Bayesian Models of Graphs, Arrays and Other Exchangeable Random Structures

Peter Orbanz and Daniel M. Roy

Abstract. The natural habitat of most Bayesian methods is data represented by exchangeable sequences of observations, for which de Finetti's theorem provides the theoretical foundation. Dirichlet process clustering, Gaussian process regression, and many other parametric and nonparametric Bayesian models fall within the remit of this framework; many problems arising in modern data analysis do not. This expository paper provides an introduction to Bayesian models of graphs, matrices, and other data that can be modeled as arrays of random variables. We describe results in probability theory that generalize de Finetti's theorem to such data and discuss the relevance of these results to nonparametric Bayesian modeling. With the basic ideas in place, we survey example models available in the literature; applications of such models include collaborative filtering, link prediction, and graph and network analysis. We also highlight connections to recent developments in graph theory and probability, and sketch the more general mathematical foundation of Bayesian methods for other types of data beyond sequences and arrays.

1. Introduction.

For data represented by exchangeable sequences, Bayesian nonparametrics has developed into a flexible and powerful toolbox of models and algorithms. Its modeling primitives—Dirichlet processes, Gaussian processes, etc.—are widely applied and well-understood, and can be used as components in hierarchical models [60] or dependent models [48] to address a wide variety of data analysis problems. One of the main challenges for Bayesian statistics and machine learning is arguably to extend this toolbox to data such as graphs, networks and relational data.

The type of data we focus on in this article are array-valued observations. By a **random d-dimensional array**, or simply **d-array**, we will mean a collection of random variables $X_{i_1 \ldots i_d}$, $(i_1, \ldots, i_d) \in \mathbb{N}^d$, indexed by d-tuples of natural numbers. A sequence is a 1-array; a matrix is a 2-array. A special case of particular importance is graph-valued data (which can represented by an adjacency matrix, and hence by a random 2-array). Array-valued data arises in problems such link prediction, citation matching, database repair, and collaborative filtering.

If we model such data naively, we encounter a variety of difficult questions: On what parameter space should we define priors on graphs? In a collaborative filtering task, what latent variables render user data conditionally independent? What can we expect to learn about an infinite random graph if we only observe a finite subgraph, however large? There are answers to these questions, and most of them can be deduced from a single result, known as the Aldous-Hoover theorem [3, 34], which gives a precise characterization of the conditional independence structure of random graphs and arrays if they satisfy an exchangeability property. Hoff [31] was the first to invoke this result in the machine learning literature.

This article explains the Aldous-Hoover theorem and its application to Bayesian modeling. The larger theme is that most Bayesian models for "structured" data can be understood as exchangeable random structures. Each type of structure comes with its own representation theorem. In the simplest case—exchangeable sequences represented by de Finetti's theorem—the Bayesian modeling approach is well-established. For more complex data, the conditional independence properties requisite to statistical inference are more subtle, and if representation results are available, they offer concrete guidance for model design. On the other hand, the theory also clarifies the limitations of exchangeable models—it shows, for example, that most Bayesian models of network data are inherently misspecified, and why developing Bayesian models for sparse structures is hard.

Contents

Section 2: reviews exchangeable random structures, their representation theorems, and the role of such theorems in Bayesian statistics.

Section 3: introduces the generalization of de Finetti's theorem to models of graph- and matrix-valued data, the Aldous-Hoover theorem, and explains how Bayesian models of such data can be constructed.

Section 4: surveys models of graph- and relational data available in the machine learning and statistics literature. Using the Aldous-Hoover representation, models can be classified and some close connections emerge between models which seem, at first glance, only loosely related.

Section 5: describes recent development in the mathematical theory of graph limits. The results of this theory refine the Aldous-Hoover representation of graphs and provide a precise understanding of how graphs converge and how random graph models are parametrized.

Section 6: explains the general Aldous-Hoover representation for higher-order arrays.

Section 7: discusses sparse random structures and networks, why these models contradict exchangeability, and open questions arising from this contradiction.

Section 8: provides references for further reading.

2. Bayesian Models of Exchangeable Structures.

The models of graph- and array-valued data described in this article are special cases of a very general approach: Bayesian models that represent data by a random structure, and use exchangeability properties to deduce valid statistical models and useful parametrizations. This section sketches out the ideas underlying this approach, before we focus on graphs and matrices in Section 3.

We are interested in random structures—sequences, partitions, graphs, functions, and so on—that possess an exchangeability property: i.e., certain components of the structure—the elements of a sequence or the rows and columns of a matrix, for example—can be rearranged without affecting the distribution of the structure. Formally speaking, the distribution is invariant to the action of some group of permutations. Borrowing language introduced by David Aldous in applied probability [6], we collectively refer to random structures with such a property as **exchangeable random structures**, even though the specific definition of exchangeability may vary considerably. Table 1 lists some illustrative examples.

The general theme is as follows: The random structure is a random variable X_∞ with values in a space $\mathbf{X}_\infty$ of infinite sequences, graphs, matrices, etc. If X_∞ satisfies an exchangeability property, this property determines a special family $\{\mathbf{p}(\,.\,,\theta) : \theta \in \mathbf{T}\}$ of distributions on $\mathbf{X}_\infty$, which are called the **ergodic distributions**. The distribution of X_∞ then has a *unique* integral decomposition

$$\mathbb{P}(X_\infty \in \,.\,) = \int_{\mathbf{T}} \mathbf{p}(\,.\,,\theta)\nu(d\theta)\ . \tag{2.1}$$

The distribution of X_∞ is completely determined by ν, and vice versa, i.e., Eq. (2.1) determines a bijection

$$\mathbb{P}(X_\infty \in \,.\,) \quad \longleftrightarrow \quad \nu\ .$$

The integral represents a *hierarchical model*: We can sample X_∞ in two stages,

$$\begin{aligned} \Theta &\sim \nu \\ X_\infty|\Theta &\sim \mathbf{p}(\,.\,,\Theta)\ . \end{aligned} \tag{2.2}$$

In Bayesian modeling, the distribution ν in Eq. (2.1) plays the role of a prior distribution, and a specific choice of ν determines a Bayesian model on $\mathbf{X}_\infty$.

Virtually all Bayesian model imply some form of exchangeability assumption, although not always in an obvious form. Eq. (2.1) and (2.2) give a first impression of why the concept is so important: If data is represented by an exchangeable random structure, the observation model is a subset of the ergodic distributions, and the parameter space of the model is either the space $\mathbf{T}$ or a subspace. Given a specific type of exchangeable structure, the representation theorem specifies these components. Perhaps the most important role is played by the ergodic distributions: The form of these distributions explains conditional independence properties of the random structure X_∞. For exchangeable sequences, observations are simply conditionally independent and identically distributed (i.i.d.) given Θ. In other exchangeable structures, the independence properties are more subtle—exchangeability generalizes beyond sequences, whereas the conditional i.i.d. assumption does not.

2.1. *Basic examples: Sequences and partitions.* Exchangeable sequences are the canonical example of exchangeable structures. An **exchangeable sequence** is an infinite sequence $X := (X_1, X_2, \dots)$ of random variables whose joint distribution satisfies

$$\begin{aligned} &\mathbb{P}(X_1 \in A_1, X_2 \in A_2, \dots) \\ &\quad = \mathbb{P}(X_{\pi(1)} \in A_1, X_{\pi(2)} \in A_2, \dots) \end{aligned} \tag{2.3}$$

for every permutation π of $\mathbb{N} := \{1, 2, \dots\}$ and collection $A_1, A_2, \dots$ of (measurable) sets. Because expressing distributional equalities this way is cumbersome, we will instead write $Y \overset{\mathrm{d}}{=} Z$ whenever two random variables Y and Z have the same distribution. Therefore, we can express Eq. (2.3) by

$$(X_1, X_2, \dots) \overset{\mathrm{d}}{=} (X_{\pi(1)}, X_{\pi(2)}, \dots)\ , \tag{2.4}$$

or simply by $(X_n) \overset{\mathrm{d}}{=} (X_{\pi(n)})$, where the range of the variable n is left implicit. If $X_1, X_2, \dots$ are exchangeable, then de Finetti's representation theorem implies they are even conditionally i.i.d.:

Theorem 2.1 (de Finetti). *Let $X_1, X_2, \dots$ be an infinite sequence of random variables with values in a space* $\mathbf{X}$.

1. *The sequence $X_1, X_2, \dots$ is exchangeable if and only if there is a random probability measure Θ on* $\mathbf{X}$*— i.e., a random variable with values in the set* $\mathbf{M}(\mathbf{X})$

Probabilistic terminology

We assume familiarity with basic notions of probability and measure theory, but highlight two key notions here: *Measurable functions* play a prominent role in the representation results, especially those of the form $f : [0,1]^d \to [0,1]$, and we encourage readers to think of such functions as "nearly continuous". More precisely, f is **measurable** if and only if, for every $\varepsilon > 0$, there is a continuous function $f_\varepsilon : [0,1]^d \to [0,1]$ such that $\mathbb{P}(f(U) \neq f_\varepsilon(U)) < \varepsilon$, where U is uniformly distributed in $[0,1]^d$. Another concept we use frequently is that of a *probability kernel*, the mathematical representation of a conditional probability. Formally, a **probability kernel p from $\mathcal{Y}$ to X** is a measurable function from $\mathcal{Y}$ to the set $\mathbf{M}(\mathbf{X})$ of probability measures on $\mathbf{X}$. For a point $y \in \mathcal{Y}$, we write $\mathbf{p}(\,.\,,y)$ for the probability measure on $\mathbf{X}$. For a measurable subset $A \subseteq \mathbf{X}$, the function $\mathbf{p}(A,\,.\,)$ is a measurable function from $\mathcal{Y}$ to $[0,1]$. Note that for every pair of random variables, e.g., in $\mathbb{R}$, there is a probability kernel $\mathbf{p}$ from $\mathbb{R}$ to $\mathbb{R}$ such that $\mathbf{p}(\,.\,,Y) = \mathbb{P}[X \in \,.\,|Y]$.

of probability distributions on $\mathbf{X}$—such that the X_i are conditionally i.i.d. given Θ and

$$\mathbb{P}(X_1 \in A_1, X_2 \in A_2, \dots) = \int_{\mathbf{M}(\mathbf{X})} \prod_{i=1}^{\infty} \theta(A_i) \nu(d\theta) \tag{2.5}$$

where ν is the distribution of Θ. We call ν the ***mixing measure*** *and Θ the* ***directing random measure****. (Some authors call ν the* de Finetti measure*.)*

2. *If the sequence is exchangeable, the empirical distributions*

$$\hat{S}_n(\,.\,) := \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}(\,.\,), \quad n \in \mathbb{N}, \tag{2.6}$$

converge to Θ as $n \to \infty$ in the sense that

$$\hat{S}_n(A) \to \Theta(A) \quad as \quad n \to \infty \tag{2.7}$$

with probability 1 under ν and for every (measurable) set A.

□

Comparing to Eq. (2.1), we see that the ergodic distributions are the factorial distributions

$$\mathbf{p}(A_1 \times A_2 \times \cdots, \theta) = \prod_{i=1}^{\infty} \theta(A_i) \;,$$

for every sequence of measurable subsets A_i of $\mathbf{X}$. The hierarchical structure is of the form:

$$\Theta \sim \nu \tag{2.8}$$

$$X_i \mid \Theta \sim_{\text{iid}} \Theta. \tag{2.9}$$

We have mentioned above that the ergodic distributions explain conditional independence properties within the random structure. Exchangeable sequences are a particularly simple case, since the elements of the sequence completely decouple given the value of Θ, but we will encounter more intricate forms of conditional independence in Section 3.

A second illustrative example of an exchangeability theorem is Kingman's theorem for exchangeable partitions, which explains the role of exchangeability in clustering problems. A clustering solution is a partition of $X_1, X_2, \dots$ into disjoint sets. A clustering solution can be represented as a partition $\pi = (b_1, b_2, \dots)$ of the index set $\mathbb{N}$. Each of the sets b_i, called **blocks**, is a finite or infinite subset of $\mathbb{N}$; every element of $\mathbb{N}$ is contained in exactly one block. An **exchangeable partition** is a random partition X_∞ of $\mathbf{N}$ which is invariant under permutations of $\mathbb{N}$. Intuitively, this means the probability of a partition depends only on the sizes of its blocks, but not on which elements are in which block.

Kingman [39] showed that exchangeable random partitions can again be represented in the form Eq. (2.1), where the ergodic distributions $\mathbf{p}(\,.\,, \theta)$ are a specific form of distribution which he referred to as **paint-boxes**. To define a paint-box, let $\theta := (s_1, s_2, \dots)$ be a sequence of scalars $s_i \in [0,1]$ such that

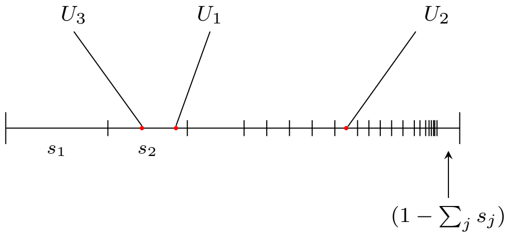

Fig 1: Sampling from a paint-box distribution with parameter $\mathbf{s} = (s_1, s_2, \dots; \bar{s})$. Two numbers i, j are assigned to the same block of the partition if the uniform variables U_i and U_j are contained in the same interval.

$$s_1 \geq s_2 \geq \dots \quad \text{and} \quad \sum_i s_i \leq 1 \;. \tag{2.10}$$

Then θ defines a partition of $[0,1]$ into intervals

$$I_j := \Big[\sum_{i=1}^{j} s_i, \sum_{i=1}^{j+1} s_i\Big) \quad \text{and} \quad \bar{I} := \Big(1 - \sum_{i=1}^{\infty} s_i, 1\Big] \;, \tag{2.11}$$

as shown in Fig. 1. The paint-box distribution $\mathbf{p}(\,.\,, \theta)$ now generates a random partition of $\mathbb{N}$ as follows:

1. Generate $U_1, U_2, \dots \sim_{\text{iid}} \text{Uniform}[0,1]$.
2. Assign $n \in \mathbb{N}$ to block b_j if $U_n \in I_j$. Assign every remaining element, i.e., those n such that $U_n \in \bar{I}$, to its own blocks of size one.

Theorem 2.2 (Kingman). *Let X_∞ be random partition of $\mathbb{N}$.*

1. *X_∞ is exchangeable if and only if*

$$\mathbb{P}(X_\infty \in \,.\,) = \int_{\mathbf{T}} \mathbf{p}(\,.\,, \theta) \nu(\theta) \;, \tag{2.12}$$

where $\mathbf{T}$ is the set of sequences $\theta = (s_1, s_2, \dots)$ as defined above, and $\mathbf{p}(\,.\,, \theta)$ is the paint-box distribution with parameter θ.

2. *If X_∞ is exchangeable, the scalars s_i can be recovered asymptotically as limiting relative block sizes*

$$s_i = \lim_{n \to \infty} \frac{|b_i \cap \{1, \dots, n\}|}{n} \;. \tag{2.13}$$

□

Example 2.3 (Chinese restaurant process). A well-known example of a random partition is the Chinese restaurant process (CRP; see e.g. [30, 54] for details). The CRP is a discrete-time stochastic process which generates a partition of $\mathbb{N}$. Its distribution is determined by a scalar concentration parameter α; different values of α correspond to different distributions $\mathbb{P}(X_\infty \in \,.\,)$ in Eq. (2.12). If X_∞ is generated by a CRP, the paint-box Θ is essentially the sequence of weights generated by the "stick-breaking" construction of the Dirichlet process [30]—with the difference that the elements of Θ are ordered by size, whereas stick-breaking weights are not. In other words, sampling from ν in Eq. (2.12) can be defined as a stick-breaking and subsequent ordering. ◁

2.2. *Exchangeability and Bayesian Theory.* A more formal description is helpful to understand the role of the exchangeability assumption and of representation theorems. It requires a brief review of the formal approach to modeling repetitive observations: Formal models represent randomness by an abstract probability space Ω, with a probability distribution $\mathbb{P}$ defined on it. A random variable is a mapping $X : \Omega \to \mathbf{X}$. A single value $\omega \in \Omega$ contains all information about a statistical experiment, and is never observed itself. Intuitively, it can be helpful to think of ω as a state of the universe; the mapping X picks out some specific aspect of ω, such as the outcome $X(\omega)$ of a coin flip.

If we record repetitive observations $X_1, X_2, \ldots \in \mathbf{X}$, all recorded values are still governed by a single value of ω, i.e., we observe $(X_1(\omega), \ldots, X_n(\omega))$. The sample can be collected in the **empirical distribution**

$$S_n(X_1, \ldots, X_n) := \frac{1}{n} \sum_{i \leq n} \delta_{X_i}. \tag{2.14}$$

A fundamental assumption of statistics is that the distribution of data can asymptotically be recovered from observations. If the infinite sequence $X^\infty = (X_1, X_2, \ldots)$ is assumed exchangeable, Theorem 2.1 shows that the empirical distribution converges to the distribution of the variables X_i as $n \to \infty$. Thus, if we define $S := \lim_n S_n$, the limiting empirical distribution $S \circ X^\infty$ coincides with the distribution of the X_i. In a frequentist setting, we would similarly assume X^∞ to be i.i.d., and convergence is then guaranteed by the law of large numbers or the Glivenko-Cantelli theorem.

An **observation model** is now a subset

$$\mathcal{P} = \{P_\theta \in \mathbf{M}(\mathbf{X}) \,|\, \theta \in \mathbf{T}\} \tag{2.15}$$

of the space $\mathbf{M}(\mathbf{X})$ of distributions on $\mathbf{X}$. To tie the various ingredients together, the following type of diagram (due to Schervish [58]) is very useful:

$$\begin{tikzcd}
\Omega \arrow[r, "X^\infty"] \arrow[rrr, bend right, "\Theta"'] & \mathbf{X}^\infty \arrow[r, "S"] & \mathbf{M}(\mathbf{X}) \supset \mathcal{P} \arrow[r, shift left, "T"] & \mathbf{T} \arrow[l, shift left, "T^{-1}"]
\end{tikzcd} \tag{2.16}$$

Each parameter value θ uniquely corresponds to a single distribution $P_\theta \in \mathbf{M}(\mathbf{X})$. The correspondence between the two is formalized by a bijective mapping T with $T(P_\theta) = \theta$, called a **parametrization**.

