\text{-}2)$. The Baxter permutations have a bijection to *diagonal rectangulations* (Figure 2, fourth from left) [63, 5, 33], which has a representative in which every rectangle's interior intersects the diagonal.

Separable permutation: $\mathrm{Av}(2413,\ 3142)$. Separable permutations have a bijection to *hierarchical and diagonal rectangulations* [59, 5], a subset of *hierarchical partitioning* (Figure 2, second from left), which are expressed as binary trees where nodes represent a vertical or horizontal separation of a rectangle into two disjoint rectangles.

Separable skew-merged permutation: $\mathrm{Av}(2143,\ 2413,\ 3142,\ 3412)$. Separable skew-merged permutations have a bijection to a special subset of hierarchical partitioning, in particular, those where only one rectangle is allowed to be further cut in each layer (Figure 2, third from left) [45].

2.2 Permuton

In this paper, we introduce *permuton* [36] as a very useful tool for handling random permutations. Before defining permuton, we begin with a geometric interpretation of a permutation (Figure 1, (e) top). To a permutation σ, we can associate a probability measure γ_σ on $[0,1]\times[0,1]$ as follows. We first divide $[0,1]\times[0,1]$ into an $n\times n$ grid of squares of size $1/n\times 1/n$. Then, we can define the density of γ_σ on the square in the i-th row and j-th column to be the constant n if $\sigma_i = j$ and 0 otherwise. Therefore, γ_σ can be regarded as a geometric representation of the permutation matrix of σ. More generally, we define a *permuton* to be a probability measure γ on $[0,1]\times[0,1]$ with uniform marginals (Figure 1, (e) bottom): $\gamma([a,b]\times[0,1]) = b-a = \gamma([0,1]\times[a,b])$ for all $0\le a\le b\le 1$. We emphasize that, for any permutations σ, the previous probability measures γ_σ is a permuton. Specifically, the permutation σ can be interpreted as the permuton γ_σ given by the sum of Lebesgue area measures [36]: $\gamma_\sigma(A) = n\sum_{i=1}^{n}\mathrm{Leb}\big([(i-1)/n, i/n]\times[(\sigma(i)-1)/n, \sigma(i)/n]\cap A\big)$, for all Borel measurable set A of $[0,1]\times[0,1]$, where $\mathrm{Leb}(\cdot)$ indicates the Lebesgue measure.

We will provide the intuition behind why permuton is useful in dealing with random permutations. We suppose that τ and σ are permutations of size k and n, respectively. We set $\widetilde{\mathrm{occ}}(\tau,\sigma) := (n!/k!(n-k)!)\cdot\#\{\text{occurrences of } \tau \text{ in } \sigma\}$. Intuitively, $\widetilde{\mathrm{occ}}(\tau,\sigma)$ means the probability to find a pattern τ in σ, when we take k elements uniformly at random in σ. Similarly, if we replace permutation σ with permuton γ, we can consider the probability of occurrence of a pattern τ: $\widetilde{\mathrm{occ}}(\tau,\gamma) := \mathbb{P}[u_1,\ldots,u_k \text{ form a pattern } \tau]$, where $u_1,\ldots,u_k$ are i.i.d. points on $[0,1]\times[0,1]$ drawn from the permuton γ. Now, a natural question is whether these two $\widetilde{\mathrm{occ}}(\tau,\sigma)$ and $\widetilde{\mathrm{occ}}(\tau,\gamma)$ are generally consistent. The following theorem gives us an answer to this question:

Theorem 2.1 *(See, e.g., Theorem* 2.5 *[30] and Lemma* 3.5 *[36]) We consider a sequence of permutations $\sigma_1,\sigma_2,\ldots,\sigma_n,\ldots$ of size tending to infinity $n\to\infty$. Then, we have $\gamma_{\sigma_n}\to\gamma \iff$* for every τ, $\widetilde{\mathrm{occ}}(\tau,\sigma_n)\to\widetilde{\mathrm{occ}}(\tau,\gamma)$. *Moreover, if τ and σ are permutations of size k and n, respectively, then $|\widetilde{\mathrm{occ}}(\tau,\sigma)-\widetilde{\mathrm{occ}}(\tau,\gamma_\sigma)|\le k(k-1)/2n$.*

As a result, we can see that random permutations belonging to some class can be controlled by the limit of permuton.

3 Permuton-induced Chinese restaurant process (PCRP)

This section proposes a new stochastic process, which is our main contribution to this paper. First, we describe the stochastic process, PCRP, in terms of a generative model and show how we can introduce

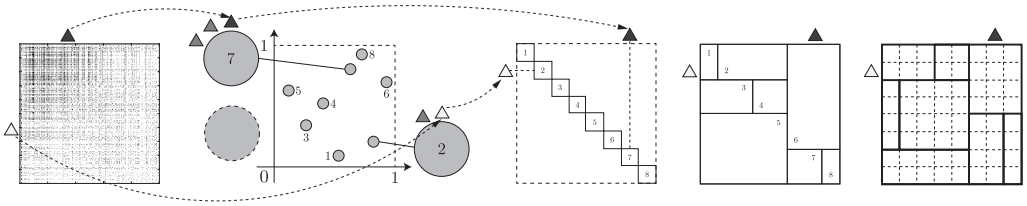


Figure 4: Overview of permuton-induced CRP as generative probabilistic model.

the notion of the permuton into CRP to obtain a model for rectangular partitioning. Next, we discuss how to design the permuton so that PCRP can represent some classes of rectangular partitioning.