The mappings in the diagram can be composed into a single mapping by defining

$$\Theta := T \circ S \circ X^\infty \,. \tag{2.17}$$

From a Bayesian perspective, this is the model parameter. If we identify Θ and P_Θ, then Θ is precisely the directing random measure in de Finetti's theorem. Its distribution $\nu(\,.\,) = \mathbb{P}(\Theta \in \,.\,)$ is the prior. If we were to observe the entire infinite sequence X^∞, then $S \circ X^\infty = T \circ \Theta$ would identify a single distribution on $\mathbf{X}$. In an actual experiment, we only observe a finite subsequence X^n, and the remaining uncertainty regarding Θ is represented by the posterior $\mathbb{P}[\Theta \in \,.\,|X^n]$.

To generalize this approach to exchangeable structures, we slightly change our perspective by thinking of X^∞ as a single random structure, rather than a collection of repetitive observations. If $P = S \circ X^\infty$ is the limiting distribution of the X_i, then by conditional independence, P^∞ is the corresponding joint distribution of X^∞. Comparing to Eq. (2.5), we see that the distributions P^∞ are precisely the ergodic measures in de Finetti's theorem. In other words, when regarded as a distribution on $\mathbf{X}^\infty$, the empirical distribution *converges to the ergodic distribution*, and we can substitute the set $\mathcal{E}$ of ergodic distributions for $\mathbf{M}(\mathbf{X})$ in diagram (2.16). Thus, the model is now a subset $\{P_\theta^\infty \in \mathbf{M}(\mathbf{X}^\infty) \,|\, \theta \in \mathbf{T}\}$ of $\mathcal{E}$.

Now suppose $\mathbf{X}_\infty$ is a space of infinite structures—infinite graphs, sequences, partitions, etc.—and X_∞ is a random element of $\mathbf{X}_\infty$ and exchangeable. We have noted above that statistical inference is based on an independence assumption. The components of exchangeable structures are not generally conditionally i.i.d. as they are for sequences, but if a representation theorem is available, it characterizes a specific form of independence by characterizing the ergodic distributions. Although the details differ,

Table 1
Exchangeable random structures

Random structure	Theorem of	Ergodic distributions $\mathbf{p}(\,.\,,\theta)$	Statistical application
Exchangeable sequences	de Finetti [19] Hewitt and Savage [29]	product distributions	most Bayesian models [e.g. 58]
Processes with exchangeable increments	Bühlmann [17]	Lévy processes	
Exchangeable partitions	Kingman [39]	"paint-box" distributions	clustering
Exchangeable arrays	Aldous [3] Hoover [34] Kallenberg [35]	sampling schemes Eq. (6.4), Eq. (6.10)	graph-, matrix- and array-valued data (e.g., [31]); see Section 4
Block-exchangeable sequences	Diaconis and Freedman [21]	Markov chains	e.g. infinite HMMs [9, 24]

the general form of a representation theorem is qualitatively as follows:

1. It characterizes a set $\mathcal{E}$ of ergodic measures for this type of structure. The ergodic measures are elements of $\mathbf{M}(\mathbf{X}_\infty)$, but $\mathcal{E}$ is "small" as a subset of $\mathbf{M}(\mathbf{X}_\infty)$. Sampling from an ergodic distribution represents some form of conditional independence between elements of the structure X_∞.
2. The distribution of X_∞ has a representation of the form Eq. (2.1), where $\mathbf{p}(\,.\,,\theta) \in \mathcal{E}$ for every $\theta \in \mathbf{T}$.
3. The (suitably generalized) empirical distribution of a substructure of size n (e.g., of a subgraph with n vertices) converges to a specific ergodic distribution as $n \to \infty$. Defining the empirical distribution of a random structure can be a challenging problem; every representation result implies a specific definition.

In the general case, the diagram now takes the form:

$$\begin{tikzcd}
\Omega \arrow[r, "X_\infty"] \arrow[rrr, bend right, "\Theta"'] & \mathbf{X}_\infty \arrow[r, "S"] & \mathcal{E} \supset \mathcal{P} \arrow[r, shift left, "T"] & \mathbf{T} \arrow[l, shift left, "T^{-1}"]
\end{tikzcd} \tag{2.18}$$

Here, S is now the limiting distribution of a suitable "random substructure", and the model $\mathcal{P}$ is again a subset of the ergodic distributions identified by the relevant representation theorem.

In Kingman's theorem 2.2, for example, the ergodic distributions (the paint-box distributions) are parametrized by the set of decreasing sequences $\theta = (s_1, s_2, \dots)$, and convergence of S_n is formulated in terms of convergence of limiting relative blocksizes to θ. The corresponding results for random graphs and matrices turn out to be more subtle, and are discussed separately in Section 3.

2.3. *"Non-exchangeable" data.* Exchangeability seems at odds with many types of data; most time series, for example, would certainly not be assumed to be exchangeable. Nonetheless, a Bayesian model of a time series will almost certainly imply an exchangeability assumption—the crucial question is which components of the overall model are assumed to be exchangeable. As the next example illustrates, these components need not be the variables representing the observations.

Example 2.4 (Lévy processes and Bühlmann's theorem). The perhaps most widely used model for time series in continuous time are Lévy processes, i.e., a stationary stochastic process with independent increments, whose paths are piece-wise continuous functions on $\mathbb{R}_+$. If we observe values $X_1, X_2, \dots$ of this process at increasing times $t_1 < t_2 < \dots$, the variables X_i are clearly not exchangeable. However, the *increments* of the process are i.i.d. and hence exchangeable. More generally, we can consider processes whose increments are exchangeable (rather than i.i.d.). The relevant representation theorem is due to Hans Bühlmann [e.g. 37, Theorem 1.19]:

If a process with piece-wise continuous paths on $\mathbb{R}_+$ has exchangeable increments, it is a mixture of Lévy processes.

Hence, each ergodic measure $\mathbf{p}(\,.\,,\theta)$ is the distribution of a Lévy process, and the measure ν is a distribution on parameters of Lévy processes or—in the parlance of stochastic process theory—on Lévy characteristics. ◁

Example 2.5 (Discrete times series and random walks). Another important type of exchangeability property is Markov exchangeability [21, 68], which is defined for sequences $X_1, X_2, \dots$ in a countable space $\mathbf{X}$. At each new observation, the sequence may remain in the current state $x \in \mathbf{X}$, or transition to another state $y \in \mathbf{X}$. It is called **Markov exchangeable** if its joint probability depends only on the initial state and the number of transitions between each pair of values x and y, but not on when these transitions occur. In other words, a sequence is Markov exchangeable if the value of X_1 and the transition counts are a sufficient statistic. Diaconis and Freedman [21] showed the following:

If a (recurrent) process is Markov exchangeable, it is a mixture of Markov chains.

(Recurrence means that each visited state is visited infinitely often if the process is run for an infinite number of steps.) Thus, each ergodic distribution $\mathbf{p}(\,.\,,\theta)$ is the distribution of a Markov chain, and a parameter value θ consists of a distribution on $\mathbf{X}$ (the distribution of the initial state) and a transition matrix. If a Markov exchangeable process is substituted for the Markov chain in a hidden Markov model, i.e., if the Markov exchangeable variables are latent variables of the model, the resulting model can express much more general dependencies than Markov exchangeability. The infinite hidden Markov model [9] is an example; see [24]. Recent work by Bacallado, Favaro, and Trippa [8] constructs prior distributions on random walks that are Markov exchangeable and can be parametrized so that the number of occurrences of each state over time has a power-law distribution. ◁

A very general approach to modeling is to assume that an exchangeability assumption holds marginally at each value of a covariate variable z, e.g., a time or a location in space: Suppose $\mathbf{X}^\infty$ is a set of structures as described above, and $\mathbf{Z}$ is a space of covariate values. A **marginally exchangeable random structure** is a random measurable mapping

$$\xi : \mathbf{Z} \to \mathbf{X}_\infty \tag{2.19}$$

such that, for each $z \in \mathbf{Z}$, the random variable $\xi(z)$ is an exchangeable random structure in $\mathbf{X}_\infty$.

Example 2.6 (Dependent Dirichlet process). A popular example of a marginally exchangeable model is the dependent Dirichlet process (DDP) of MacEachern [48]. In this case, for each $z \in \mathbf{Z}$, the random variable $\xi(z)$ is a random probability measure whose distribution is a

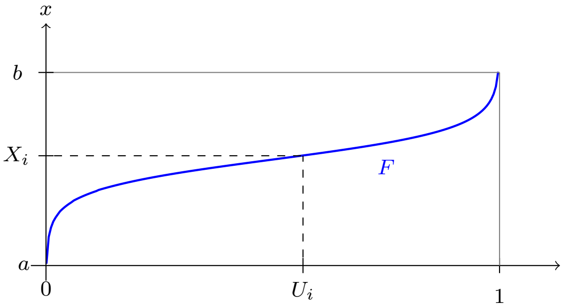

Fig 2: de Finetti's theorem expressed in terms of random functions: If F is the inverse CDF of the random measure Θ in the de Finetti representation, X_i can be generated as $X_i := F(U_i)$, where $U_i \sim \text{Uniform}[0,1]$.

Dirichlet process. More formally, $\mathbf{Y}$ is some sample space, $\mathbf{X}_\infty = \mathbf{M}(\mathbf{Y})$, and the DDP is a distribution on mappings $\mathbf{Z} \to \mathbf{M}(\mathbf{Y})$; thus, the DDP is a random probability kernel. Since $\xi(z)$ is a Dirichlet process if z is fixed, samples from $\xi(z)$ are exchangeable. ◁

Eq. (2.19) is, of course, just another way of saying that ξ is a $\mathbf{X}_\infty$-valued stochastic process indexed by $\mathbf{Z}$, although we have made no specific requirements on the paths of ξ. The path structure is more apparent in the next example.

Example 2.7 (Coagulation- and fragmentation models). If ξ is a coagulation or fragmentation process, $\mathbf{X}_\infty$ is the set of partititions of $\mathbb{N}$ (as in Kingman's theorem), and $\mathbf{Z} = \mathbb{R}_+$. For each $z \in \mathbb{R}_+$, the random variable $\xi(z)$ is an exchangeable partition—hence, Kingman's theorem is applicable marginally in time. Over time, the random partitions become consecutively finer (fragmentation processes) or coarser (coagulation processes): At random times, a randomly selected block is split, or two randomly selected blocks merge. We refer to [10] for more details and to [62] for applications to Bayesian nonparametrics. ◁

2.4. *Random functions vs random measures.* De Finetti's theorem can be equivalently formulated in terms of a random function, rather than a random measure, and this formulation provides some useful intuition for Section 3. Roughly speaking, this random function is the inverse CDF of the random measure Θ in de Finetti's theorem; see Fig. 2.

More precisely, suppose that $\mathbf{X} = [a,b]$. A measure on $[a,b]$ can by expressed by its cumulative distribution function (CDF). Hence, sampling the random measure Θ in de Finetti's theorem is equivalent to sampling a random CDF ψ. A CDF is not necessarily an invertible function, but it always admits a so-called right-continuous inverse $\overline{\psi^{-1}}$, given by

$$\overline{\psi^{-1}}(u) = \inf\{x \in [a,b] \mid \psi(x) \geq u\} . \tag{2.20}$$

This function inverts ψ in the sense that $\psi \circ \overline{\psi^{-1}}(u) = u$ for all $u \in [0,1]$. It is well-known that any scalar random variable X_i with CDF ψ can be generated as

$$X_i \overset{\text{d}}{=} \overline{\psi^{-1}}(U_i) \qquad \text{where } U_i \sim \text{Uniform}[0,1] . \tag{2.21}$$

In the special case $\mathbf{X} = [a,b]$, de Finetti's theorem therefore translates as follows: If $X_1, X_2, \dots$ is an exchangeable sequence, then there is a random function $F := \overline{\Psi^{-1}}$ such that

$$(X_1, X_2, \dots) \overset{\text{d}}{=} (F(U_1), F(U_2), \dots) , \tag{2.22}$$

where $U_1, U_2, \dots$ are i.i.d. uniform variables.

It is much less obvious that the same should hold on an arbitrary sample space, but that is indeed the case:

Corollary 2.8. *Let $X_1, X_2, \dots$ be an infinite, exchangeable sequence of random variables with values in a space $\mathbf{X}$. Then there exists a random function F from $[0,1]$ to $\mathbf{X}$ such that, if $U_1, U_2, \dots$ is an i.i.d. sequence of uniform random variables,*

$$(X_1, X_2, \dots) \overset{d}{=} (F(U_1), F(U_2), \dots). \tag{2.23}$$

□

As we will see in the next section, this random function representation generalizes to the more complicated case of array data, whereas the random measure representation in Eq. (2.5) does not. The result is formulated here as a corollary, since it formally follows from the more general theorem of Aldous and Hoover which we have yet to describe.

3. Exchangeable Graphs and Matrices.

Representing data as a matrix is a natural choice only if the subdivision into rows and columns carries information. A useful notion of exchangeability for matrices should hence preserve rows and columns, rather than permuting entries arbitrarily. There are two possible definitions: We could permute rows and columns separately, or simultanously. Both have important applications in modeling. Since rows and columns intersect, the exchangeable components are not disjoint as in de Finetti's theorem, and the entries of an exchangeable matrix are not conditionally i.i.d.

3.1. *Defining exchangeability of matrices.* We consider observations that can be represented by a **random matrix**, or **random 2-array**, i.e., a collection of random variables X_{ij}, where $i, j \in \mathbb{N}$. All variables X_{ij} take values in a common sample space $\mathbf{X}$. Like the sequences characterized by de Finetti's theorem, the matrix has infinite size, and we denote it by $(X_{ij})_{i,j\in\mathbb{N}}$, or by (X_{ij}) for short.

Definition 3.1 (Separately exchangeable array). A random array (X_{ij}) is called **separately exchangeable** if

$$(X_{ij}) \overset{\text{d}}{=} (X_{\pi(i)\pi'(j)}) \tag{3.1}$$

holds for every pair of permutations π, π' of $\mathbb{N}$. ◁

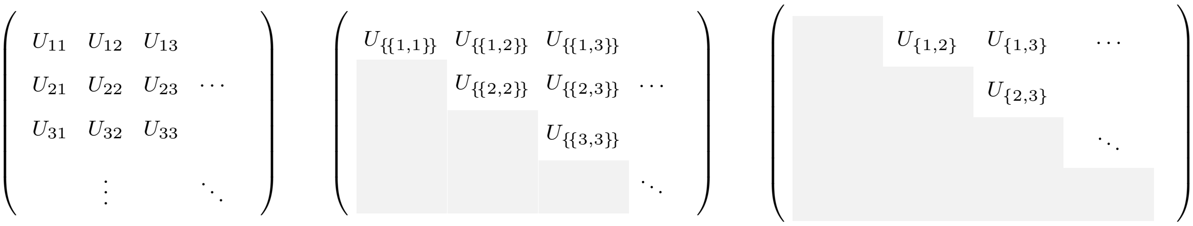
U_{11} U_{12} U_{13}
U_{21} U_{22} U_{23} $\cdots$
U_{31} U_{32} U_{33}
$\vdots$ $\ddots$
$U_{\{\{1,1\}\}}$ $U_{\{\{1,2\}\}}$ $U_{\{\{1,3\}\}}$
$U_{\{\{2,2\}\}}$ $U_{\{\{2,3\}\}}$ $\cdots$
$U_{\{\{3,3\}\}}$
$\ddots$
$U_{\{1,2\}}$ $U_{\{1,3\}}$ $\cdots$
$U_{\{2,3\}}$
$\ddots$

Fig 3: The uniform random variables $U_{\{\{i,j\}\}}$ or U_{ij} can themselves be arranged in a matrix. *Left:* In the separately exchangeable case (Corollary 3.7), the variables form an infinite matrix and are indexed as U_{ij}. *Middle:* The jointly exchangeable case (Theorem 3.4) implies a symmetric matrix $U_{ij} = U_{ji}$, which is expressed by the multiset index notation $U_{\{\{i,j\}\}}$. The subset of variables which is actually random can hence be arranged in an upper triangular matrix, which in turn determines the variables in the shaded area by symmetry. *Right:* In the special case of exchangeable random graphs (Example 3.5), the diagonal is also non-random, and variables can be indexed as $U_{\{i,j\}}$.

Applying two separate permutations to the rows and the columns is appropriate if rows and columns represent two *distinct* sets of entities, such as in a collaborative filtering problem, where rows may correspond to users and columns to movies. It is less adequate if (X_{ij}) is, for example, the adjacency matrix of a graph: In this case, there is only a single set of entities—the vertices of the graph—each of which corresponds both to a row and a column of the matrix.

Definition 3.2 (Jointly exchangeable array). A random array $(X_{ij})_{i,j\in\mathbb{N}}$ is called **jointly exchangeable** if

$$(X_{ij}) \stackrel{\text{d}}{=} (X_{\pi(i)\pi(j)}) \tag{3.2}$$

holds for every permutation π of $\mathbb{N}$. ◁

Example 3.3 (Exchangeable graph). Suppose $g = (v, e)$ is an undirected graph with an infinite (but countable) vertex set. We can label the vertices by the elements of $\mathbb{N}$. The graph can be represented by its adjacency matrix $x = (x_{ij})$, a binary matrix in which $x_{ij} = 1$ indicates that the edge between nodes i and j is present in the graph. Since the graph is undirected, the matrix is symmetric. If we replace the matrix x by a *random* matrix $X = (X_{ij})$, the edge set e is replaced by a random edge set E, and the graph becomes a random graph $G = (\mathbb{N}, E)$. We call G an **exchangeable random graph** if its adjacency matrix is a jointly exchangeable array. Thus, G is exchangeable if its distribution is invariant under relabeling of the vertices. Intuitively, this means that the probability of seeing a particular graph depends only on which patterns occur in the graph and how often—how many edges there are, how many triangles, how many five-stars, etc.—but not on where in the graph they occur. ◁

3.2. *The Representation Theorems.*

The analogue of de Finetti's theorem for exchangeable matrices is the *Aldous-Hoover theorem* [e.g. 37, Theorem 7.22]. It has two separate versions, for jointly and for separately exchangeable arrays.