3.1 Model description of PCRP from generative probabilistic point of view

PCRP is composed of an input matrix, the concentration parameter $\eta > 0$, and the permuton γ, i.e., a probability measure on $[0, 1] \times [0, 1]$. The specific design of the permuton will be discussed in the next subsection, thus the reader may continue to read assuming, for example, $\gamma(\cdot) = \mathrm{Leb}\big(\{\cdot\}\big)$, that is, the uniform distribution on $[0, 1] \times [0, 1]$. The following four steps represent the overall picture of PCRP from the viewpoint of a probabilistic generative model.

Step 1: Table assignment process of all row and column elements of the input matrix by a single CRP (Figure 4, first and second from left) - Each row (white triangle) and each column (black triangle) of the input matrix chooses either an existing table to sit at with probability $N_i/(N+\eta)$ or a new table with probability $\eta/(N + \eta)$, as in a standard CRP [9], where N is the total number of customers seated at the table, N_i is the number of customers seated at the i-th table. If a new table is chosen, we draw the coordinates of the table on $[0, 1] \times [0, 1]$ from the permuton γ. Here we emphasize that unlike IRM [39] (i.e., the product of CRPs), PCRP uses only one CRP; IRM generally uses separate CRPs for clustering row elements and column elements, while PCRP sends row elements and column elements to a single CRP table assignment.

Step 2: Numbering the CRP tables and placement of customers on the sides of a rectangle (Figure 4, second and third from left) - The CRP tables are numbered according to the vertical order of the coordinates, which leads to a permutation $\sigma = (\sigma_1, \ldots, \sigma_n)$. Note that we recursively perform this numbering each time we add a new table. Next, we consider a rectangle (referred to as the outer rectangle) and its main diagonal and divide it into n equal-sized line segments numbered $1, \ldots n$, from top left to bottom right. The customer corresponding to each row of the input matrix is moved to the left side position of the outer rectangle corresponding to the table number it belongs to. Similarly, we move the customer corresponding to each column of the input matrix to the top side position of the outer rectangle corresponding to the table number it belongs to.

Step 3: Transforming permutation to diagonal rectangulation (Figure 4, fourth from left) - Given the permutation $\sigma = (\sigma_1, \ldots, \sigma_n)$, we obtain the diagonal rectangulation as follows [53] (See also Figure 5). For the i-th step $(i = 1, \ldots, n)$, we add the i-th rectangle such that it intersects the main diagonal in the σ_i-th line segment. Specifically, considering the stair-stepping nature of the union of the first through i-th rectangles (the lower left boundary is an L-shaped line and the upper right boundary is a non-increasing polygonal line), we add the i-th rectangle so that the i-th rectangle is maximal. It should be emphasized here that every permutation can be mapped to a diagonal rectangulation. More specifically, there is a non-injective surjective map from permutations to diagonal rectangulations [53]. In short, it is guaranteed that for every permutation σ represented by the random CRP table coordinates, σ can indeed be converted to a diagonal rectangulation.

Step 4: Transforming diagonal rectangulation to generic rectangulation (Figure 4, right) - We first introduce two notions, *edge* and *wall* (See also Figure 6, left) in this step. An *edge* of the rectangulation is a line segment contained in the side of some rectangles such that the endpoints of the line segment are vertices and the segment has no vertices in its interior. A *wall* of the rectangulation is a maximal union of edges forming a line segment. We then focus on the vertices on each wall (Figure 6, third from left). If the vertex is in the upper left corner of the rectangle, then we assign a rectangle index to the vertex. Otherwise, we number the vertex with the index of the rectangle whose bottom right is the vertex (Figure 6, fourth from left). Finally, for each wall, if it is horizontal, reorder

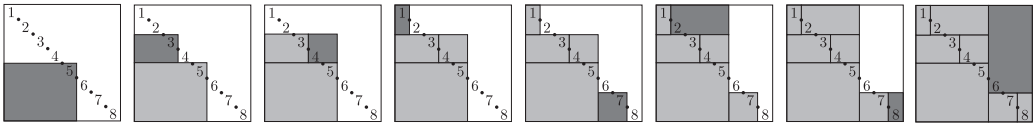


Figure 5: Construction of diagonal rectangulation of $\sigma = 53417286$.

the vertices so that their numbers match the original permutation from left to right. Otherwise (if it is vertical), reorder the vertices so that their numbers match the original permutation from bottom to top (Figure 6, right). For example, Figure 6 shows the case of $\sigma = 53417286$. The horizontal wall of its diagonal rectangulation has the vertex indexes $\mathbf{3}, \mathbf{1}, \mathbf{4}, \mathbf{2}$ from left to right. Then we extract these indexes from $\sigma = 5\mathbf{3}4\mathbf{1}7\mathbf{2}86$, and obtain the order of $\mathbf{3}, \mathbf{4}, \mathbf{1}, \mathbf{2}$. Finally, we reorder the vertices so that their numbers match the order of $\mathbf{3}, \mathbf{4}, \mathbf{1}, \mathbf{2}$. We apply these procedures to all walls, and obtain the generic rectangulation of $\sigma = 53417286$. We also emphasize that there is a non-injective surjective map from permutations to generic rectangulations [53]. As a result, PCRP can be used as a model of rectangular partitioning, regardless of the choice of the permuton.

Intuitive interpretation of PCRP: We recall that PCRP consists of two components: (1) random table assignments to all the row and column elements of the input matrix, and (2) random permutations represented by the table coordinates. The former table assignments serve to roughly group the rows and columns of the input data, and the latter random table coordinates serve to transform those rough groupings into detailed rectangular partitioning. The practical advantage of PCRP is that when a new table of CRPs is added, it can be inserted into every possible place of the rectangular partition (with the consistency that the whole is a rectangular partition). This allows us to eliminate the problem of the disadvantage of the conventional BBP, where the insertion of a new rectangle is restricted to a specific place in the lower right corner (Figure 3).

3.2 Design of permuton to provide various representational capabilities to PCRP

We discuss how to design the permuton γ in a specific way. We start with the simplest case of $\gamma(\cdot) = \mathrm{Leb}\big(\{\cdot\}\big)$, called the *uniform permuton* [36]. As is well known, if σ_n is a uniform random permutation of size n, then γ_{σ_n} converges in distribution to this uniform permuton. We recall here that every permutation can be transformed into some rectangular partition. Thus it seems that PCRP with the uniform permuton can be used as a probabilistic model for arbitrary rectangular partitioning. However, there is one crucial point to note: The mapping between permutations (a set of all permutations) and generic rectangulations is not a one-to-one correspondence. This means that even if the random permutations are uniform, the corresponding random rectangular partitions they are transformed into are not uniform and have some bias. Namely, there are some biases in PCRP with the uniform permuton, such that some rectangular partitions are more likely to appear and some are less likely to appear. We could ignore this bias for practical use, but theoretically, it is essential to know how to eliminate it. We briefly introduce some permuton designs that are uniform for some class of rectangular partitioning. Here is a brief list, and details are given in the supplementary material.

Hierarchical partitioning with one place deeper in each layer (Figure 2, third from left) - Uniform partitioning of this class can be obtained by restricting the random permutation of coordinates in the CRP tables to uniform separable skew-merged. If σ_n is a uniform separable skew-merged permutation of size n, then γ_{σ_n} converges in distribution to a deterministic permuton γ (See, e.g., Theorem 3.3 and Section 3.2.1 [13]): Intuitively, the permuton γ has an X-shaped density on $[0, 1] \times [0, 1]$.

Hierarchical and diagonal rectangulation (Figure 2, second and fourth from left) - Uniform hierarchical and diagonal rectangulations can be expressed by uniform separable permutations. If σ_n is a uniform separable permutation of size n, then γ_{σ_n} converges in distribution to the *Brownian separable permuton* [12, 11, 45].

Diagonal rectangulation (Figure 2, fourth from left) - Uniform diagonal rectangulations can be expressed by uniform Baxter permutations. If σ_n is a uniform Baxter permutation of size n, then γ_{σ_n} converges in distribution to the *Baxter permuton* [19].

Generic rectangulation (Figure 2, right) - To the best of our knowledge, there is still no way to construct a permuton that corresponds to 2-clumped permutations, which have a bijection to generic rectangulations. This means that it is currently tricky to explicitly design a clever permuton that will

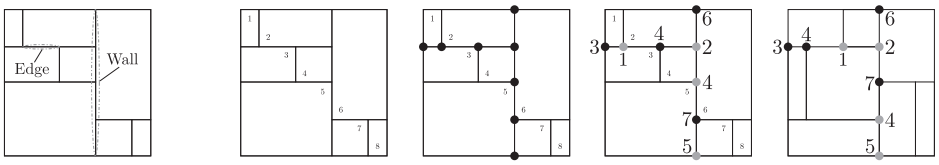


Figure 6: Transformation from diagonal rectangulation of $\sigma = 53417286$ to generic rectangulation.

always restrict its samples to 2-clumped permutations. One possible strategy is to generate a very large random permutation in advance and use it as an approximation to the limit of the permuton, as suggested by Theorem 2.1. However, this strategy is contrary to the original aim of CRP, which is to generate as many intermediate parameters as necessary depending on the input data size. Therefore, another strategy is to relax the permuton itself to one that can generate a broad class of permutations (e.g., $\gamma(\cdot) = \mathrm{Leb}\big(\{\cdot\}\big)$), and instead use *rejection sampling* to restrict the generated permutations so that they are actually 2-clumped. As a result, this method only requires parameters for the random coordinates of the table that are needed by the CRP. Details are given in the supplementary material.