Theorem 3.4 (Jointly exchangeable matrices). *A random array $(X_{ij})_{ij\in\mathbb{N}}$ is jointly exchangeable if and only if it can be represented as follows: There is a random measurable function $F : [0,1]^3 \to \mathbf{X}$ such that*

$$(X_{ij}) \stackrel{d}{=} (F(U_i, U_j, U_{\{\{i,j\}\}})) , \tag{3.3}$$

where $(U_i)_{i\in\mathbb{N}}$ and $(U_{\{\{i,j\}\}})_{i,j\in\mathbb{N}}$ are, respectively, a sequence and an array of i.i.d. Uniform$[0, 1]$ *random variables.* □

If the function F is symmetric in its first two arguments—if $F(x, y, .) = F(y, x, .)$ for all x and y—Eq. (3.3) implies the matrix X is symmetric, but a jointly exchangeability matrix X need *not* be symmetric in general.

To understand Eq. (3.3), we have to clarify various different ways of indexing variables: Roughly speaking, the variables $U_{\{\{i,j\}\}}$ account for the randomness in row-column interactions, and hence must be indexed by two values, a row index i and a column index j. Indexing them as U_{ij} would mean that, in general, U_{ij} and U_{ji} are two distinct quantities. This is not necessary in Theorem 3.4: To represent jointly exchangeable matrices, it is sufficient to sample only U_{ij}, and then set $U_{ji} := U_{ij}$. This is usually expressed in the literature by using the set $\{i, j\}$ as an index, since such sets are unordered, i.e., $\{i, j\} = \{j, i\}$. This is not quite what we need here, since a diagonal element of the matrix would have to be indexed $\{i, i\}$, but sets do not distinguish multiple occurrences of the same element—in other words, $\{i, i\} = \{i\}$. On the other hand, *multisets*, commonly denoted $\{\{i, j\}\}$, distinguish multiple occurrences. See also Fig. 3.

Example 3.5 (Exchangeable graphs, cont.). If X is a random graph, the variables U_i are associated with vertices—i.e., U_i with vertex i—and the variables $U_{\{\{i,j\}\}}$ with edges. We consider undirected graphs without self-loops. Then (X_{ij}) is symmetric, and the diagonal entries of the adjacency matrix are non-random and zero. Hence, we can neglect the diagonal variables $U_{\{\{i,i\}\}}$, and can therefore index by ordinary sets as $U_{\{i,j\}}$. Since X is binary, i.e., $X_{ij} \in \{0, 1\}$, it can be represented as follows: There is a two-argument, symmetric random function $W : [0,1]^2 \to [0,1]$ such that

$$X_{ij} \overset{\mathrm{d}}{=} F(U_i, U_j, U_{\{i,j\}}) \overset{\mathrm{d}}{=} \mathbb{I}\{U_{\{i,j\}} < W(U_i, U_j)\} \tag{3.4}$$

(where $\mathbb{I}$ denotes the indicator function). This follows directly from Eq. (3.3): For fixed values of U_i and U_j, the function $F(U_i, U_j, \,.\,)$ is defined on $[0,1]$. In the graph case, this function is binary, and takes value 1 on some set $A \subset [0,1]$ and value 0 on the complement of A. Since $U_{\{i,j\}}$ is uniform, the probability that F is 1 is simply $|A| =: W(U_i, U_j)$. The sampling scheme defined by Eq. (3.4) is visualized in Fig. 4. ◁

Theorem 3.4 is also applicable to directed graphs. However, in the directed case, (X_{ij}) is asymmetric, which changes the conditional independence structure: X_{ij} and X_{ji} are now distinct variables, but since $\{\{i,j\}\} = \{\{j,i\}\}$, the representation (3.3) implies that both are still represented by the same variable $U_{\{\{i,j\}\}}$. Thus, X_{ij} and X_{ji} are not conditionally independent.

Remark 3.6 (Non-uniform sampling schemes). The random variables U_i, U_{ij}, etc used in the representation need not be uniform. The resemblance between functions on $[0,1]^2$ and empirical graph distributions (see Fig. 4) makes the unit square convenient for purposes of exposition, but for modeling problems or sampling algorithms, we could for example choose i.i.d. Gaussian variables on $\mathbb{R}$ instead. In this case, F would be a different random function of the form $\mathbb{R}^3 \to \mathbf{X}$, rather than $[0,1]^3 \to \mathbf{X}$. More generally, any atomless probability measure on a standard Borel space can be substituted for the Uniform$[0,1]$ distribution. ◁

For separately exchangeable arrays, the Aldous-Hoover representation differs from the jointly exchangeable case:

Corollary 3.7 (Separately exchangeable matrices). *A random array $(X_{ij})_{ij\in\mathbb{N}}$ is separately exchangeable if and only if it can be represented as follows: There is a random measurable function $F : [0,1]^3 \to \mathbf{X}$ such that*

$$(X_{ij}) \overset{d}{=} (F(U_i^{row}, U_j^{col}, U_{ij}))\ , \tag{3.5}$$

where $(U_i^{row})_{i\in\mathbb{N}}$, (U_j^{col}) and $(U_{ij})_{i,j\in\mathbb{N}}$ are, respectively, two sequences and a matrix of i.i.d. Uniform$[0,1]$ *random variables.* □

Since separate exchangeability treats rows and columns independently, the single sequence (U_i) of random variables in Eq. (3.3) is replaced by two distinct sequences $(U_i^{\mathrm{row}})_{i\in\mathbb{N}}$ and $(U_j^{\mathrm{col}})_{j\in\mathbb{N}}$, respectively. Additionally, we now need an entire random matrix (U_{ij}) to account for interactions. The index structure of the uniform random variables is the only difference between the jointly and separately exchangeable case.

Example 3.8 (Collaborative filtering). In the prototypical version of a collaborative filtering problem, users assign scores to movies. Scores may be binary ("like/don't like", $X_{ij} \in \{0,1\}$), have a finite range ("one to five stars", $X_{ij} \in \{1,\dots 5\}$), etc. Separate exchangeability then simply means that the probability of seeing any particular realization of the matrix does not depend on the way in which either the users or the movies are ordered. ◁

Remark 3.9. We have stated the separately exchangeable case as a Corollary of Theorem 3.4. The implication is perhaps not obvious, and most easily explained for binary matrices: If such a matrix X is separately exchangeable, we can interpret it as a graph, but since rows and columns are separate entities, the graph has two separate sets V^{rows} and V^{cols} of vertices. Each vertex represents either a row or a column. Hence, entries of X represent edges *between* these two sets, and the graph is bipartite. If the bipartite surrogate graph satisfies Eq. (3.2) for all permutations π of $\mathbb{N}$, then it does so in particular for all permutations that affect only one of the two sets V^{rows} or V^{cols}. Hence, joint exchangeability of the bipartite graph implies separate exchangeability of the original graph. In Eq. (3.3), the two separate sets V^{rows} and V^{cols} of vertices are represented by two separate sets U_i^{row} and U_j^{col} of uniform variables. Similarly, X_{ij} and X_{ji} are represented in the bipartite graphs by two separate edges between two distinct pairs of vertices—row i and column j versus row j and column i—and hence represented by two distinct variables U_{ij} and U_{ji}, which results in Eq. (3.5). ◁

3.3. *Application to Bayesian Models.* The representation results above have fundamental implications for Bayesian modeling—in fact, they provide a general characterization of Bayesian models of array-valued data:

If array data is exchangeable (jointly or separately), any prior distribution can be represented as the distribution of a random measurable function of the form $[0,1]^3 \to [0,1]$.

More concretely, suppose we are modeling matrix-valued data represented by a random matrix X. If we can make the case that X is jointly exchangeable, Theorem 3.4 states that there is a uniquely defined distribution μ on measurable functions such that X can be generated by sampling

$$F \sim \mu \tag{3.6}$$

$$\forall i \in \mathbb{N}: \qquad U_i \sim_{\mathrm{iid}} \mathrm{Uniform}[0,1] \tag{3.7}$$

$$\forall i,j \in \mathbb{N}: \qquad U_{\{\{i,j\}\}} \sim_{\mathrm{iid}} \mathrm{Uniform}[0,1] \tag{3.8}$$

and computing X as

$$\forall i,j \in \mathbb{N}: \qquad X_{ij} := F(U_i, U_j, U_{\{\{i,j\}\}})\ . \tag{3.9}$$

Another (very useful) way to express this sampling scheme is as follows: For every measurable function $f : [0,1]^3 \to \mathbf{X}$, we define a probability distribution $\mathbf{p}(X \in \,.\,, f)$ as the distribution obtained by sampling (U_i) and $(U_{\{\{i,j\}\}})$ as in Eq. (3.7)-Eq. (3.8) and then defining

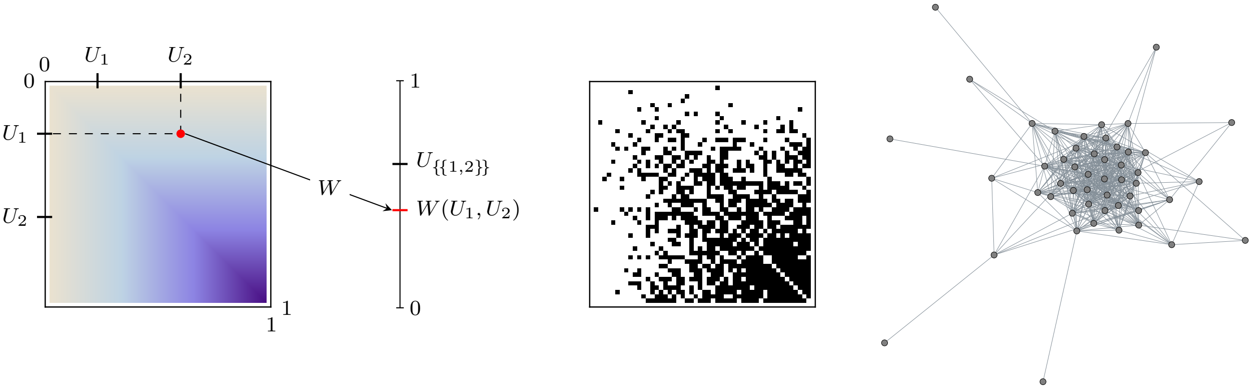

Fig 4: Sampling an exchangeable random graph according to Eq. (3.4). *Left:* An instance of the random function W, chosen here as $W = \min\{x, y\}$, as a heat map on $[0,1]^2$. In the case depicted here, the edge $(1,2)$ is not present in the graph, since $U_{\{\{1,2\}\}} > W(U_1, U_2)$. *Middle:* The adjacency matrix of a 50-vertex random graph, sampled from the function on the left. Rows in the matrix are ordered according the value, rather than the index, of U_i, resulting in a matrix resembling W. *Right:* A plot of the random graph sample. The highly connected vertices plotted in the center correspond to values lower right region in $[0,1]^2$.

$X_{ij} := f(U_i, U_j, U_{\{\{i,j\}\}})$. Thus, $\mathbf{p}(\,.\,, f)$ is a family of distributions parametrized by f, or more formally, a probability kernel. X is then sampled as

$$F \sim \mu \tag{3.10}$$
$$X|F \sim \mathbf{p}(\,.\,, F)\ . \tag{3.11}$$

In Bayesian modeling terms, μ is a prior distribution, F a parameter variable, and $\mathbf{p}$ the observation model.

If X is separately exchangeable, we similarly sample

$$F \sim \mu \tag{3.12}$$
$$\forall i \in \mathbb{N}: \qquad U_i^{\text{row}} \sim_{\text{iid}} \text{Uniform}[0,1] \tag{3.13}$$
$$\forall j \in \mathbb{N}: \qquad U_j^{\text{col}} \sim_{\text{iid}} \text{Uniform}[0,1] \tag{3.14}$$
$$\forall i,j \in \mathbb{N}: \qquad U_{ij} \sim_{\text{iid}} \text{Uniform}[0,1] \tag{3.15}$$

and set

$$\forall i,j \in \mathbb{N}: \qquad X_{ij} := F(U_i^{\text{row}}, U_j^{\text{col}}, U_{ij})\ . \tag{3.16}$$

Analogous to $\mathbf{p}$, we define a probability kernel $\mathbf{q}(X \in \,.\,, f)$ which summarizes Eq. (3.13)-Eq. (3.15), and obtain

$$F \sim \mu \tag{3.17}$$
$$X|F \sim \mathbf{q}(\,.\,, F)\ . \tag{3.18}$$

Bayesian models are usually defined by defining a prior and a sampling distribution (i.e., likelihood). We hence have to stress here that, in the representation above, the sampling distributions $\mathbf{p}$ and $\mathbf{q}$ are generic—any jointly or separately exchangeable matrix can be represented with these sampling distributions, and specifying the model is equivalent to specifying the prior, i.e., the distribution of F.

Remark 3.10 (Non-exchangeable arrays). Various types of array-valued data depend on time or some other covariate. In this case, joint or separate exchangeability can be assumed to hold marginally, as described in Section 2.3. For time-dependent graph data, for example, one would assume that joint exchangeability holds marginally at each point in time. In this case, the random mapping ξ in (2.19) becomes a time-indexed array. The random function $W(\,.\,,\,.\,)$ in Eq. (3.4) then turns into a function $W(\,.\,,\,.\,,t)$ additionally dependent on time—which raises new modeling questions, e.g., whether the stochastic process $(W(\,.\,,\,.\,,t))_t$ should be smooth. More generally, the discussion in 2.3 applies to joint and separate exchangeability just as it does to exchangeable sequences.

There is a much deeper reason why exchangeability may not be an appropriate assumption—too oversimplify, because exchangaeble models of graphs may generate too many edges—which is discussed in depth in Section 7. ◁

3.4. *Uniqueness of representations.* In the representation Eq. (3.4), random graph distributions are parametrized by measurable functions $w : [0,1]^2 \to [0,1]$. This representation is not unique, as illustrated in Fig. 5. In mathematics, the lack of uniqueness causes a range of technical difficulties. In statistics, it means that w, when regarded as a model parameter, is not identifiable. It is possible, though mathematically challenging, to treat the estimation problem up to equivalence of functions; Kallenberg [35, Theorem 4] has solved this problem for a large class of exchangeable arrays (see also [18, §4.4] for recent related work). For now, we will only explain the problem; a unique parametrizations exists, but it is based on the notion of a graph limit, and has to be postponed until Section 5.

To see that the representation by w is not unique, note that the only requirement on the random variables U_i in Theorem 3.4 is that they are uniformly distributed. Suppose we define a bijective function $\phi : [0,1] \to [0,1]$ with the property that, if U is a uniform random variable, $\phi(U)$ is still uniformly distributed. Such a mapping is called a **measure-preseving transformation** (MPT), because it preserves the uniform probability measure. Intuitively, an MPT generalizes the concept of permuting the nodes of a graph to the representation of graphs by functions on

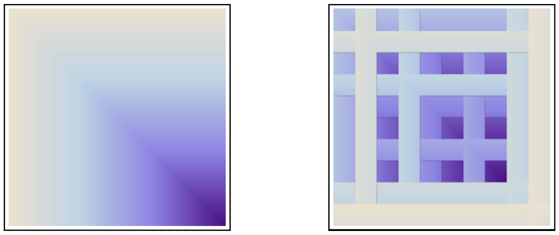

Fig 5: Non-uniqueness of representations: The function on the left parametrizes a random graph as in Fig. 4. On the right, this function has been modified by dividing the unit square into 10×10 blocks and applying the same permutation of the set $\{1, \ldots, 10\}$ simultaneously to rows and columns. Since the random variables U_i in Eq. (3.4) are i.i.d., sampling from either function defines one and the same distribution on random graphs.

a continous set. There is an infinite number of such mappings. For example, we could define ϕ by partitioning $[0,1]$ into any number of blocks, and then permute these blocks, as illustrated in Fig. 5

In the sampling procedure Eq. (3.4), we can apply ϕ simultaneously to both axes of $[0,1]^2$—formally, we apply the mapping $\phi \otimes \phi$—without changing the distribution of the resulting random graph, since the $\phi(U_i)$ are still uniform. Equivalently, we can leave the U_i untouched, and instead apply $\phi \otimes \phi$ to the function w. The resulting function $(\phi \otimes \phi) \circ w$ parametrizes the same random graph as w.

Remark 3.11 (Monotonization is not applicable). A question which often arises in this context is whether a unique representation can be defined through "monotonization": On the interval, every bounded real-valued function can be transformed into a monotone left-continuous functions by a measure-preserving transformation, and this left-continuous representation is unique [e.g. 45, Proposition A.19]. It is well known in combinatorics that the same does *not* hold on $[0,1]^2$ [15, 45]. More precisely, one might attempt to monotonize w on $[0,1]^2$ by first projecting onto the axes, i.e., by defining $w_1(x) := \int w(x,y)dy$ and $w_2(y) := \int w(x,y)dx$. The function w_1 can be transformed into a monotone representation by a unique MPT ϕ_1, and so can w_2 by ϕ_2. We could then use $(\phi_1 \otimes \phi_2) \circ w$ as a representative of w, but this approach does not yield a canonical representation: Fig. 6 shows two distinct functions w and w', which have indentical projections $w_1 = w_2 = w'_1 = w'_2$ (the constant function 1/2) and determine identical MPTs ϕ_1 and ϕ_2 (the identity map). The monotonizations of w and w' are hence again w and w', which are still distinct, even though w and w' parametrize the same graph. ◁

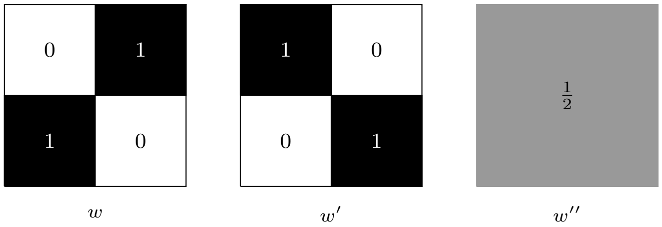

Fig 6: The functions w and w' are distinct but parametrize the same random graph (an almost surely bipartite graph). Both remain invariant and hence distinct under monotonization, which illustrates that monotonization does not yield a canonical representation (see Remark 3.11 for details). Additionally, function w'' shows that the projections do not distinguish different random graphs: w'' projects to the same constant functions as w and w', but parametrizes a different distribution (an Erdös-Renyi graph with edge probability 1/2).

4. Literature Survey.

The representation theorems show that any Bayesian model of an exchangeable array can be specified by a prior on functions. Models can therefore be classified according to the type of random function they employ. This section surveys several common categories of such random functions, including random piece-wise constant (p.w.c.) functions, which account for the structure of models built using Chinese restaurant processes, Indian buffet processes and other combinatorial stochastic processes; and random continuous functions with, e.g., Gaussian process priors. Special cases of the latter include a range of matrix factorization and dimension reduction models proposed in the machine learning literature. Table 2 summarizes the classes in terms of restrictions on the random function and the values it takes.

4.1. *Cluster-based models.*

Cluster-based models assume that the rows and columns of the random array $X := (X_{ij})$ can be partitioned into (disjoint) classes, such that the probabilistic structure between every row- and column-class is homogeneous. Within social science, this idea is captured by assumptions underlying **stochastic block models** [33, 65].

The collaborative filtering problem described in Example 3.8 is a prototypical application: here, a cluster-based model would assume that the users can be partitioned into classes/groups/types/kinds (of users), and likewise, the movies can also be partitioned into classes/groups/types/kinds (of movies). Having identified the underlying partition of users and movies, each class of user would be assumed to have a prototypical preference for each class of movie.

Because a cluster-based model is described by two partitions, this approach to modeling exchangeable arrays is closely related to clustering, and many well-known nonparametric Bayesian stochastic processes—e..g, the Dirichlet process and Pitman-Yor process, or their combinatorial counterpart, the Chinese restaurant process—are common components of cluster-based models. Indeed, we will begin by describing the Infinite Relational Model [38, 66], the canonical nonparametric, cluster-based, Bayesian model for arrays.

To our knowledge, the Infinite Relational Model, or simply IRM, was the first nonparametric Bayesian model of an exchangeable array. The IRM was introduced in 2006 independently by Kemp, Tenenbaum, Griffiths, Yamada and Ueda [38], and then by Xu, Tresp, Yu and Kriegel

Model class	Random function F	Distribution of values
Cluster-based (Section 4.1)	p.w.c. on random product partition	exchangeable
Feature-based (Section 4.2)	p.w.c. on random product partition	feature-exchangeable
Piece-wise constant (Section 4.3)	p.w.c. general random partition	arbitrary
Gaussian process-based (Section 4.4)	continuous	Gaussian

TABLE 2
Important classes of exchangeable array models. (Note that p.w.c. stands for piecewise constant.)

[66]. (Xu et al. referred to their model as the Infinite Hidden Relational Model, but we will refer to both simply by IRM.) The IRM can be seen as a nonparametric generalization of parametric stochastic block models introduced by Holland, Laskey and Leinhardt [33] and Wasserman and Anderson [65]. In the following example, we describe the model for the special case of a $\{0,1\}$-valued array.

EXAMPLE 4.1 (Infinite Relational Model). Under the IRM, the generative process for a finite subarray of binary random variables X_{ij}, $i \leq n$, $j \leq m$, is as follows: To begin, we partition the rows (and then columns) into clusters according to a **Chinese restaurant process**, or simply CRP. (See Pitman's excellent monograph [54] for a in-depth treatment of the CRP and related processes.) In particular, the first and second row are chosen to belong to the same cluster with probability proportional to 1 and to belong to different clusters with probability proportional to a parameter $c > 0$. Subsequently, each row is chosen to belong to an existing cluster with probability proportional to the current size of the cluster, and to a new cluster with probability proportional to c. Let $\Pi := \{\Pi_1, \dots, \Pi_\kappa\}$ be the random partition of $\{1, \dots, n\}$ induced by this process, where Π_1 is the cluster containing 1, and Π_2 is the cluster containing the first row not belonging to Π_1, and so on. Note that the number of clusters, κ, is also a random variable. Let $\Pi' := \{\Pi'_1, \dots, \Pi'_{\kappa'}\}$ be the random partition of $\{1, \dots, m\}$ induced by this process on the *columns*, possibly with a different parameter $c' > 0$ determining the probability of creating new clusters. Next, for every pair (k, k') of cluster indices, $k \leq \kappa$, $k' \leq \kappa'$, we generate an independent beta random variable $\theta_{k,k'}$.[1] Finally, we generate each X_{ij} independently from a Bernoulli distribution with mean $\theta_{k,k'}$, where $i \in \Pi_k$ and $j \in \Pi'_{k'}$. As we can see, $\theta_{k,k'}$ represents the probability of links arising between elements in clusters k and k'.

The Chinese restaurant process (CRP) generating Π and Π' is known to be exchangeable in the sense that the distribution of Π is invariant to a permutation of the underlying set $\{1, \dots, n\}$. It is then straightforward to see that the distribution on the subarray is exchangeable. In addition, it is straightforward to verify that, were we to have generated an $n+1 \times m+1$ array, the marginal distribution on the $n \times m$ subarray would have agreed with that of the above process. This implies that we have defined a so-called projective family and so results from probability theory imply that there exists an infinite array and that the above process described every finite subarray. ◁

The IRM model can be seen to be a special case of exchangeable arrays that we will call **cluster-based**. We will define this class formally, and then return to the IRM example, re-describing it in this new language where the exchangeability is manifest. To begin, we first introduce a subclass of cluster-based models, called **simple cluster-based** models:

DEFINITION 4.2. We say that a Bayesian model of an exchangeable array is *simple cluster-based* when, for some random function F representing X, there are random partitions $B_1, B_2, \dots$ and $C_1, C_2, \dots$ of the unit interval $[0,1]$ such that:

1. On each block $A_{i,j} := B_i \times C_j \times [0,1]$, F is constant. Let f_{ij} be the value F takes on block $A_{i,j}$.
2. The block values (f_{ij}) are themselves an exchangeable array, and independent from (B_i) and (C_j).

We call an array simple cluster-based if its distribution is.[2] ◁

Most examples of simple cluster-based models in the literature—including, e.g., the IRM—take the block values f_{ij} to be conditionally i.i.d. (and so the array (f_{ij}) is then trivially exchangeable). As an example of a more flexible model for (f_{ij}), which is merely exchangeable, consider the following:

EXAMPLE 4.3 (exchangeable link probabilities). For every block i in the row partition, let u_i be an independent and identically distributed Gaussian random variable. Similarly, let (v_j) be an i.i.d. sequence of Gaussian random variables for the column partitions. Then, for every row and column block i, j, put $f_{ij} := \mathrm{sig}(u_i + v_j)$, where $\mathrm{sig}\colon \mathbb{R} \to [0,1]$ is a sigmoid function. The array (f_{ij}) is obviously exchangeable. ◁

Like with cluster-based models of exchangeable sequences, if the number of classes in each partition is bounded, then a simple cluster-based model of an exchangeable array is a mixture of a finite-dimensional family of ergodic distributions. Therefore, mixtures of an infinite-dimensional family must place positive mass on partitions with arbitrarily many classes.

[1] For simplicity, assume that we fix the hyperparameters of the beta distribution, although this assumption can be relaxed if one is careful not to break exchangeability or projectivity.

[2] Those familiar with the theory of exchangeable partitions might note that our model does not allow for singleton blocks (aka *dust*). This is a straightforward generalization, but complicates the presentation.

In order to define the more general class of cluster-based models, we relax the piecewise constant nature of the random function. In particular, we will construct an exchangeable array (X_{ij}) from a corresponding array (θ_{ij}) of parameters, which will have a simple cluster-based model. The parameter θ_{ij} could, e.g., determine the probability of an interaction $X_{ij} \in \{0,1\}$. More generally, the parameters index a family of distributions on $\mathbf{X}$.

To precisely define such models, we adapt the notion of a *randomization* from probability theory [36]. Intuitively, given a random variable θ_i in $\mathbf{T}$ and a probability kernel P from $\mathbf{T}$ to $\mathbf{X}$, we can generate a random variable Y_i from $P(.,\theta_i)$. The following definition generalizes this idea to an indexed collection of random variables.

Definition 4.4 (randomization). Let $\mathbf{T}$ be a parameter space, let P be a probability kernel from $\mathbf{T}$ to $\mathbf{X}$, and let $\theta := (\theta_i : i \in I)$ be a collection of random variables taking values in $\mathbf{T}$, indexed by elements of a set I. (E.g., $I = \mathbb{N}^2$) We say that a collection $Y := (Y_i : i \in I)$ of random variables, indexed by the same set I, is a **P-randomization of θ** when the elements Y_i are conditionally independent given θ, and

$$\forall i \in I: \qquad Y_i \mid \theta \sim P(\,.\,,\theta_i). \tag{4.1}$$

◁

Thus, a generative model for the collection Y is to first generate θ, and then, for each $i \in I$, to sample Y_i independently from the distribution $P(\,.\,,\theta_i)$. It is straightforward to prove that, if θ is an exchangeable array and Y is a randomization of θ, then Y is exchangeable. We may therefore define:

Definition 4.5 (cluster-based models). We say that a Bayesian model for an exchangeable array $X := (X_{ij})$ in $\mathbf{X}$ is **cluster-based** when X is a P-randomization of a simple cluster-based exchangeable array $\theta := (\theta_{ij})$ taking values in a space $\mathbf{T}$, for some probability kernel P from $\mathbf{T}$ to $\mathbf{X}$. We say an array is cluster-based when its distribution is. ◁

The intuition is that the cluster membership of two individuals i, j determines a distribution, parametrized by θ_{ij}. The actual observed relationship X_{ij} is then a sample from this distribution.

Let X, θ, $\mathbf{T}$ and P be defined as above. We may characterize the random function F for X as follows: Let $\phi : \mathbf{T}\times[0,1] \to \mathbf{X}$ be such that $\phi(t,U)$ is $P(\,.\,,t)$-distributed for every $t \in \mathbf{T}$, when U is uniformly distributed in $[0,1]$. (Such a function ϕ is sometimes called a **sampling function**.) Then, if G is the random function representing the exchangeable array (θ_{ij}) then

$$F(x,y,z) = \phi(G(x,y,z),z) \tag{4.2}$$

is a function representing X. (Recall that $G(x,y,z) = G(x,y,z')$ for almost all x,y,z,z' by part 1 of Definition 4.2.)

The next example describes a model which generates the random partitions using a Dirichlet process.

Example 4.6 (Infinite Relational Model continued). We may alternatively describe the IRM distribution on exchangeable arrays as follows: Let P be a probability kernel from $\mathbf{T}$ to $\mathbf{X}$ (e.g., a Bernoulli likelihood mapping $[0,1]$ to distributions on $\{0,1\}$) and let H be a prior distribution on the parameter space $[0,1]$ (e.g., a Beta distribution, so as to achieve conjugacy). The IRM model of an array $X := (X_{ij})$ is cluster-based, and in particular, is a P-randomization of a simple, cluster-based exchangeable array $\theta := (\theta_{ij})$ of parameters in $\mathbf{T}$.

In order to describe the structure of θ, it suffices to describe the distribution of the partitions (B_k) and (C_k) as well as that of the block values. For the latter, the IRM simply chooses the block values to be i.i.d. draws from the distribution H. (While the block values can be taken to be merely exchangeable, we have not seen this generalization in the literature.) For the partitions, the IRM utilizes the stick-breaking construction of a Dirichlet process [59].

In particular, let $W_1, W_2, \dots$ be an i.i.d. sequence of $\mathrm{Beta}(1,\alpha)$ random variables, for some concentration parameter $\alpha > 0$. For every $k \in \mathbb{N}$, we then define

$$P_k := (1-W_1)\cdots(1-W_{k-1})W_k. \tag{4.3}$$

With probability one, we have $P_k \geq 0$ for every $k \in \mathbb{N}$ and $\sum_{k=1}^{\infty} P_k = 1$ almost surely, and so the sequence (P_k) characterizes a (random) probability distribution on $\mathbb{N}$. We then let (B_k) be a sequence of contiguous intervals that partition of $[0,1]$, where B_k is the half-open interval of length P_k. In the jointly exchangeable case, the random partition (C_k) is usually chosen either as a copy of (B_k), or as partition sampled independently from the same distribution as (B_k).

The underlying discrete partitioning of G induces a partition on the rows and columns of the array under the IRM model. In the IRM papers themselves, the clustering of rows and columns is described directly in terms of a Chinese restaurant process (CRP) as we did in the first IRM example, rather than in terms of an explicit list of probabilities. To connect the random probabilities (P_k) for the rows with the CRP, note that P_k is the limiting fraction of rows in the kth cluster Π_k as the number of rows tends to infinity. ◁

4.2. *Feature-based models.* Feature-based models of exchangeable arrays have similar structure to cluster-based models. Like cluster-based models, feature-based models partition the rows and columns into clusters, but unlike cluster-based models, feature-based models allow the rows and columns to belong to multiple clusters simultaneously. The set of clusters that a row belongs to are then called its **features**. The interaction between row i and column j is then determined by the features that the row and column possess.

The stochastic process at the heart of most existing feature-based models of exchangeable arrays is the Indian

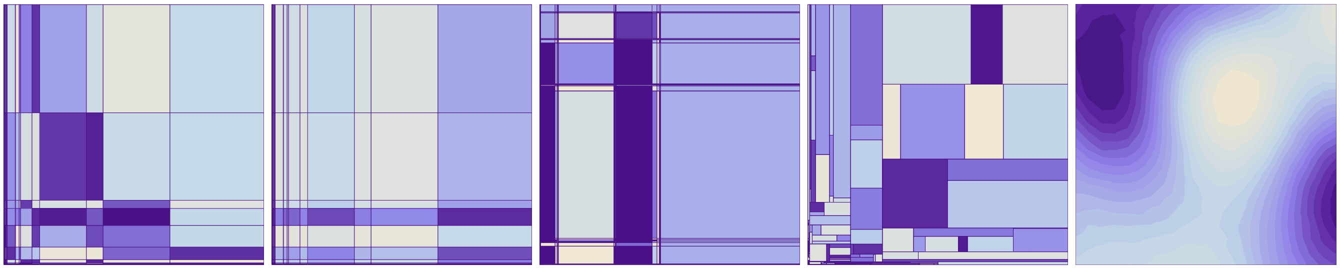

Fig 7: Typical directing random functions underlying, from left to right, 1) an IRM (where partitions correspond with a Chinese restaurant process) with conditionally i.i.d. link probabilities; 2) a more flexible variant of the IRM with merely *exchangeable* link probabilities as in Example 4.3; 3) a LFRM (where partitions correspond with an Indian buffet process) with feature-exchangeable link probabilities as in Example 4.10; 4) a Mondrian-process-based model with a single latent dimension; 5) a Gaussian-processed-based model with a single latent dimension. (Note that, in practice, one would use more than one latent dimension in the last two examples, although this complicates visualization. In the first four figures, we have truncated each of the "stick-breaking" constructions at a finite depth, although, at the resolution of the figures, it is very difficult to notice the effect.)

buffet process, introduced by Griffiths and Ghahramani [28]. The Indian buffet process (IBP) produces an allocation of features in a sequential fashion, much like the Chinese restaurant process produces a partition in a sequential fashion. In the follow example, we will describe the Latent Feature Relational Model (LFRM) of Miller et al. [49], one of the first nonparametric, feature-based models of exchangeable arrays. For simplicity, we will describe the special case of a $\{0,1\}$-valued, separately-exchangeable array.

Example 4.7 (Latent Feature Relational Model). Under the LFRM, the generative process for a finite subarray of binary random variables X_{ij}, $i \le n$, $j \le m$, is as follows: To begin, we allocate features to the rows (and then columns) according to an IBP. In particular, the first row is allocated a Poisson number of features, with mean $\gamma > 0$. Each subsequent row will, in general, share some features with earlier rows, and possess some features not possessed by any earlier row. Specifically, the second row is also allocated a Poisson number of altogether new features, but with mean $\gamma/2$, and, for every feature possessed by the first row, the second row is allocated that feature, independently, with probability $1/2$. In general, the kth row: is allocated a Poisson number of altogether new features, with mean γ/k; and, for every subset $K \subseteq \{1,\dots,k-1\}$ of the previous rows, and every feature possessed by exactly those rows in K, is allocated that feature, independently, with probability $|K|/n$. (We use the same process to allocate a distinct set of features to the m columns, though potentially with a different constant $\gamma' > 0$ governing the overall number of features.)

We now describe how the features possessed by the rows and columns come to generate the observed subarray. First, we number the row- and column- features arbitrarily, and for every row i and column j, we let $N_i, M_j \subseteq \mathbb{N}$ be the set of features they possess, respectively. For every pair (k,k') of a row- and column- feature, we generate an independent and identically distributed Gaussian random variable $w_{k,k'}$. Finally, we generate each $X_{i,j}$ independently from a Bernoulli distribution with mean $\mathrm{sig}(\sum_{k\in N_i}\sum_{k'\in M_j} w_{k,k'})$. Thus a row and column that possess feature k and k', respectively, have an increased probability of a connection as $w_{k,k'}$ becomes large and positive, and a decreased probability as $w_{k,k'}$ becomes large and negative.

The exchangeability of the subarray follows from the exchangeability of the IBP itself. In particular, define the family of counts Π_N, $N \subseteq \{1,\dots,n\}$, where Π_N is the number of features possessed by exactly those row in N. We say that $\Pi := (\Pi_N)$ is a **random feature allocation** for $\{1,\dots,n\}$. (Let Π' be the random feature allocation for the columns induced by the IBP.) The IBP is exchangeable is the sense that

$$(\Pi_N) \overset{\mathrm{d}}{=} (\Pi_{\sigma(N)}) \tag{4.4}$$

for every permutation π of $\{1,\dots,n\}$, where $\sigma(N) := \{\sigma(n) : n \in N\}$. Moreover, the conditional distribution of the subarray given the feature assignments (N_i, M_j) is the same as the conditional distribution given the feature allocations (Π_N, Π'_M). It is then straightforward to verify that the subarray is itself exchangeable. Like with the IRM example, the family of distributions on subarrays of different sizes is projective, and so there exists an infinite array and the above process describes the distribution of every subarray. $\triangleleft$

We will cast the LFRM model as a special case of a class of models that we will call **feature-based**. From the perspective of simple cluster-based models, simple feature-based models also have a block structured representing function, but relax the assumption that values of each block form an exchangeable array. To state the definition of this class more formally, we begin by generalizing the notion of a partition of $[0,1]$. (See [16] for recent work characterizing exchangeable feature allocations.)