4 Intermediate level representation between SBP and CRP

PCRP would be very useful in the practical use of relational data analysis due to the following two advantages. (1) The model itself adjusts the model complexity in a data-driven manner according to the quality and size of the input data. (2) The design of permuton allows us to handle various classes of rectangular partitioning in a unified manner. In this section, we explore the former point more deeply from the viewpoint of the theory of Bayesian nonparametrics. A precise explanation would require the notions of *projective system*, *projectivity*, and *exchangeability*, which are far from the scope of this paper. Therefore, we would like to focus on an intuitive explanation while providing a rigorous discussion in the supplementary material.

Motivation - As well as the standard CRP for sequence partitioning being an infinitely exchangeable model, PCRP also has projectivity and exchangeability in terms of the CRP table assignments of data. However, from the perspective of rectangular partitioning model in relational data analysis, the projectivity[1] and exchangeability of PCRP, unfortunately, do not hold. The implications of this fact will be explained in terms of the practical intuition for relational data analysis. We consider a pair of matrices $I \subset J \subset \{1, 2, \ldots, \} \times \{1, 2, \ldots, \}$. Namely, a smaller matrix I is an extraction of some rows and columns of a larger matrix J. As it turns out, the only theoretical imperfection of PCRP can be explained as follows: the random rectangular partitioning of the smaller matrix I that follows PCRP cannot be equivalent to the random rectangular partitioning of I when the random rectangular partitioning of the larger matrix J follows PCRP. In other words, PCRP does not have perfect self-similarity according to the size of the data in terms of the rectangular partitioning model, but only in the imaginary object of the CRP table assignments. This implies that PCRP may have to tune its model complexity (e.g., the concentration parameter of CRP) according to the size of the data. Nevertheless, in a practical sense, we believe that this is not such a critical issue. Rather, local optima and slow mixing problems in Bayesian inference have a more critical impact on the performance of the model. However, readers familiar with the theory of Bayesian nonparametrics may wish to have a BNP model that is strictly projective and exchangeable in terms of rectangular partitioning. For such readers, we also propose a new way to create an exact infinite exchangeable model while using PCRP. Specifically, we apply PCRP as a bridging state to existing infinitely exchangeable relational models, inspired by the MCMC inference via bridging [42].

We are interested in a set of rectangular partitioning of a matrix $I \subset \{1, 2, \ldots, \} \times \{1, 2, \ldots, \}$ whose rows and columns are indexed by natural numbers. We consider probability measures μ_X^I on $(\mathcal{X}_I, \mathbf{2}^{\mathcal{X}_I})$ where $\mathcal{X}_I$ is a set of rectangular partitioning of a matrix I (i.e., a constrained combinatorial space), and hope to find an exact BNP model (Figure 7, left). Needless to say, a number of the various BNP models have actually been proposed as the exact BNP model μ_X^I, including the IRM [39], the MP-based relational model [56], and the BBP-based relational model [47]. However, conventionally, these exact BNP models have significantly struggled with local optima and slow mixing problems

[1]This is because the choice of the *projector* of the projective system is strongly related to the success or failure of the projectivity of the BNP model.

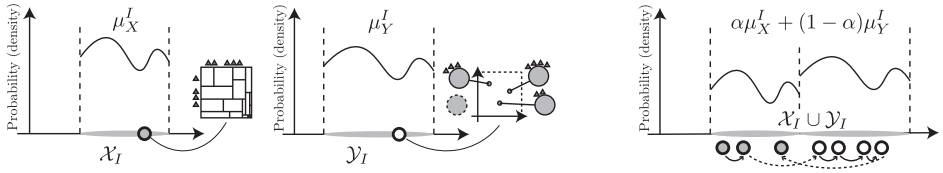


Figure 7: Illustration of exact BNP model with bridging states. **Left:** Existing exact BNP model on rectangular partitioning $\mathcal{X}_I$. **Middle:** PCRP on imaginary table assignments of data $\mathcal{Y}_I$. **Right:** BNP model on extended space $\mathcal{X}_I \cup \mathcal{Y}_I$. Sampling from μ_X^I is equivalent to drawing samples (gray circles) from $\mathcal{X}_I \cup \mathcal{Y}_I$ via the joint chain and discarding those (white circles) from $\mathcal{Y}_I$.

in Bayesian inference. For example, in the standard MCMC inference phase, we wish to sample from the distribution μ_X^I over a constrained combinatorial space $\mathcal{X}_I$. Typically, by using some local moves, we can derive a Markov chain with a transition matrix, which may have slow mixing or even be non-ergodic. To mitigate these issues, we propose to introduce auxiliary spaces to relax the local movements in combinatorial spaces and to connect different regions of the sample space where communication is difficult or impossible.