Definition 4.8 (feature allocation). Let U be a uniformly-distributed random variable and $E := (E_1, E_2, \dots)$ a sequence of (measurable) subsets of $[0,1]$. Given E, we say that U has feature n when $U \in E_n$. We call the sequence E a **feature allocation** if

$$\mathbb{P}\Big\{U \notin \textstyle\bigcup_{k\ge n} E_k\Big\} \to 1 \quad \text{as} \quad n \to \infty. \tag{4.5}$$

◁

The definition probably warrants some further explanation: A partition is a special case of a feature allocation, in which the sets E_n are disjoint and represent blocks of a partition. The relation $U \in E_k$ then indicates that an object represented by the random variable U is in block k of the partition. In a feature allocation, the sets E_k may overlap. The relation $U \in E_n$ now indicates that the object has feature n. Because the sets may overlap, the object may possess multiple features. However, condition Eq. (4.5) ensures that the number of features per object remains finite (with probability 1).

A feature allocation induces a partition if we equate any two objects that possess exactly the same features. More carefully, for every subset $N \subset \mathbb{N}$ of features, define

$$E_{(N)} := \bigcap_{i \in N} E_i \cap \bigcap_{j \notin N} ([0,1] \setminus E_j) \,. \tag{4.6}$$

Then, two objects represented by random variables U and U' are equivalent iff $U, U' \in E_{(N)}$ for some finite set $N \subset \mathbb{N}$. As before, we could consider a simple, cluster-based representing function where the block values are given by an $(f_{N,M})$, indexed now by finite subsets $N, M \subseteq \mathbb{N}$. Then $f_{N,M}$ would determine how two objects relate when they possess features N and M, respectively.

However, if we want to capture the idea that the relationships between objects depend on the individual features the objects possess, we would not want to assume that the entries of $f_{N,M}$ formed an exchangeable array, as in the case of a simple, cluster-based model. E.g., we might choose to induce more dependence between $f_{N,M}$ and $f_{N',M}$ when $N \cap N' \neq \emptyset$ than otherwise. The following definition captures the appropriate relaxation of exchangeability:

Definition 4.9 (feature-exchangeable array). Let $Y := (Y_{N,M})$ be an array of random variables indexed by pairs $N, M \subseteq \mathbb{N}$ of finite subsets. For a permutation π of $\mathbb{N}$ and $N \subseteq \mathbb{N}$, write $\pi(N) := \{\pi(n) : n \in N\}$ for the image. Then, we say that Y is **feature-exchangeable** when

$$(Y_{N,M}) \overset{\mathrm{d}}{=} (Y_{\pi(N),\pi(M)}), \tag{4.7}$$

for all permutations π of $\mathbb{N}$. ◁

Informally, an array Y indexed by sets of features is feature-exchangeable if its distribution is invariant to permutations of the underlying feature labels (i.e., of $\mathbb{N}$). The following is an example of a feature-exchangeable array, which we will use when we re-describe the Latent Feature Relational Model in the language of feature-based models:

Example 4.10 (feature-exchangeable link probabilities). Let $w := (w_{ij})$ be a conditionally i.i.d. array of random variables in $\mathbb{R}$, and define $\theta := (\theta_{N,M})$ by

$$\theta_{N,M} = \mathrm{sig}(\textstyle\sum_{i \in N} \sum_{j \in M} w_{ij}), \tag{4.8}$$

where $\mathrm{sig}\colon \mathbb{R} \to [0,1]$ maps real values to probabilities via, e.g., the sigmoid or probit functions. It is straightforward to verify that θ is feature-exchangeable. ◁

We can now define simple feature-based models:

Definition 4.11. We say that a Bayesian model of an exchangeable array X is **simple feature-based** when, for some random function F representing X, there are random feature allocations B and C of the unit interval $[0,1]$ such that, for every pair $N, M \subseteq \mathbb{N}$ of finite subsets, F takes the constant value $f_{N,M}$ on the block

$$A_{N,M} := B_{(N)} \times C_{(M)} \times [0,1], \tag{4.9}$$

and the values $f := (f_{N,M})$ themselves form a feature-exchangeable array, independent of B and C. We say an array is simple feature-based if its distribution is. ◁

We can relate this definition back to cluster-based models by pointing out that simple feature-based arrays are simple cluster-based arrays when either i) the feature allocations are partitions or ii) the array f is exchangeable. The latter case highlights the fact that feature-based arrays relax the exchangeability assumption of the underlying block values.

As in the case of simple cluster-based models, nonparametric simple feature-based models will place positive mass on feature allocations with an arbitrary number of distinct sets. As we did with general cluster-based models, we will define general feature-based models as randomizations of simple models:

Definition 4.12 (feature-based models). We say that a Bayesian model for an exchangeable array $X := (X_{ij})$ in $\mathbf{X}$ is **feature-based** when X is a P-randomization of a simple, feature-based, exchangeable array $\theta := (\theta_{ij})$ taking values in a space T, for some probability kernel P from T to $\mathbf{X}$. We say an array is feature-based when its distribution is. ◁

Comparing Definitions 4.5 and 4.12, we see that the relationship between random functions representing θ and X are the same as with cluster-based models. We now return to the LFRM model, and describe it in the language of feature-based models:

Example 4.13 (Latent Feature Relational Model continued). The random feature allocations underlying the LFRM can be described in terms of so-called "stick-breaking" constructions of the Indian buffet process. One of the simplest stick-breaking constructions, and the one we will use here, is due to Teh, Görür, and Ghahramani [61]. (See also [63], [52] and [53].)

Let $W_1, W_2, \dots$ be an i.i.d. sequence of $\mathrm{Beta}(\alpha, 1)$ random variables for some concentration parameter $\alpha > 0$. For every n, we define $P_n := \prod_{j=1}^{n} W_j$. (The relationship between this construction and Eq. (4.3) highlights one of several relationships between the IBP and CRP.) It follows that we have $1 \geq P_1 \geq P_2 \geq \cdots \geq 0$. The allocation of features then proceeds as follows: for every $n \in \mathbb{N}$, we assign the feature with probability P_n, independently of all other features. It can be shown that $\sum_n P_n$ is finite with probability one, and so every object has a finite number of features with probability one.

We can describe a feature allocation (B_n) corresponding with this stick-breaking construction of the IBP as follows: Put $B_1 = [0, P_1)$, and then inductively, for every $n \in \mathbb{N}$, put

$$B_{n+1} := \bigcup_{j=1}^{2^n-1} [b_j, (b_{j+1} - b_j) \cdot P_{n+1}) \tag{4.10}$$

where $B_n = [b_1, b_2) \cup [b_3, b_4) \cup \cdots \cup [b_{2^n-1}, b_{2^n})$. (As one can see, this representation obscures the conditional independence inherent in the feature allocation induced by the IBP.)

Having described the distribution of the random feature allocations underlying the LFRM model, it suffices to specify the distribution of the underlying feature-exchangeable array and the probability kernel P of the randomization. The latter is simply the map $p \mapsto \text{Bernoulli}(p)$ taking a probability to the Bernoulli distribution, and the former is the feature-exchangeable array of link probabilities described in Example 4.10. ◁

4.3. *Piece-wise constant models.* Simple partition- and feature-based models have piecewise-constant structure, which arises because both types of models posit prototypical relationships on the basis of a *discrete* set of classes or features assignments, respectively. More concretely, a partition of $[0,1]^3$ is induced by partitions of $[0,1]$.

An alternative approach is to consider partitions of $[0,1]^3$ directly, or partitions of $[0,1]^3$ induced by partitions of $[0,1]^2$. Rather than attempting a definition capturing a large, natural class of such models, we present an illustrative example:

Example 4.14 (Mondrian-process-based models [57]). A Mondrian process is a partition-valued stochastic process introduced by Roy and Teh [57]. (See also Roy [56, Chp. V] for a formal treatment.) More specifically, a **homogeneous Mondrian process on** $[0,1]^2$ is a continuous-time Markov chain $(M_t : t \geq 0)$, where, for every time $t \geq 0$, M_t is a floorplan-partition of $[0,1]^2$—i.e., a partition of $[0,1]^2$ comprised of axis-aligned rectangles of the form $A = B \times C$, for intervals $B, C \subseteq [0,1]$. It is assumed that M_0 is the trivial partition containing a single class.

Every continuous-time Markov chain is characterized by the mean waiting times between jumps and the discrete-time Markov process of jumps (i.e., the *jump chain*) embedded in the continuous-time chain. In the case of a Mondrian process, the mean waiting time from a partition composed of a finite set of rectangles $\{B_1 \times C_1, \ldots, B_k \times C_k\}$ is $\sum_{j=1}^k (|B_j| + |C_j|)$. The jump chain of the Mondrian process is entirely characterized by its transition probability kernel, which is defined as follows: From a partition $\{B_1 \times C_1, \ldots, B_k \times C_k\}$ of $[0,1]^2$, we choose to "cut" exactly one rectangle, say $B_j \times C_j$, with probability proportional to $|B_j| + |C_j|$; Choosing j, we then cut the rectangle vertically with probability proportional to $|C_j|$ and horizontally with probability proportional to $|B_j|$; Assuming the cut is horizontal, we partition B_j into two intervals $B_{j,1}$ and $B_{j,2}$, uniformly at random; The jump chain then transitions to the partition where $B_j \times C_j$ is replaced by $B_{j,1} \times C_j$ and $B_{j,2} \times C_j$; The analogous transformation occurs in the vertical case.

As is plain to see, each partition is produced by a sequence of cuts that hierarchically partition the space. The types of floorplan partitions of this form are called **guillotine partitions**. Guillotine partitions are precisely the partitions represented by kd-trees, the classical data structure used to represent hierarchical, axis-aligned partitions.

The Mondrian process possesses several invariances that allow one to define a Mondrian process M_t^* on all of $\mathbb{R}^2$. The resulting process is no longer a continuous-time Markov chain. In particular, for all $t > 0$, M_t^* has a countably infinite number of classes with probability one. Roy and Teh [57] use this extended process to produce a nonparametric prior on random functions as follows:

Let $\phi : (0,1] \to \mathbb{R}$ be the embedding $\phi(x) = -\log x$, let M be a Mondrian process on $\mathbb{R}^2$, and let (A_n) be the countable set of rectangles comprising the partition of $\mathbb{R}^2$ given by M_c for some constant $c > 0$. A random function $F : [0,1]^3 \to [0,1]$ is then defined by $F(x,y,z) = \psi_n$ where n is such that $A_n \ni (\phi(x), \phi(y))$, and where (ψ_n) is an exchangeable sequence of random variables in $\mathbf{X}$, independent of M. As usual, one generally considers a randomization. In particular, Roy and Teh present results in the case where the ψ_n are Beta random variables, and the data are modeled via a Bernoulli likelihood. An interesting property of the above construction is that the partition structure along any axis-aligned slice of the random function agrees with the stick-breaking construction of the Dirichlet process, presented in the IRM model example. (See [57] and [56] for more details.) ◁

4.4. *Gaussian-process-based models.* Up until now, we have discussed classes of models for exchangeable arrays whose random functions have piece-wise constant structure. In this section we briefly discuss a large and important class of models that relax this restriction by modeling the random function as a Gaussian process.

We begin by recalling the definition of a Gaussian process [e.g. 55]. Let $G := (G_i : i \in I)$ be an indexed collection of $\mathbb{R}$-valued random variables. We say that G is a **Gaussian process on** I when, for all finite sequences of indices $i_1, \ldots, i_k \in I$, the vector $(G(i_1), \ldots, G(i_k))$ is Gaussian, where we have written $G(i) := G_i$ for notational convenience. A Gaussian process is completely specified by two function-valued parameters: a **mean function** $\mu : I \to \mathbb{R}$, satisfying

$$\mu(i) = \mathbb{E}\big(G(i)\big), \quad i \in I, \tag{4.11}$$

and a positive semidefinite **covariance function** $\kappa : I \times$

$I \to \mathbb{R}_+$, satisfying

$$\kappa(i,j) = \operatorname{cov}(G(i), G(j)). \qquad (4.12)$$

Definition 4.15 (Gaussian-process-based exchangeable arrays). We say that a Bayesian model for an exchangeable array $X := (X_{ij})$ in $\mathbf{X}$ is **Gaussian-process-based** when, for some random function F representing X, the process $F = (F_{x,y,z};\ x,y,z \in [0,1])$ is Gaussian on $[0,1]^3$. We will say that an array X is Gaussian-process-based when its distribution is. ◁

In the language of Eq. (3.17), a Gaussian-process-based model is one where a Gaussian process prior is placed on the random function F. The definition is stated in terms of the space $[0,1]^3$ as domain of the uniform random variables U to match our statement of the Aldous-Hoover theorem and of previous models. In the case of Gaussian processes, however, it is arguably more natural to use the real line instead of $[0,1]$, and we note that this is indeed possible: Given an embedding $\phi : [0,1]^3 \to J$ and a Gaussian process G on J, the process G' on $[0,1]^3$ given by $G'_{x,y,z} = G_{\phi(x,y,z)}$ is Gaussian. More specifically, if the former has a mean function μ and covariance function κ, then the latter has mean $\mu \circ \phi$ and covariance $\kappa \circ (\phi \otimes \phi)$. We can therefore talk about Gaussian processes on spaces J that can be put into correspondence with the unit interval. Note that the particular embedding also induces a distribution on the J.

The above definition also implies that the array X is conditionally Gaussian, ruling out, e.g., the possibility of $\{0,1\}$-valued arrays. This restriction is overcome by considering randomizations of Gaussian-process-based arrays. Indeed, in the $\{0,1\}$-valued case, the most common type of randomization can be described as follows:

Definition 4.16 (noisy sigmoidal/probit likelihood). For every mean $m \in \mathbb{R}$, variance $v \in \mathbb{R}_+$, and sigmoidal function $\sigma : \mathbb{R} \to [0,1]$, we can construct a probability kernel L from $\mathbb{R}$ to $\{0,1\}$ as follows: for each real $r \in \mathbb{R}$, let $L(r)$ be the distribution of Bernoulli random variable with mean $\mathbb{E}\big(\sigma(r+\xi)\big)$, where ξ is itself Gaussian with mean m and variance v. ◁

Many of the most popular parametric models for exchangeable arrays of random variables can be constructed as (randomizations of) Gaussian-process-based arrays. For a catalog of such models and several nonparametric variants, as well as their covariance functions, see [43]. Here we will focus on the parametric **eigenmodel**, introduced by Hoff [31, 32], and its nonparametric cousin, introduced Xu, Yan and Qi [67]. To simplify the presentation, we will consider the case of a $\{0,1\}$-valued array.

Example 4.17 (Eigenmodel [31, 32]). In the case of a $\{0,1\}$-valued array, both the eigenmodel and its nonparametric extension can be interpreted as an L-randomizations of a Gaussian-process-based array $\theta := (\theta_{ij})$, where L is given as in Definition 4.16 for some mean, variance and sigmoid. To complete the description, we define the Gaussian processes underlying θ.

The eigenmodel is best understood in terms of a zero-mean Gaussian process G on $\mathbb{R}^d \times \mathbb{R}^d$. (The corresponding embedding $\phi : [0,1]^3 \to \mathbb{R}^d \times \mathbb{R}^d$ is $\phi(x,y,z) = \Phi^{-1}(x)\Phi^{-1}(y)$, where Φ^{-1} is defined so that $\Phi^{-1}(U) \in \mathbb{R}^d$ is a vector independent doubly-exponential (aka Laplacian) random variables, when U is uniformly distributed in $[0,1]$.) The covariance function $\kappa : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$ of the Gaussian process G underlying the eigenmodel is simply

$$\kappa(u,v;x,y) = \langle u,x\rangle\langle v,y\rangle, \quad u,v,x,y \in \mathbb{R}^d, \qquad (4.13)$$

where $\langle .,.\rangle \colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ denotes the dot product, i.e., Euclidean inner product. This corresponds with a more direct description of G: in particular,

$$G(x,y) = \langle x,y\rangle_\Lambda \qquad (4.14)$$

where $\Lambda \in \mathbb{R}^{d\times d}$ is a $d \times d$ array of independent standard Gaussian random variables and $\langle x,y\rangle_A = \sum_{n,m} x_n y_m A_{n,m}$ is an inner product. ◁

A nonparametric counterpart to the eigenmodel was introduced by Xu *et al.* [67]:

Example 4.18. The Infinite Tucker Decomposition model [67] defines the covariance function on $\mathbb{R}^d \times \mathbb{R}^d$ to be

$$\kappa(u,v;x,y) = \kappa'(u,x)\kappa'(v,y), \quad u,v,x,y \in \mathbb{R}^d, \qquad (4.15)$$

where $\kappa' : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is some positive semi-definite covariance function on $\mathbb{R}^d$. This change can be understood as generalizing the inner product in Eq. (4.13) from $\mathbb{R}^d$ to a (potentially, infinite-dimensional) reproducing kernel Hilbert space (RKHS). In particular, for every such κ', there is an RKHS $\mathcal{H}$ such that

$$\kappa'(x,y) = \langle \phi(x), \phi(y)\rangle_{\mathcal{H}}, \quad x,y \in \mathbb{R}^d. \qquad (4.16)$$

◁

A related nonparametric model for exchangeable arrays, which places fewer restrictions on the covariance structure and is derived directly from the Aldous-Hoover representation, is described in [43].