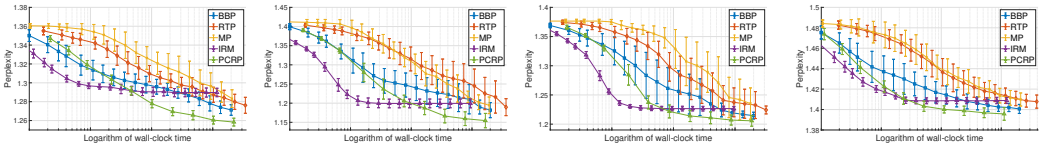
Exact BNP model with bridging states - First, as the auxiliary space $\mathcal{Y}_I$, we introduce the table assignments of I by CRP and a set of coordinates on $[0,1] \times [0,1]$ of the tables. The corresponding probability model μ_Y^I on $(\mathcal{Y}_I, \mathbf{2}^{\mathcal{Y}_I})$ should be PCRP described in the previous section (Figure 7, middle). Then, connecting the bridging state in the auxiliary space $\mathcal{Y}_I$ with samples in $\mathcal{X}_I$, we consider the combined probability model over the union space $\mathcal{X}_I \cup \mathcal{Y}_I$. To address probabilistic models on the union space $\mathcal{X}_I \cup \mathcal{Y}_I$, we modify the previous probabilistic models μ_X^I and μ_Y^I as follows: $\hat{\mu}_X^I(\cdot \in \mathbf{2}^{\mathcal{X}_I \cup \mathcal{Y}_I}) := \mu_X^I(\cdot \cap \mathcal{X}_I)$ and $\hat{\mu}_Y^I(\cdot \in \mathbf{2}^{\mathcal{X}_I \cup \mathcal{Y}_I}) := \mu_Y^I(\cdot \cap \mathcal{Y}_I)$. For simplicity of notation, the modified probabilistic models $\hat{\mu}_X^I$ and $\hat{\mu}_Y^I$ will be denoted as original μ_X^I and μ_Y^I in the following. Finally, we consider a probabilistic model μ_+^I (Figure 7, right) in form of $\mu_+^I = \alpha\mu_X^I + (1-\alpha)\mu_Y^I$, where $0 \leq \alpha \leq 1$ is a tunable real variable. Fortunately, we can treat this new model μ_+^I as a BNP model, that is, there uniquely exists a projective limit of μ_+^I $(I \to \mathbb{N} \times \mathbb{N})$ (See the supplementary materials for details). In summary, the BNP model with bridging states has the following advantages: (1) The family $\langle \mu_+^I \rangle_{I \in E}$ can be treated in the same way as the usual BNP model, i.e., the model itself has self-similarity (self-consistency) regardless of the size of the observed inputs. (2) If we restrict the space of interest to $\mathcal{X}_I$ (i.e. forget about the auxiliary space $\mathcal{Y}_I$), it becomes equivalent to the exact BNP model on rectangular partitions. More specifically, we have $\mu_X^I(\cdot) = (1/\alpha)\mu_+^I(\cdot \cap \mathcal{X}_I)$.

5 Application to relational data analysis

Relational models - The PCRP-based relational model is applied to the input observation matrix $\boldsymbol{Z} = (Z_{i,j})_{N \times M}$. We suppose that the input matrix $\boldsymbol{Z}$ consists of categorical elements, i.e., $Z_{i,j} \in \{1, 2, \ldots, D\}$ $(D \in \mathbb{N})$. The generative model can be constructed as follows. First, we draw a random rectangular partitioning from PCRP. Each block has a latent Dirichlet random variable $\vartheta_k \sim \mathrm{Dirichlet}(\boldsymbol{\alpha}_0)$ $(k = 1, 2, \ldots)$, where $\boldsymbol{\alpha}_0$ is a D-dimensional non-negative hyper parameter. Each element $Z_{i,j}$ is drawn from the multinomial distribution with the parameter $\vartheta_{\mathbf{k}(i,j)}$, where $\mathbf{k}(i,j)$ indicates the block index to which the entry with the i-th row and the j-th column belongs.

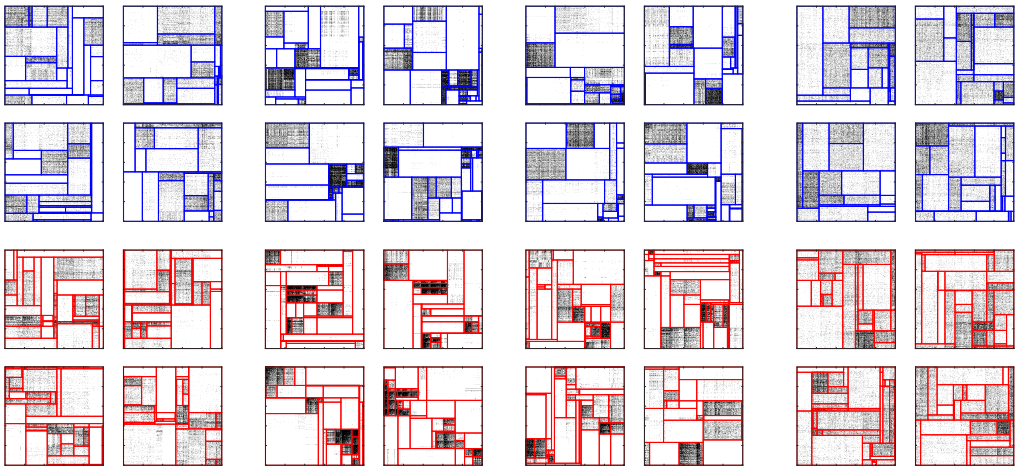
We compare PCRP with the existing BNP models for rectangular partitioning, IRM [39], MP [56], RTP [48], and BBP [47], whose hyper-parameter settings and MCMC inference methods are provided in the supplementary material. For PCRP, in order to evaluate its original capability in the main body, we set $\alpha = 0$ and use the case of generic rectangulations, which have the highest expressive capability. Our code will be available on https://SupplementaryMaterial.

Datasets - Four social network datasets [40]: **Wiki** [1], **Facebook** [2], **Twitter** [3], and **Epinions** [4]. All data is public and does not contain any personally identifiable information (See [41] for license). We selected the top 1000 active nodes based on their interactions with others; subsequently we randomly sampled 500×500 matrix to construct the relational data, as in [26, 47]. We held out 20% cells of the input data for testing, and each model was trained by the MCMC using the remaining 80% of the cells. We evaluated the models using perplexity as a criterion: $\mathrm{perp}(\hat{Z}) = \exp(-(\log p(\hat{Z}))/E)$, where E is the number of non-missing cells in the partitioned matrix $\hat{Z}$.



Wiki data Facebook data Twitter data Epinion data

Figure 8: Experimental results of perplexity comparison for four real world data, relationship between test perplexity (mean$\pm$std) evolution and logarithm of wall-clock time.



Wiki data Facebook data Twitter data Epinion data

Figure 9: **Top:** Four examples of BBP analysis results for each data. **Bottom:** Four examples of PCRP analysis results for each data. While BBP is indeed able to reveal the coarse clusters, we can confirm that PCRP is able to avoid some local optima and capture more details of hidden fine clusters.

Experimental results - We ran 10 trials of analysis for each method on each data set. Figure 8 summarizes the test perplexity comparison results. We recall that IRM and MP are limited in the class of rectangular partitions they can represent, and only PCRP, BBP, and RTP can represent arbitrary rectangular partitioning. Indeed, it can be seen that these arbitrary rectangular partitioning models have good prediction accuracy when MCMC converges. Among them, PCRP in particular achieves the best prediction accuracy for all the data. As another perspective, BBP, RTP, and MP are SBP-type models, while PCRP and IRM are CRP-type models. It can be seen that PCRP improves the perplexity from an early stage compared to SBP-like methods. In light of the above, we can confirm that PCRP can reduce the problems of local solutions and slow mixing in Bayesian inference. In terms of the qualitative aspects of the analysis, Figure 9 shows four samples of analysis for each data set. We can confirm that PCRP can extract more detailed clusters than BBP.

6 Conclusion and Discussion

Summary - This paper has proposed a new stochastic process for relational data analysis. Our main contributions are as follows: (1) We introduce the notion of the permuton to the CRP and obtain a probabilistic model of rectangular partitioning of a matrix. (2) We show a unified framework for PCRP to various classes of rectangular partitioning based on the design of the permuton. (3) We discuss the theoretical imperfection of PCRP for projectivity and exchangeability in terms of rectangular partitioning and propose a new representation of the BNP model as a way to overcome this imperfection that applies PCRP as a bridging state to existing BNP relational models.

Ethical perspective - Our work does not encourage unethical aspects of machine learning technologies. However, as is often the case with clustering methods, our proposal can be misused in a variety of situations. Since PCRP may reveal hidden clusters from any input matrices, unanticipated cues can lead to unanticipated results. This issue highly depends on the choice of input data. Therefore, what is suitable as input data needs to be carefully considered from an ethical perspective.

References

[1] http://snap.stanford.edu/data/wiki-Vote.html

[2] http://snap.stanford.edu/data/ego-Facebook.html

[3] http://snap.stanford.edu/data/ego-Twitter.html

[4] http://snap.stanford.edu/data/soc-Epinions1.html

[5] Ackerman, E., Barequet, G., Pinter, R.Y.: A bijection between permutations and floorplans, and its applications. Discrete Applied Mathematics **154**(12), 1674–1684 (2006)

[6] Adams, R.P., Jordan, M.I., Ghahramani, Z.: Tree-structured stick breaking for hierarchical data. In: Lafferty, J., Williams, C., Shawe-Taylor, J., Zemel, R., Culotta, A. (eds.) Advances in Neural Information Processing Systems. vol. 23. Curran Associates, Inc. (2010)

[7] Airoldi, E.M., Costa, T.B., Chan, S.H.: Stochastic blockmodel approximation of a graphon: Theory andconsistent estimation. In: Advances in Neural Information Processing Systems (2013)

[8] Aldous, D.J.: Representations for partially exchangeable arrays of random variables. Journal of Multivariate Analysis **11**, 581–598 (1981)

[9] Aldous, D.J.: Exchangeability and related topics. École d'Été St Flour, Lecture Notes in Mathematics **1117**, 1–198 (1985)

[10] Avis, D., Fukuda, K.: Reverse search for enumeration. Discrete Applied Mathematics **65**(1), 21–46 (1996)