5. Limits of graphs. *We have already noted that the parametrization of random arrays by functions in the Aldous-Hoover theorem is not unique. Our statement of the theorem also lacks an asymptotic convergence result such as the convergence of the empirical measure in de Finetti's theorem. The tools to fill these gaps have only recently become available in a new branch of combinatorics which studies objects known as graph limits. This section summarizes a few elementary notions of this rapidly*

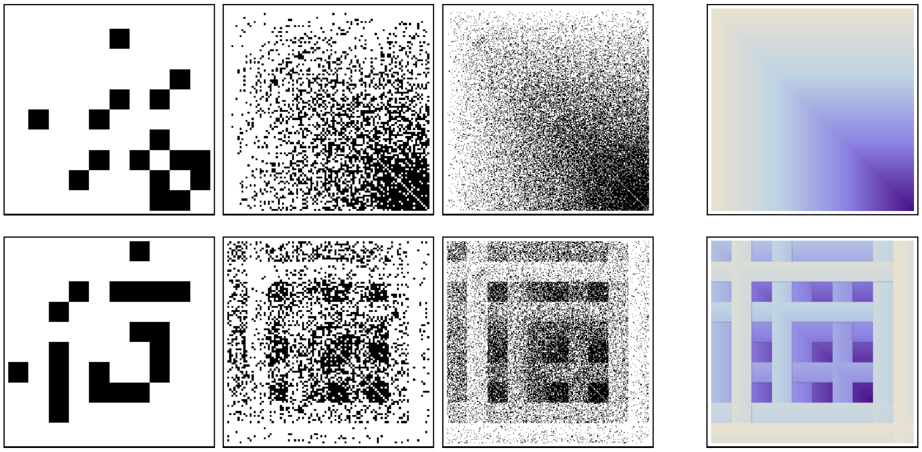

Fig 8: For graph-valued data, the directing random function F in the Aldous-Hoover representation can be regarded as a limit of adjacency matrices: The adjacency matrix of a graph of size n can be represented as a function on $[0,1]^2$ by dividing the square into $n \times n$ patches of equal size. On each patch, the representing function is constant, with value equal to the corresponding entry of the adjacency matrix. (In the figure, a black patch indicates a value of one and hence the presence of an edge.) As the size of the graph increases, the subdivision becomes finer, and converges to the function depicted on the right for $n \to \infty$. Convergence is illustrated here for the two functions from Fig. 5. Since the functions are equivalent, the two random graphs within each column are equal in distribution.

emerging field and shows how they apply to the Aldous-Hoover theorem for graphs.

Graph limit theory is based on a simple idea: Given a finite graph with n vertices, we subdivide $[0,1]^2$ into $n \times n$ square patches, resembling the $n \times n$ adjacency matrix. We then define a function w_n with constant value 0 or 1 on each patch, equal to the corresponding entry of the adjacency matrix. A plot of w_n is a checkerboard image as in Fig. 8. If we increase the size n of the graph, the resulting functions w_n are defined on finer and finer subdivisions of $[0,1]^2$, and it is not hard to imagine that they converge to a (possibly smooth) function $w : [0,1]^2 \to [0,1]$ as $n \to \infty$. This function is interpreted as the limit of the graph sequence $(g_n)_{n\in\mathbb{N}}$. There are two important ways to give a precise definition of this notion of convergence, and we will briefly discuss both definitions and some of their consequences.

5.1. *Metric definition of convergence.* The technically most convenient way to define convergence is, whenever possible, using a metric: If d is a distance measure, we can define w as the limit of w_n if $d(w, w_n) \to 0$ as $n \to \infty$. The metric on functions which has emerged as the "right" choice for graph convergence is called the **cut metric**, and is defined as follows: We first define a norm as

$$\|w\|_{\square} := \sup_{S,T\subset[0,1]} \int_{S\times T} w(x,y)d\mu(x)d\mu(y)\ . \tag{5.1}$$

The measure μ in the integral is Lebesgue measure $[0,1]$, i.e., the distribution of the uniform variables U_i in Eq. (3.3). S and T are arbitrary measurable sets. Intuitively—if we assume for the moment that w can indeed be thought of as a limiting adjacency matrix—S and T are subsets of nodes. The integral (5.1) measures the total number of edges between S and T in the "graph" w. Since a partition of the vertices of a graph into two sets is called a **cut**, $\|\,.\,\|_{\square}$ is called the **cut norm**. The distance measure defined by $d_{\square}(w,w') := \|w - w'\|_{\square}$ is called the **cut distance**.

Suppose w and w' are two distinct functions which parametrize the same random graph. The distance $d_{\square}$ in general perceives such functions as different: The functions in Fig. 8, for instance, define the same graph, but have non-zero distance under $d_{\square}$. Hence, if we were to use $d_{\square}$ to define convergence, the two sequences of graphs in the figure would converge to two different limits. We therefore modify $d_{\square}$ by defininig:

$$\delta_{\square}(w,w') := \inf_{\phi\ \mathrm{MPT}} d_{\square}(w, w' \circ (\phi \otimes \phi))\ . \tag{5.2}$$

MPT is the set of measure-preserving transformations (see Section 3.4). In words, before we measure the distance between w and w' using $d_{\square}$, we push w' through the MPT that best aligns w' to w. In Fig. 5, this optimal ϕ would simply be the mapping which reverses the permutation of blocks, so that the two functions would look identical.

Definition 5.1. We say that a sequence $(g_n)_{n\in\mathbb{N}}$ of graphs converges if $\delta_{\square}(w_{g_n}, w) \to 0$ for some measurable function $w : [0,1]^2 \to [0,1]$. The function w is called the limit of (g_n), and often referred to as a **graph limit** or **graphon**. ◁

The function $\delta_{\square}$ is called the **cut pseudometric**: It is not an actual metric, since it can take value 0 for two distinct functions. It does, however, have all other properties of a metric. By definition, $\delta_{\square}(w,w') = 0$ nolds if and only if w and w' parametrize the same random graph.

The properties of $\delta_{\square}$ motivate the definition of a "quotient space": We begin with the space $\mathbf{W}$ of all graphons, i.e., all measurable functions $[0,1]^2 \to [0,1]$, and regard two functions w, w' as equivalent if $\delta_{\square}(w, w') = 0$. The equivalence classes form a partition of $\mathbf{W}$. We then define a new space $\widehat{\mathbf{W}}$ by collapsing each equivalence class to a single point. Each element $\widehat{w} \in \widehat{\mathbf{W}}$ corresponds to all functions in one equivalence class, and hence to one specific random graph distribution. The pseudometric $\delta_{\square}$ turns into a metric on $\widehat{\mathbf{W}}$. The metric space $(\widehat{\mathbf{W}}, \delta_{\square})$ is one of the central objects of graph limit theory and has remarkable analytic properties [45].

5.2. *Probabilistic definition of convergence.* A more probabilistic definition reduces convergence of non-random graphs to the convergence of random graphs by means of sampling: We use each *non-random* graph g_n to define the distribution of a *random* graph, and then say that (g_n) converges if the resulting distributions do.

More precisely, let g be a finite graph with vertex set $\mathbf{V}(g)$. We can sample a random graph $G(k, g)$ of size k by sampling k vertices of g uniformly at random, without replacement. We then construct $G(k, g)$ as the induced subgraph (the graph consisting of the randomly selected subset of vertices and all edges between them which are present in g.) Formally, this procedure is well-defined even if $k \geq |\mathbf{V}(g)|$, in which case $G(k, g) = g$ with probability 1. Clearly, the distribution of $G(k, g)$ is completely defined by g.

Definition 5.2. Let (g_n) be a sequence of graphs, and let $P(k, g_n)$ be the distribution of $G(k, g_n)$. We say that the graph sequence (g_n) converges if the sequence of distributions $(P(k, g_n))_{n \in \mathbb{N}}$ converges for all k (in the sense of weak convergence of probability measures). ◁

We can obviously as why we should prefer one particular definition of convergence over another one; remarkably, both definitions given above, and also several other definitions studied in the literature, turn out to be equivalent:

Fact 5.3. Definitions 5.1 and 5.2 are equivalent: $d_{\square}(w_{g_n}, w) \to 0$ holds if and only if $P(k, g_n)$ converges weakly for all k. ◁

5.3. *Unique parametrization in the Aldous-Hoover theorem.* The non-uniqueness problem in the Aldous-Hoover theorem is that each random graph is parametrized by an infinite number of distinct functions (Section 3.4). Since the space $\widehat{\mathbf{W}}$ of unique graph limits contains precisely one element for each each exchangeable random graph distribution, we can obtain a unique parametrization by using $\mathbf{T} := \widehat{\mathbf{W}}$ as a parameter space: If $w \in \mathbf{W}$ is a graphon and $\widehat{w}$ the corresponding element of $\widehat{\mathbf{W}}$—the element to which w was collapsed in the definition of $\widehat{\mathbf{W}}$—we define a probability kernel $\mathbf{p}(\,.\,, \widehat{w})$ as the distribution parametrized by w according to the uniform sampling scheme Eq. (3.4). Although the existence of such a probability kernel is not a trivial fact, it follows from a technical result of Orbanz and Szegedy [51]. The Aldous-Hoover theorem for a random graph G can now be written as a mixture

$$\mathbb{P}(G \in \,.\,) = \int_{\widehat{\mathbf{W}}} \mathbf{p}(\,.\,, \widehat{w}) \nu(d\widehat{w}) \,, \tag{5.3}$$

in analogy to the de Finetti representation. As for the other representation results, we now also obtain a diagram

$$\begin{tikzcd}
\Omega \arrow[r, "G"] \arrow[rrr, bend right=15, "\Theta"'] & \mathcal{G} \arrow[r, "S"] & \mathbf{M}(\mathcal{G}) \supset \mathcal{P} \arrow[r, shift left, "T"] & \widehat{\mathbf{W}} \arrow[l, shift left, "T^{-1}"]
\end{tikzcd} \tag{5.4}$$

where $T^{-1}(\widehat{w}) = \mathbf{p}(\,.\,, \widehat{w})$. In this case, G is a random infinite graph; observing a finite sample means observing a finite subgraph G_n of G.

The convergence of the "empirical graphons", the checkerboard functions w_n, to a graph limit corresponds to the convergence of the empirical measure in de Finetti's theorem and of the relative block sizes in Kingman's theorem. The set of graph limits is larger than the set of graphs: Although each graph g has a representation as a measurable function $w_g : [0,1]^2 \to [0,1]$, not each such function represents a graph. Each is, however, the *limit* of a sequence of graphs. The analogy in the de Finetti case is that not each probability distribution represents an empirical measure (since empirical measures are discrete), but every probability measure is the limit of a sequence of empirical measures.

5.4. *Regularity and Concentration.* Asymptotic statistics and empirical process theory provides a range of concentration results which show that the empirical distribution converges with high probability. These results require independence properties, but are model free; adding model assumptions then typically yields more bespoke results with stronger guarantees. Graph limit theory provides a similar type of results for graphs, which are again model free, and based on exchangeability.

Underlying these ideas is one of the deepest and perhaps most surprising results of modern graph theory, Szemeredi's regularity lemma, which shows that for every very large graph g, there is a small, weighted graph $\hat{g}$ that summarizes all essential structure in g. The only condition is that g is sufficiently large. In principle, this means that $\hat{g}$ can be used as an approximation or summary of g, but unfortunately, the result is only valid for graphs which are much larger than possible in most conceivable applications. There are, however, weaker forms of this result which hold for much smaller graphs.

To define $\hat{g}$ for a given graph g, we proceed as follows: Suppose $\Pi := \{V_1, \ldots, V_k\}$ is a partition of $\mathbf{V}(g)$ into k sets. For any two sets V_i and V_j, we define p_{ij} as the probability that two vertices $v \in V_i$ and $v' \in V_j$, each chosen uniformly at random from its set, are connected by an edge. That is,

$$p_{ij} := \frac{\#\text{ edges between } V_i, V_j}{|V_i| \cdot |V_j|} \,. \tag{5.5}$$

The graph $\hat{g}_\Pi$ is now defined as the weighted graph with vertex set $\{1,\dots,k\}$ and edge weights p_{ij} for edge (i,j). To compare this graph to g, it can be helpful to blow it up to a graph g_Π of the same size as g, constructed as follows:

- Each node i is replaced by a clique of size $|V_i|$ (with all edges weighted by 1).
- For each pair V_i and V_j, all possible edges between the sets are inserted and weighted by p_{ij}.

If we measure how much two graphs differ in terms of the distance $d_\square$ defined above, g can be approximated by g_Π as follows:

Theorem 5.4 (Weak regularity lemma [27]). *Let $k \in \mathbb{N}$ and let g be any graph. There is a partition Π of $\mathbf{V}(g)$ into k sets such that $d_\square(g, g_\Pi) \leq 2(\sqrt{\log(k)})^{-1}$.* □

This form of the result is called "weak" since it uses a less restrictive definition of what it means for g and g_Π to be close then Szemerédi's original result. The weaker hypothesis makes the theorem applicable to graphs that are, by the standards of combinatorics, of modest size.

A prototypical concentration result based on Theorem 5.4 is the following:

Theorem 5.5 ([44, Theorem 8.2]). *Let f be a real-valued function on graphs, which is smooth in the sense that $|f(g) - f(g')| \leq d_\square(g, g')$ for any two graphs g and g' defined on the same vertex set. Let $G(k, g)$ be a random graph of size k sampled uniformly from g (see Section 5.2). Then the distribution of $f(G(k, g))$ concentrates around some value $f_0 \in \mathbb{R}$, in the sense that*

$$\mathbb{P}\Big\{|f(G(k,g)) - f_0| > \frac{20}{\sqrt{k}}\Big\} < 2^{-k}\ . \tag{5.6}$$

□

A wide range of similar results for graphs and other random structures is available in graph limit theory and combinatorics, and collectively known under the term *property testing*. Lovász [45, Chapter 15] gives a clear and authorative exposition.

6. Exchangeability in higher-dimensional arrays.

The theory of exchangeable arrays extends beyond 2-dimensional arrays, and, indeed, some of the more exciting implications and applications of the theory rely on the general results. In this section we begin by defining the natural extension of (joint) exchangeability to higher dimensions, and then give higher-dimensional analogues of the theorems of Aldous and Hoover due to Kallenberg. These theorems introduce exponentially-many additional random variables as the dimension increases, but a theorem of Kallenberg's shows that only a linear number are necessary to produce an arbitrarily good approximation. The presentation owes much to Kallenberg [35].

Definition 6.1 (jointly exchangeable d-arrays). Let $(X_{k_1,\dots,k_d})$ be a d-dimensional array (or simply d-array) of random variables in $\mathbf{X}$. We say that X is **jointly exchangeable** when

$$(X_{k_1,\dots,k_d}) \overset{\mathrm{d}}{=} (X_{\pi(k_1),\dots,\pi(k_d)}) \tag{6.1}$$

for every permutation π of $\mathbb{N}$. ◁

As in the 2-dimensional representation result, a key ingredient in the characterization of higher-dimensional jointly exchangeable d-arrays will be an indexed collection U of i.i.d. latent random variables. In order to define the index set for U, let $\tilde{\mathbb{N}}^d$ be the space of multisets $J \subseteq \mathbb{N}$ of cardinality $|J| \leq d$. E.g., $\{\{1,1,3\}\} \in \tilde{\mathbb{N}}^3 \subseteq \tilde{\mathbb{N}}^4$. Rather than two collections—a sequence (U_i) indexed by $\mathbb{N}$, and a triangular array $(U_{\{\{i,j\}\}})$ indexed by multisets of cardinality 2—we will use a single i.i.d. collection U indexed by elements of $\tilde{\mathbb{N}}^d$. For every $I \subseteq [d] := \{1,\dots,d\}$, we will write $\tilde{k}_I$ for the multiset

$$\{\{k_i : i \in I\}\} \tag{6.2}$$

and write

$$(U_{\tilde{k}_I};\ I \in 2^{[d]} \setminus \emptyset) \tag{6.3}$$

for the element of the function space $[0,1]^{2^{[d]}\setminus\emptyset}$ that maps each nonempty subset $I \subseteq [d]$ to the real $U_{\tilde{k}_I}$, i.e., the element in the collection U indexed by the multiset $\tilde{k}_I \in \tilde{\mathbb{N}}^{|I|} \subseteq \tilde{\mathbb{N}}^d$.

Theorem 6.2 (Aldous, Hoover). *Let U be an i.i.d. collection of uniform random variables indexed by multisets $\tilde{\mathbb{N}}^d$. A random d-array $X := (X_k;\ k \in \mathbb{N}^d)$ is jointly exchangeable if and only if there is random measurable function $F : [0,1]^{2^{[d]}\setminus\emptyset} \to \mathbf{X}$ such that*

$$(X_k;\ k \in \mathbb{N}^d) \overset{d}{=} (F(U_{\tilde{k}_I};\ I \in 2^{[d]} \setminus \emptyset);\ k \in \mathbb{N}^d). \tag{6.4}$$

□

When $d = 2$, we recover Theorem 3.4 characterizing two-dimensional exchangeable arrays. Indeed, if we write $U_i := U_{\{\{i\}\}}$ and $U_{ij} := U_{\{\{i,j\}\}}$ for notational convenience, then the right hand side of Eq. (6.4) reduces to

$$(F(U_i, U_j, U_{ij});\ i,j \in \mathbb{N}) \tag{6.5}$$

for some random $F : [0,1]^3 \to \mathbf{X}$. When $d = 3$, we instead have

$$(F(U_i, U_j, U_k, U_{ij}, U_{ik}, U_{jk}, U_{ijk});\ i,j,k \in \mathbb{N}) \tag{6.6}$$

for some random $F : [0,1]^7 \to \mathbf{X}$, where we have additionally taken $U_{ijk} := U_{\{\{i,j,k\}\}}$ for notational convenience. (One may be concerned with the apparent exponential blowup in the number of random variables; We will later describe a result due to Kallenberg that shows that, in a certain technical sense which we will define, the distributions of d-arrays can be arbitrarily well approximated with a random function on $[0,1]^d$.)