[11] Bassino, F., Bouvel, M., Feray, V., Gerin, L., Maazoun, M., Pierrot, A.: Universal limits of substitution-closed permutation classes. Journal of the European Mathematical Society **22**(11), 3565–3639 (2019)

[12] Bassino, F., Bouvel, M., Féray, V., Gerin, L., Pierrot, A.: The Brownian limit of separable permutations. The Annals of Probability **46**(4), 2134—-2189 (2018)

[13] Bassino, F., Bouvel, M., Féray, V., Gerin, L., Maazoun, M., Pierrot, A.: Scaling limits of permutation classes with a finite specification: a dichotomy. arXiv:1903.07522 (2019)

[14] Baxter, G.: On fixed points of the composite of commuting functions. Proceedings of American Mathematical Society **15**, 851–855 (1964)

[15] Blei, D., Frazier, P.: Distance dependent Chinese restaurant processes. In: International Conference on Machine Learning (2010)

[16] Blei, D., Frazier, P.: Distance dependent Chinese restaurant processes. Journal of Machine Learning Reseach **12**, 2461–2488 (2011)

[17] Blei, D.M., Jordan, M.I., Griffiths, T.L., Tenenbaum, J.B.: Hierarchical topic models and the nested Chinese restaurant process. pp. 17–24 (2003)

[18] Blei, D.M., Griffiths, T.L., Jordan, M.I.: The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. J. ACM **57**(2) (2010)

[19] Borga, J., Maazoun, M.: Scaling and local limits of Baxter permutations and bipolar orientations through coalescent-walk processes. arXiv:2008.09086 (2020)

[20] Caldas, J., Kaski, S.: Bayesian biclustering with the plaid model. In: 2008 IEEE Workshop on Machine Learning for Signal Processing. pp. 291–296 (2008)

[21] Choi, D.S., Wolfe, P.J.: Co-clustering separately exchangeable network data. Annals of Statistics **42**, 29–63 (2014)

[22] Fan, X., Li, B., Sisson, S.A.: The binary space partitioning-tree process. In: International Conference on Artificial Intelligence and Statistics. pp. 1859–1867 (2018)

[23] Fan, X., Li, B., Luo, L., Sisson, S.A.: Bayesian nonparametric space partitions: A survey. arXiv:2002.11394 (2020)

[24] Fan, X., Li, B., Sisson, S.: Rectangular bounding process. In: Advances in Neural Information Processing Systems. pp. 7631–7641 (2018)

[25] Fan, X., Li, B., Sisson, S.A.: Online binary space partitioning forests. arXiv:2003.00269 (2020)

[26] Fan, X., Li, B., Sisson, S.A.: Binary space partitioning forests. arXiv:1903.09348 (2019)

[27] Fan, X., Li, B., Wang, Y., Wang, Y., Chen, F.: The Ostomachion Process. In: AAAI Conference on Artificial Intelligence. pp. 1547–1553 (2016)

[28] Ferguson, T.: Bayesian analysis of some nonparametric problems. Annals of Statistics **2**(1), 209–230 (1973)

[29] de Finetti, B.: Funzione caratteristica di un fenomeno aleatorio. Memorie della R. Academia Nazionale dei Lincei **4**(5), 251–299 (1930)

[30] Frédérique, B., B. Mathilde, F.V., Lucas, G., M. Mickaël, P.A.: Universal limits of substitution-closed permutation classes. Journal of European Mathematical Society **22**, 3565–3639 (2020)

[31] Ge, S., Wang, S., Teh, Y.W., Wang, L., Elliott, L.: Random tessellation forests. In: Advances in Neural Information Processing Systems 32, pp. 9575–9585 (2019)

[32] Ghosh, S., Ungureunu, A., Sudderth, E., Blei, D.: Spatial distance dependent Chinese restaurant processes for image segmentation. In: Advances in Neural Information Processing Systems (2011)

[33] He, B.D.: A simple optimal binary representation of mosaic floorplans and Baxter permutations. Theoretical Computer Science **532**(1), 40–50 (2014)

[34] Hong, X., Huang, G., Cai, Y., Gu, J., Dong, S., Cheng, C., Gu, J.: Corner block list: an effective and efficient topological representation of non-slicing floorplan. In: Proceedings of the IEEE/ACM International Conference on Computer-Aided Design (2000)

[35] Hoover, D.N.: Relations on probability spaces and arrays of random variables. Tech. rep., Institute of Advanced Study, Princeton (1979)

[36] Hoppen, C., Kohayakawa, Y., Moreira, C.G., Ráth, B., Sampaio, R.M.: Limits of permutation sequences. Journal of Combinatorial Theory, Series B **103**(1), 93—-113 (2013)

[37] Ishiguro, K., Sato, I., Nakano, M., Kimura, A., Ueda, N.: Infinite plaid models for infinite bi-clustering. pp. 1701–1708 (2016)

[38] Kallenberg, O.: Symmetries on random arrays and set-indexed processes. Journal of Theoretical Probability **5**(4), 727–765 (1992)

[39] Kemp, C., Tenenbaum, J.B., Griffiths, T.L., Yamada, T., Ueda, N.: Learning systems of concepts with an infinite relational model. In: AAAI Conference on Artificial Intelligence. pp. 381–388 (2006)

[40] Leskovec, J., Huttenlocher, D., Kleinberg, J.: Predicting positive and negative links in online social networks. In: Proceedings of the 19th International Conference on World Wide Web. pp. 641–650 (2010)

[41] Leskovec, J., Krevl, A.: SNAP Datasets: Stanford large network dataset collection. `http://snap.stanford.edu/data` (Jun 2014)

[42] Lin, D., Fisher, J.: Efficient sampling from combinatorial space via bridging. In: International Conference on Artificial Intelligence and Statistics (2012)

[43] Lloyd, J., Orbanz, P., Ghahramani, Z., Roy, D.M.: Random function priors for exchangeable arrays with applications to graphs and relational data. In: Advances in Neural Information Processing Systems (2012)

[44] Lovász, L.: Very large graphs. Current Developments in Mathematics **11**, 67–128 (2009)

[45] Maazoun, M.: On the brownian separable permuton. Combinatorics, Probability and Computing **29**(2), 241–266 (2019)

[46] Miller, K., Jordan, M.I., Griffiths, T.L.: Nonparametric latent feature models for link prediction. pp. 1276–1284 (2009)

[47] Nakano, M., Kimura, A., Yamada, T., Ueda, N.: Baxter permutation process. In: Advances in Neural Information Processing Systems (2020)

[48] Nakano, M., Ishiguro, K., Kimura, A., Yamada, T., Ueda, N.: Rectangular tiling process. In: Proceedings of the 31st International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 32, pp. 361–369 (2014)

[49] Nakano, S.: Enumerating floorplans with n rooms. In: The 12th International Symposium on Algorithms and Computation (2001)

[50] Orbanz, P., Roy, D.M.: Bayesian models of graphs, arrays and other exchangeable random structures. IEEE Transactions on Pattern Analysis and Machine Intelligence **37**, 437–461 (2013)

[51] Pitman, J., Yor, M.: The two-parameter poisson-dirichlet distribution derived from a stable subordinator. IBM Journal of Research and Development **2**(25), 855–900 (1997)

[52] Rasmussen, C.: The infinite gaussian mixture model. In: Advances in Nueral Information Processing Systems (2000)

[53] Reading, N.: Generic rectangulations. European Journal of Combinatorics **33**(4), 610–623 (2012)

[54] Rodriguez, A., Ghosh, K.: Nested partition models. Tech. rep., JackBaskin School of Engineering (2009)

[55] Roy, D.M.: Computability, inference and modeling in probabilistic programming. Ph.D. thesis, Massachusetts Institute of Technology (2011)

[56] Roy, D.M., Teh, Y.W.: The Mondrian process. In: Advances in Neural Information Processing Systems (2009)

[57] Sethuraman, J.: A constructive definition of Dirichlet priors. Statistica Sinica **4**, 639–650 (1994)

[58] Shan, H., Banerjee, A.: Bayesian co-clustering. In: IEEE International Conference on Data Mining. pp. 530–539 (2008)

[59] Shen, Z.C., Chu, C.C.N.: Bounds on the number of slicing, mosaic, and general floorplans. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems **22**(10), 1354–1361 (2003)

[60] Takagi, M., Nakano, S.: Listing all rectangular drawings with certain properties. Systems and Computers in Japan **35**(4), 1–8 (2004)

[61] Tan, X., Rao, V., Neville, J.: Nested crp with hawkes-gaussian processes. In: Storkey, A., Perez-Cruz, F. (eds.) Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, vol. 84, pp. 1289–1298 (2018)

[62] Teh, Y.W., Jordan, M.I., Beal, M., Blei, D.: Hierarchical Dirichlet processes. Journal of the AmericanStatistical Association **101**, 1566–1581 (2006)

[63] Yao, B., Chen, H., Cheng, C.K., Graham, R.L.: Floorplan representations: Complexity and connections. ACM Transactions on Design Automation of Electronic Systems **8**(1), 55–80 (2003)

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See Section 1. We have clarified the main contributions of this paper under three headings.
   (b) Did you describe the limitations of your work? [Yes] See Section 4. We have simultaneously clarified the technical imperfection of our proposed PCRP and proposed a reasonable proposal to solve it.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See **Ethical perspective** in Section 6.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] Section 4 is an intuitive sketch of the theoretical results. Details are given in the supplementary material.
   (b) Did you include complete proofs of all theoretical results? [Yes] See Supplementary material.
3. If you ran experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Our code is provided in Supplementary material

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See experimental settings in Supplementary material

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See experimental settings in Supplementary material

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [Yes] See **Relational models** in Section 5

(b) Did you mention the license of the assets? [Yes] See **Datasets** in Section 5

(c) Did you include any new assets either in the supplemental material or as a URL? [Yes] We plan to make all the code available via GitHub. For the sake of anonymity in peer review, they are provided in the supplementary material

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] See **Datasets** in Section 5

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] See **Datasets** in Section 5

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]