6.1. *Separately exchangeable d-arrays.* As in the two-dimensional case, arrays with certain additional symmetries can be treated as special cases. In this section, we consider separate exchangeability in the setting of d-arrays, and in the next section we consider further generalizations. We begin by defining:

Definition 6.3 (separately exchangeable d-arrays). We say that d-array X is **separately exchangeable** when

$$(X_{k_1,\dots,k_d}) \overset{\mathrm{d}}{=} (X_{\pi_1(k_1),\dots,\pi_d(k_d)}) \tag{6.7}$$

for every collection $\pi_1,\dots,\pi_d$ of permutations of $\mathbb{N}$. ◁

For every $J \subseteq [d]$, let 1_J denote its characteristic function (i.e., $1_J(x) = 1$ when $x \in J$ and 0 otherwise), and let the vector $k_J \in \mathbb{Z}_+^d := \{0,1,2,\dots\}^d$ be given by

$$k_J := (k_1\,1_J(1),\dots,k_d\,1_J(d)). \tag{6.8}$$

In order to represent separately exchangeable d-arrays, we will use a collection U of i.i.d. uniform random variables indexed by vectors $\mathbb{Z}_+^d$. Similarly to above, we will write

$$(U_{k_I};\ I \in 2^{[d]} \setminus \emptyset) \tag{6.9}$$

for the element of the function space $[0,1]^{2^{[d]}\setminus\emptyset}$ that maps each nonempty subset $I \subseteq [d]$ to the real U_{k_I}, i.e., the element in the collection U indexed by the vector k_I. Then we have:

Corollary 6.4. *Let U be an i.i.d. collection of uniform random variables indexed by vectors $\mathbb{Z}_+^d$. A random d-array $X := (X_k;\ k \in \mathbb{N}^d)$ is separately exchangeable if and only if there is random measurable function $F : [0,1]^{2^{[d]}\setminus\emptyset} \to \mathbf{X}$ such that*

$$(X_k;\ k \in \mathbb{N}^d) \overset{d}{=} (F(U_{k_I};\ I \in 2^{[d]} \setminus \emptyset);\ k \in \mathbb{N}^d). \tag{6.10}$$

□

We can consider the special cases of $d = 2$ and $d = 3$ arrays. Then we have, respectively,

$$(F(U_{i0}, U_{0j}, U_{ij});\ i,j \in \mathbb{N}) \tag{6.11}$$

for some random $F : [0,1]^3 \to \mathbf{X}$; and

$$(F(U_{i00}, U_{0j0}, U_{00k}, U_{ij0}, U_{i0k}, U_{0jk}, U_{ijk});\ i,j,k \in \mathbb{N}) \tag{6.12}$$

for some random $F : [0,1]^7 \to \mathbf{X}$. As we can see, jointly exchangeable arrays, which are required to satisfy fewer symmetries than their separately exchangeable counterparts, may take $U_{ij0} = U_{0ij} = U_{i0j} = U_{ji0} = \dots$. Indeed, one can show that these additional assumptions make jointly exchangeable arrays a strict superset of separately exchangeable arrays, for $d \geq 2$.

6.2. *Further generalizations.* In applications, it is common for the distribution of an array to be invariant to permutations that act simultaneously on *some but not all* of the dimensions. E.g., if the first two dimensions of an array index into the same collection of users, and the users are *a priori* exchangeable, then a sensible notion of exchangeability for the array would be one for which these first two dimensions could be permuted jointly together, but separately from the remaining dimensions.

More generally, we consider arrays that, given a partition of the dimensions of an array into classes, are invariant to permutations that act jointly within each class and separately across classes. More carefully:

Definition 6.5 (π-exchangeable d-arrays). Let $\pi = \{I_1,\dots,I_m\}$ be a partition of $[d]$ into disjoint classes, and let $p = (p^I;\ I \in \pi)$ be a collection of permutations of $\mathbb{N}$, indexed by the classes in π. We say that a d-array X is π-exchangeable when

$$(X_{k_1,\dots,k_d};\ k \in \mathbb{N}^d) \overset{\mathrm{d}}{=} (X_{p^{\pi_1}(k_1),\dots,p^{\pi_d}(k_d)};\ k \in \mathbb{N}^d), \tag{6.13}$$

for every collection p of permutations, where π_i denotes the subset $I \in \pi$ containing i. ◁

We may now cast both jointly and separately exchangeable arrays as π-exchangeable arrays for particular choices of partitions π. In particular, when $\pi = \{[d]\}$ we recover joint exchangeability, and when $\pi = \{\{1\},\dots,\{d\}\}$, we recover separate exchangeability. Just as we characterized jointly and separately exchangeable arrays, we can characterize π-exchangeable arrays.

Let π be a partition of $[d]$. In order to describe the representation of π-exchangeable d-arrays, we will again need a collection U of i.i.d. uniform random variables, although the index set is more complicated than before: Let $\mathcal{V}(\pi) := \times_{I\in\pi} \tilde{\mathbb{N}}^{|I|}$ denote the space of functions taking classes $I \in \pi$ to multisets $J \subseteq \mathbb{N}$ of cardinality $J \leq |I|$. We will then take U to be a collection of i.i.d. uniform random variables indexed by elements in $\mathcal{V}(\pi)$.

It is worth spending some time giving some intuition for $\mathcal{V}(\pi)$. When $\pi = \{[d]\}$, $\mathcal{V}(\pi)$ is equivalent to the space $\tilde{\mathbb{N}}^d$ of multisets of cardinality no more than d, in agreement with the index set in the jointly exchangeable case. The separately exchangeable case is also instructive: there $\pi = \{\{1\},\dots,\{d\}\}$ and so $\mathcal{V}(\pi)$ is equivalent to the space of functions from $[d]$ to $\tilde{\mathbb{N}}^1$, which may again be seen to be equivalent to the space $\mathbb{Z}_+^d$ of vectors, where 0 encodes the empty set $\emptyset \in \tilde{\mathbb{N}}^1 \cap \tilde{\mathbb{N}}^0$. For a general partition π of $[d]$, an element in $\mathcal{V}(\pi)$ is a type of generalized vector, where, for each class $I \in \pi$ of dimensions that are jointly exchangeable, we are given a multiset of indices.

For every $I \subseteq [d]$, let $\tilde{k}_{\pi I} \in \mathcal{V}(\pi)$ be given by

$$\tilde{k}_{\pi I}(J) = \tilde{k}_{I\cap J}, \quad J \in \pi, \tag{6.14}$$

where $\tilde{k}_J$ is defined as above for jointly exchangeable arrays. We will write

$$(U_{\tilde{k}_{\pi I}};\ I \in 2^{[d]} \setminus \emptyset) \tag{6.15}$$

for the element of the function space $[0,1]^{2^{[d]}\setminus\emptyset}$ that maps each nonempty subset $I \subseteq [d]$ to the real $U_{\tilde{k}_{\pi I}}$, i.e., the element in the collection U indexed by the generalized vector $\tilde{k}_{\pi I}$. Then we have:

Corollary 6.6 (Kallenberg [35]). *Let π be a partition of $[d]$, and let U be an i.i.d. collection of uniform random variables indexed by generalized vectors $\mathcal{V}(\pi)$. A random d-array $X := (X_k;\ k \in \mathbb{N}^d)$ is π-exchangeable if and only if there is random measurable function $F : [0,1]^{2^{[d]}\setminus\emptyset} \to \mathbf{X}$ such that*

$$(X_k;\ k \in \mathbb{N}^d) \overset{d}{=} (F(U_{\tilde{k}_{\pi I}};\ I \in 2^{[d]} \setminus \emptyset);\ k \in \mathbb{N}^d). \qquad (6.16)$$

□

6.3. *Approximations by simple arrays.* These representational results require a number of latent random variables exponential in the dimension of the array, i.e., roughly twice as many latent variables are needed as the entries generated in some subarray. Even if a d-array is sparsely observed, each observation requires the introduction of potentially 2^d variables. (In a densely observed array, there will be overlap, and most latent variables will be reused.)

Regardless of whether this blowup poses a problem for a particular application, it is interesting to note that exchangeable d-arrays can be approximated by arrays with much simpler structure, known as **simple arrays**.

Definition 6.7 (simple d-arrays). Let $U = (U_k^I;\ I \in \pi, k \in \mathbb{N})$ be an i.i.d. collection of uniform random variables. We say that a π-exchangeable d-array X is **simple** when there is a random function $F\colon [0,1]^{[d]} \to \mathbf{X}$ such that

$$(X_k;\ k \in \mathbb{N}^d) \overset{\mathrm{d}}{=} (F(U_{k_1}^{\pi_1}, \dots, U_{k_d}^{\pi_d});\ k \in \mathbb{N}^d), \qquad (6.17)$$

where π_j is defined as above. ◁

Again, it is instructive to study special cases: in the jointly exchangeable case, taking $U_j := U_j^{\{[d]\}}$, we get

$$(F(U_{k_1}, \dots, U_{k_d}); k \in \mathbb{N}^d) \qquad (6.18)$$

and, in the separately exchangeable case, we get

$$(F(U_{k_1}^1, \dots, U_{k_d}^d);\ k \in \mathbb{N}^d), \qquad (6.19)$$

taking $U_j^i := U_j^{\{i\}}$. We may now state the relationship between general arrays and simple arrays:

Theorem 6.8 (simple approximations, Kallenberg [35, Thm. 2]). *Let X be a π-exchangeable d-array. Then there exists a sequence of simple π-exchangeable arrays $X^1, X^2, \dots$ such that, for all finite sub-arrays $X_J := (X_k; k \in J)$, $J \subseteq \mathbb{N}^d$, the distributions of X_J and X_J^n are mutually absolutely continuous, and the associated densities tend* uniformly to 1 *as $n \to \infty$ for fixed J.* □

7. Sparse random structures and networks.

Exchangeable random structures are not "sparse". In an exchangeable infinite graph, for example, the expected number of edges attached to each node is either infinite or zero. In contrast, graphs representing network data typically have a finite number of edges per vertex, and exhibit properties like power-laws and "small-world phenomena", which can only occur in sparse graphs. Hence, even though exchangeable graph models are widely used in network analysis, they are inherently misspecified. We have emphasized previously that most Bayesian models are based on exchangeability. The lack of sparseness, however, is a direct mathematical consequence of exchangeability. Thus, networks and sparse random structures pose a problem that seems to require genuinely non-exchangeable models. The development of a coherent theory for sparse random graphs and structures is, despite intense efforts in mathematics, a largely unsolved problem, and so is the design of Bayesian models for networks data. In this section, we make the problem more precise and describe how, at least in principle, exchangeability might be substituted by other symmetry properties. We also briefly summarize a few specific results on sparse graphs. The topic raises a host of challenging questions to which, in most cases, we have no answers.

7.1. *Dense vs Sparse Random Structures.* In an exchangeable structure, events either never occur, or they occur infinitely often with a fixed, constant (though unknown) probability. The simplest example is an exchangeable binary sequence: Since the order of observations is irrelevant, the probability of observing a one is the same for all entries in the sequence. If this probability is $p \in [0,1]$, and we sample infinitely often, the fraction of ones in the infinite sequence will be precisely p. Therefore, we either observe a constant proportion of ones (if $p > 0$) or no ones at all (if $p = 0$). In an exchangeable graph, rather than ones and zeros, we have to consider the possible subgraphs (single edges, triangles, five-stars, etc). Each possible subgraph occurs either never, or infinitely often.

Since an infinite graph may have infinitely many edges even if it is sparsely connected, the number of edges is best quantified in terms of a rate:

Definition 7.1. Let $g = (v, e)$ be an infinite graph with vertex set $\mathbb{N}$ and let $g_n = (v_n, e_n)$ be the subgraph on $\{1, \dots, n\}$. We say that g is **sparse** if, as n increases, $|e_n|$ is of size $\Omega(n)$ (is upper-bounded by $c \cdot n$ for some constant c). It is called **dense** if $|e_n| = \Theta(n^2)$ (lower-bounded by $c \cdot n^2$ for some constant c). ◁

Many important types of graph and array data are inherently sparse: In a social network with billions of users, individual users do not, on average, have billions of friends.

Fact 7.2. Exchangeable graphs are not sparse. If a random graph is exchangeable, it is either dense or empty. ◁

The argument is simple: Let G_n be an n-vertex random undirected graph sampled according to Eq. (3.4). The expected proportion of edges in present in G_n, out of all $\binom{n}{2} = \frac{n(n-2)}{2}$ possible edges, is independent of n and given by $\varepsilon := \frac{1}{2}\int_{[0,1]^2} W(x,y)dxdy$. (The factor $\frac{1}{2}$ occurs since W is symmetric.) If $\varepsilon = 0$, it follows that G_n is empty with probability one and therefore trivially sparse. On the other hand, if $\varepsilon > 0$, we have $\varepsilon \cdot \binom{n}{2} = \Theta(n^2)$ edges in expectation and so, by the law of large numbers, G_n is dense with probability one.

Remark 7.3 (Graph limits are dense). The theory of graph limits described in Section 5 is intimately related to exchangeability, and is inherently a theory of dense graphs: If we construct a sequence of graphs with sparsely growing edge sets, convergence in cut metric is still well-defined, but the limit object is always the empty graphon, i.e., a function on $[0,1]^2$ which vanishes almost everywhere. ◁

The theory of dense graphs, as described in this article, is well-developed; the theory of sparse graphs, in contrast, is not, and the practical importance of such graphs therefore raises crucial questions for further research.

7.2. *Beyond exchangeability: Symmetry and ergodic theory.* Exchangeability is a specific form of probabilistic symmetry: Mathematically, symmetries are expressed as invariance under a group. Exchangeability is the special case where this group is either the infinite symmetric group (as in de Finetti's theorem), or a under a suitable subgroup (as in the Aldous-Hoover theorem). A very general mathematical result, the *ergodic decomposition theorem*, shows that integral decompositions of the form (2.1) are a general consequence of symmetry properties, rather than specifically of exchangeability. The general theme is that there is some correspondence of the form

$$\text{invariance property} \quad \longleftrightarrow \quad \text{integral decomposition} \;.$$

In principal, Bayesian models can be constructed based on any type of symmetry, as long as this symmetry defines a useful set of ergodic distributions.

The following statement of the ergodic decomposition theorem glosses over various technical details; for a precise statement, see e.g., [37, Theorem A1.4].

Theorem 7.4 (Varadarajan [64]). *If the distribution of a random structure X_∞ is invariant under a nice group G (= has a symmetry property), it has a representation of the form*

$$\mathbb{P}(X_\infty \in \,.\,) = \int_{\mathbf{T}} \mathbf{p}(\,.\,,\theta)\nu(\theta) \;. \tag{7.1}$$

The group G defines a set $\mathcal{E}$ of ergodic distributions on $\mathbf{X}_\infty$, and $\mathbf{p}(\,.\,,\theta)$ is a distribution in $\mathcal{E}$ for each $\theta \in \mathbf{T}$. □

Following the discussion in Section 2, the components of the theorem will look familiar. In Bayesian terms, $\mathbf{p}(\,.\,,\theta)$ again corresponds to the observation distribution and ν to the prior. Geometrically, integral representations like Eq. (7.1) can be regarded as convex combinations (as illustrated in Fig. 9 for a toy example with three ergodic measures).

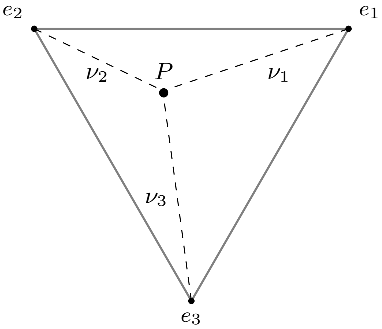

Fig 9: If $\mathcal{E}$ is finite, the de Finetti mixture representation Eq. (2.5) and the more general representation Eq. (7.1) reduce to a finite convex combination. The points inside the set—i.e., the distributions P with the symmetry property defined by the group G—can be represented as convex combinations $P = \sum_{e_i\in\mathcal{E}} \nu_i e_i$, with coefficients $\nu_i \geq 0$ satisfying $\sum_i \nu_i = 1$. When $\mathcal{E}$ is infinite, an integral is substituted for the sum.

A special case of this result is well-known in Bayesian theory as a result of David Freedman [25, 26].

Example 7.5 (Freedman's theorem). Consider a sequence $X_1, X_2, \dots$ as in de Finetti's theorem. Now replace invariance under permutations by a stronger condition: Let $O(n)$ be the group of rotations and reflections on $\mathbb{R}^n$, i.e., the set of $n \times n$ orthogonal matrices. We now demand that, if we regard any initial sequence of n variables as a random vector in $\mathbb{R}^n$, then rotating this vector does not change the distribution of the sequence: For any $n \in \mathbb{N}$ and any $M \in O(n)$,

$$(X_1, X_2, \dots) \overset{\text{d}}{=} (M(X_1, \dots, X_n), X_{n+1}, X_{n+2} \dots) \;. \tag{7.2}$$

In the language of Theorem 7.4, the group G is the set of all rotations of any length, $G = \cup_{n\in\mathbb{N}} O(n)$. If X^∞ satisfies Eq. (7.2), its distribution is a scale mixture of Gaussians:

$$\mathbb{P}(X^\infty \in \,.\,) = \int_{\mathbb{R}_+} \Big(\prod_{n=1}^{\infty} N_\sigma(X_n)\Big) d\nu_{\mathbb{R}_+}(\sigma) \tag{7.3}$$

Thus, $\mathcal{E}$ contains all factorial distributions of zero-mean normal distributions on $\mathbb{R}$, $\mathbf{T}$ is the set $\mathbb{R}_{>0}$ of variances, and ν a distribution on $\mathbb{R}_{>0}$. ◁

Compared to de Finetti's theorem, the of the group G has been increased: Any permutation can be represented as an orthogonal matrix, but here rotations have been added as well. In other words, we are strenghtening the hypothesis by imposing more constraints on the distribution of X^∞. As a result, the set $\mathcal{E}$ of ergodic measures shrinks from all factorial measures to the set of factorials of zero-mean Gaussians. This is again an example of a general theme:

$$\text{larger group} \quad \longleftrightarrow \quad \text{more specific representation}$$

In contrast, the Aldous-Hoover theorem *weakens* the hypothesis of de Finetti's theorem—in the matrix case, for instance, the set of all permutations of the index set $\mathbb{N}^2$ is restricted to those which preserve rows and columns—and hence yields a more general representation.

REMARK 7.6 (Symmetry and sufficiency). An alternative way to define symmetry in statistical models is through sufficient statistics: Intuitively, a symmetry property identifies information which is not relevant to the statistical problem; so does a sufficient statistic. For example, the empirical distribution retains all information about a sample except for the order in which observations are recorded. A model for random sequences is hence exchangeable if and only if the empirical distribution is a sufficient statistic. In an exchangeable graph model, the empirical graphon (the checkerboard function in Fig. 8) is a sufficient statistic. If the sufficient statistic is finite-dimensional and computes an average $\frac{1}{n}\sum_i S_0(x_i)$ over observations for some function S_0, the ergodic distributions are exponential family models [41]. A readable introduction to this topic is given by Diaconis [20]. The definitive reference is the monograph of Lauritzen [42], who refers to the set $\mathcal{E}$ of ergodic distributions as an *extremal family.* ◁

The ergodic decomposition theorem does not, unfortunately, solve all foundational problems of Bayesian inference. To be useful to statistics, a symmetry principle must satisfy two conditions:

1. The set $\mathcal{E}$ of ergodic measures should be a "small" subset of the set of symmetric measures.
2. The measures $\mathbf{p}(\,.\,,\theta)$ should have a tractable representation, such as Kingman's paint-box or the Aldous-Hoover sampling scheme.

Theorem 7.4 guarantees neither. If (1) is not satisfied, the representation is useless for statistical purposes: The integral representation Eq. (7.1) means that the information in X_∞ is split into two parts, the information contained in the parameter value θ (which a statistical procedure tries to extract) and the randomness represented by $\mathbf{p}(\,.\,,\theta)$ (which the statistical procedure discards). If the set $\mathcal{E}$ is too large, Θ contains almost all the information in X_∞, and the decomposition becomes meaningless. We will encounter an appealing notion of symmetry for sparse networks in the next section—which, however, seems to satisfy neither condition (1) or (2). It is not clear at present whether there are useful types of symmetries based on groups which are not isomorphic to a group of permutations. In light of the apparent contradiction between sparseness and exchangeability, this question, despite its abstraction, seems to be of some importance to the Bayesian paradigm.

7.3. *Stationary networks and involution invariance.*

A class of sparse random structures of particular interest are networks. There is a large and rapidly growing literature on this subject in applied probability, which defines and studies specific graph distributions and their probabilistic properties; [23] is a good survey. Similarly, a huge literature available on applications [e.g. 50]. Lacking at present are both a proper statistical understanding of such models, and a mathematical theory similarly coherent as that provided by graph limits for dense graphs. This final section describes some concepts at the intersection of network problems and exchangeable random structures.

One possible way to generate sparse graphs is of course to modify the sampling scheme for exchangeable graphs to generate fewer edges.

EXAMPLE 7.7 (The BJR model). There is a very simple way to translate the Aldous-Hoover approach into a sparse graph: Suppose we sample rows and columns of the matrix consecutively. At the nth step, we sample X_{nj} for all $j < n$. Now we multiply the probability in our usual sampling scheme by $1/n$:

$$X_{nj} \sim \text{Bernoulli}\Big(\frac{1}{n}w(U_n, U_j)\Big)\,. \tag{7.4}$$

Comparison with our argument why exchangeable graphs are dense immediately shows that a graph sampled this way is sparse. This class of random graphs was introduced by Bollobás, Janson, and Riordan [13]. The BJR model contains various interesting models as special cases; for instance, setting $w(x,y) := \frac{c}{\sqrt{xy}}$ yields the mean-field version of the well-known Barabási-Albert model (though not the Barabási-Albert model itself) [12]. A moment estimator for the edge density under this model is studied by Bickel, Chen, and Levina [11]. ◁

An obvious limitation of the BJR model is that it does not actually attempt to model network structure; rather, it modifies a model of exchangeable structure to fit a first-order statistic (the number of edges) of the network.

A crucial difference between network structures and exchangeable graphs is that, in most networks, location in the graph matters. If conditioning on location is informative, exchangeability is broken. Probabilistically, location is modeled by marking a distinguished vertex in the graph. A **rooted graph** (g, v) is simply a graph g in which a particular vertex v has been marked as the root. A very natural notion of invariance for networks modeled by rooted graphs is the following:

DEFINITION 7.8. Let P be the distribution of a random rooted graph, and define a distribution $\tilde{P}$ as follows: A sample $(G, w) \sim \tilde{P}$ is generated by sampling $(G, v) \sim P$, and then sampling w uniformly from the neighbors of v in G. The distribution P is called **involution invariant** if $P = \tilde{P}$. ◁

The definition says that, if an observer randomly walks along the graph G by mfoving to a uniformly selected neighbor in each step, the *distribution* of the network around the observer remains unchanged (although the actual neighborhoods in a sampled graph may vary). This is can be thought of as a network analogue of a stationary stochastic process.

An equivalent (though more technical) definition of introduces a shift mapping, which shifts the root v to a neighbor w [2]. Involution invariance then means that P is invariant under such shifts, just as exchangeable distribution are invariant under permutations. In particular, it is a symmetry property, and involution invariant graphs admit an ergodic decomposition. Aldous and Lyons [1] have characterized the ergodic measures.

This characterization is abstract, however, and there is no known "nice" representation resembling, for example, the sampling scheme for exchangeable graphs. Thus, of the two desiderata described in Section 7.2, property (2) does not seem to hold. We believe that property (1) does not hold either: Although we have no proof at present, we conjecture that every involution invariant distribution can be closely approximated by an ergodic measure (i.e., the set of ergodic distributions is a "large" subset of the involution invariant distributions). Involution invariance is the only reasonably well-studied notion of invariance for sparse graphs, but despite its intuitive appeal, it seems to constitute and example of a symmetry that is too weak to yield useful statistical models.

8. Further References.

Excellent non-technical references on the general theory of exchangeable arrays and other exchangeable random structures are two recent surveys by Aldous [5, 6]. His well-known lecture notes [4] also cover exchangeable arrays. The most comprehensive available reference on the general theory is the monograph by Kallenberg [37] (which presupposes in-depth knowledge of measure-theoretic probability). Kingman's original article [39] provides a concise reference on exchangeable random partitions. A thorough, more technical treatment of exchangeable partitions can be found in [10].

Schervish [58] gives an insightful discussion of the application of exchangeability to Bayesian statistics. There is a close connection between symmetry principles (such as exchangeability) and sufficient statistics, which is covered by a substantial literature. See Diaconis [20] for an introduction and further references. For applications of exchangeability results to machine learning models, see [24], who discuss applications of the partial exchangeability result of Diaconis and Freedman [21] to the infinite hidden Markov model [9].

The theory of graph limits in its current form was initiated by Lovász and Szegedy [46, 47] and Borgs *et al.* [14]. It builds on work of Frieze and Kannan [27], who introduced both the weak regularity lemma (Theorem 5.4) and the cut norm $d_\square$. In the framework of this theory, the Aldous-Hoover representation of exchangeable graphs can be derived by purely analytic means [46, Theorem 2.7]. The connection between graph limits and Aldous-Hoover theory was established, independently of each other, by Diaconis and Janson [22] and by Austin [7]. A lucid introduction to the analytic perspective is the survey Lovász [44], which assumes basic familiarity with measure-theoretic probability and functional analysis, but is largely non-technical.

Historically, the Aldous-Hoover representation was established in independent works of David Aldous and Douglas N. Hoover in the late 1970s. Aldous proof used probability-theoretic methods, whereas Hoover, a logician, leveraged techniques from model theory. In 1979, Kingman [40] writes

> ...a general solution has now been supplied by Dr David Aldous of Cambridge. [...] The proof is at present very complicated, but there is reason to hope that the techniques developed can be applied to more general experimental designs.

Aldous' paper [3], published in 1981, attributes the idea of the published version of the proof to Kingman. The results were later generalized considerably by Olav Kallenberg [35].

References.

[1] Aldous, D. and Lyons, R. (2007). Processes on unimodular random networks. *Electron. J. Probab.*, **12**, no. 54, 1454–1508.

[2] Aldous, D. and Steele, J. M. (2004). The objective method: Probabilistic combinatorial optimization and local weak convergence. In H. Kesten, editor, *Probability on Discrete Structures*, volume 110 of *Encyclopaedia Math. Sci.*, pages 1–72. Springer.

[3] Aldous, D. J. (1981). Representations for partially exchangeable arrays of random variables. *J. Multivariate Anal.*, **11**(4), 581–598.

[4] Aldous, D. J. (1985). Exchangeability and related topics. In P. L. Hennequin, editor, *École d'Été de Probabilités de Saint-Flour XIII - 1983*, number 1117 in Lecture Notes in Mathematics, pages 1–198. Springer.

[5] Aldous, D. J. (2009). More uses of exchangeability: Representations of complex random structures.

[6] Aldous, D. J. (2010). Exchangeability and continuum limits of discrete random structures. In *Proceedings of the International Congress of Mathematicians*.

[7] Austin, T. (2008). On exchangeable random variables and the statistics of large graphs and hypergraphs. *Probab. Surv.*, **5**, 80–145.

[8] Bacallado, S. A., Favaro, S., and Trippa, L. (2013). Bayesian nonparametric analysis of reversible Markov chains. *Ann. Statist.* To appear.

[9] Beal, M. J., Ghahramani, Z., and Rasmussen, C. (2002). The infinite hidden Markov model. In T. G. Dietterich, S. Becker, and Z. Ghrahmani, editors, *Advances in Neural Information Processing Systems*, volume 14, pages 577–584.

[10] Bertoin, J. (2006). *Random Fragmentation and Coagulation Processes*. Cambridge University Press.

[11] Bickel, P. J., Chen, A., and Levina, E. (2011). The method of moments and degree distributions for network models. *Ann. Statist.*, **39**(5), 2280–2301.

[12] Bollobás, B. and Riordan, O. (2009). Random graphs and branching processes. In *Handbook of large-scale random networks*, volume 18 of *Bolyai Soc. Math. Stud.*, pages 15–115. Springer, Berlin.

[13] Bollobás, B., Janson, S., and Riordan, O. (2007). The phase transition in inhomogeneous random graphs. *Random Structures Algorithms*, **31**(1), 3–122.

[14] Borgs, C., Chayes, J., Lovász, L., Sós, V. T., Szegedy, B., and Vesztergombi, K. (2005). Graph limits and parameter testing. In *Topics in discrete mathematics*, volume 25 of *Algorithms Combin.* Springer.

[15] Borgs, C., Chayes, J., and Lovász, L. (2010). Moments of two-variable functions and the uniqueness of graph limits. *Geometric And Functional Analysis*, **19**(6), 1597–1619.

[16] Broderick, T., Jordan, M. I., and Pitman, J. (2013). Feature allocations, probability functions, and paintboxes. To appear in *Bayesian Anal.*, arXiv:1301.6647.

[17] Bühlmann, H. (1960). *Austauschbare stochastische Variabeln und ihre Grenzwertsätze*. Ph.D. thesis. University of California Press, 1960.

[18] Chatterjee, S. (2012). Matrix estimation by universal singular value thresholding.

[19] de Finetti, B. (1931). Fuzione caratteristica di un fenomeno

aleatorio. *Atti della R. Academia Nazionale dei Lincei*, **4**, 251–299.

[20] Diaconis, P. (1992). Sufficiency as statistical symmetry. In F. Browder, editor, *Proc. 100th Anniversary Americal Mathematical Society*, pages 15–26. American Mathematical Society.

[21] Diaconis, P. and Freedman, D. (1980). De Finetti's theorem for Markov chains. *The Annals of Probability*, **8**(1), pp. 115–130.

[22] Diaconis, P. and Janson, S. (2008). Graph limits and exchangeable random graphs. *Rendiconti di Matematica, Serie VII*, **28**, 33–61.

[23] Durrett, R. (2006). *Random Graph Dynamics*. Cambridge University Press.

[24] Fortini, S. and Petrone, S. (2012). Predictive construction of priors in Bayesian nonparametrics. *Braz. J. Probab. Stat.*, **26**(4), 423–449.

[25] Freedman, D. A. (1962). Invariants under mixing which generalize de Finetti's theorem. *Ann. Math. Statist.*, **33**, 916–923.

[26] Freedman, D. A. (1963). Invariants under mixing which generalize de Finetti's theorem. *Ann. Math. Statist.*, **34**(1194–1216).

[27] Frieze, A. and Kannan, R. (1999). Quick approximation to matrices and applications. *Combinatorica*, **19**(2), 175–220.

[28] Griffiths, T. L. and Ghahramani, Z. (2006). Infinite latent feature models and the Indian buffet process. In *Adv. in Neural Inform. Processing Syst. 18*, pages 475–482. MIT Press, Cambridge, MA.

[29] Hewitt, E. and Savage, L. J. (1955). Symmetric measures on Cartesian products. *Transactions of the American Mathematical Society*, **80**(2), 470–501.

[30] Hjort, N., Holmes, C., Müller, P., and Walker, S., editors (2010). *Bayesian Nonparametrics*. Cambridge University Press.

[31] Hoff, P. (2008). Modeling homophily and stochastic equivalence in symmetric relational data. In *Adv. Neural Inf. Process. Syst. 2007*.

[32] Hoff, P. D. (2011). Separable covariance arrays via the Tucker product, with applications to multivariate relational data. *Bayesian Analysis*, **6**(2), 179–196.

[33] Holland, P. W., Laskey, K. B., and Leinhardt, S. (1983). Stochastic blockmodels: First steps. *Social Networks*, **5**(2), 109–137.

[34] Hoover, D. N. (1979). Relations on probability spaces and arrays of random variables. Technical report, Institute of Advanced Study, Princeton.

[35] Kallenberg, O. (1999). Multivariate sampling and the estimation problem for exchangeable arrays. *J. Theoret. Probab.*, **12**(3), 859–883.

[36] Kallenberg, O. (2001). *Foundations of Modern Probability*. Springer, 2nd edition.

[37] Kallenberg, O. (2005). *Probabilistic Symmetries and Invariance Principles*. Springer.

[38] Kemp, C., Tenenbaum, J., Griffiths, T., Yamada, T., and Ueda, N. (2006). Learning systems of concepts with an infinite relational model. In *Proc. of the Nat. Conf. on Artificial Intelligence*, volume 21, page 381.

[39] Kingman, J. F. C. (1978). The representation of partition structures. *J. London Math. Soc.*, **2**(18), 374–380.

[40] Kingman, J. F. C. (1979). Discussion of: "on the reconciliation of probability assessments" by d. v. lindley, a. tversky and r. v. brown. *Journal of the Royal Statistical Society. Series A (General)*, **142**(2), 171.

[41] Küchler, U. and Lauritzen, S. L. (1989). Exponential families, extreme point models and minimal space-time invariant functions for stochastic processes with stationary and independent increments. *Scand. J. Stat.*, **16**, 237–261.

[42] Lauritzen, S. L. (1988). *Extremal Families and Systems of Sufficient Statistics*. Lecture Notes in Statistics. Springer.

[43] Lloyd, J. R., Orbanz, P., Roy, D. M., and Ghahramani, Z. (2012). Random function priors for exchangeable arrays. In *Adv. in Neural Inform. Processing Syst. 25*.

[44] Lovász, L. (2009). Very large graphs. In D. Jerison, B. Mazur, T. Mrowka, W. Schmid, R. Stanley, and S. T. Yau, editors, *Current Developments in Mathematics*, pages 67–128. International Press.

[45] Lovász, L. (2013). *Large Networks and Graph Limits*. American Mathematical Society.

[46] Lovász, L. and Szegedy, B. (2006). Limits of dense graph sequences. *J. Combin. Theory Ser. B*, **96**, 933–957.

[47] Lovász, L. and Szegedy, B. (2007). Szemerédi's lemma for the analyst. *Geometric And Functional Analysis*, **17**(1), 252–270.

[48] MacEachern, S. N. (2000). Dependent Dirichlet processes. Technical report, Ohio State University.

[49] Miller, K. T., Griffiths, T. L., and Jordan, M. I. (2009). Nonparametric latent feature models for link prediction. *Advances in Neural Information Processing Systems (NIPS)*, pages 1276–1284.

[50] Newman, M. (2009). *Networks. An Introduction*. Oxford University Press.

[51] Orbanz, P. and Szegedy, B. (2012). Borel liftings of graph limits. Preprint.

[52] Paisley, J., Zaas, A., Woods, C., Ginsburg, G., and Carin, L. (2010). A stick-breaking construction of the beta process. In *Proc. Int. Conf. on Machine Learning*.

[53] Paisley, J., Blei, D., and Jordan, M. (2012). Stick-breaking beta processes and the poisson process. In *Proc. Int. Conf. on A.I. and Stat.*

[54] Pitman, J. (2006). *Combinatorial stochastic processes*, volume 1875 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin. Lectures from the 32nd Summer School on Probability Theory held in Saint-Flour, July 7–24, 2002, With a foreword by Jean Picard.

[55] Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. The MIT Press.

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

[57] Roy, D. M. and Teh, Y. W. (2009). The Mondrian process. In *Advances in Neural Information Processing Systems*, volume 21, page 27. Citeseer.

[58] Schervish, M. J. (1995). *Theory of Statistics*. Springer.

[59] Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statist. Sinica*, **4**(2), 639–650.

[60] Teh, Y. W. and Jordan, M. I. (2010). Hierarchical Bayesian nonparametric models with applications. In N. L. Hjort, C. Holmes, P. Müller, and S. G. Walker, editors, *Bayesian Nonparametrics*. Cambridge University Press.

[61] Teh, Y. W., Görür, D., and Ghahramani, Z. (2007). Stick-breaking construction for the Indian buffet process. In *Proc. of the 11th Conf. on A.I. and Stat.*

[62] Teh, Y. W., Blundell, C., and Elliott, L. (2011). Modelling genetic variations using fragmentation-coagulation processes. In *Adv. Neural Inf. Process. Syst.*, pages 819–827.

[63] Thibaux, R. and Jordan, M. I. (2007). Hierarchical beta processes and the Indian buffet process. In *Proc. of the 11th Conf. on A.I. and Stat.*

[64] Varadarajan, V. S. (1963). Groups of automorphisms of Borel spaces. *Transactions of the American Mathematical Society*, **109**(2), pp. 191–220.

[65] Wasserman, S. and Anderson, C. (1987). Stochastic a posteriori blockmodels: Construction and assessment. *Social Networks*, **9**(1), 1–36.

[66] Xu, Z., Tresp, V., Yu, K., and Kriegel, H.-P. (2006). Infinite hidden relational models. In *Proc. of the 22nd Int. Conf. on Uncertainity in Artificial Intelligence (UAI)*.

[67] Xu, Z., Yan, F., and Qi, Y. (2012). Infinite Tucker Decomposition: Nonparametric Bayesian Models for Multiway Data Analysis. In *Proceedings of the 29th International Conference on Machine Learning*.

[68] Zabell, S. L. (1995). Characterizing Markov exchangeable sequences. *J. Theoret. Probab.*, **8**(1), 175–178.