The Annals of Probability
1997, Vol. 25, No. 3, 1437–1450

THE REPRESENTATION OF COMPOSITION STRUCTURES

BY ALEXANDER V. GNEDIN

University of Göttingen

A composition structure is a sequence of consistent probability distributions for compositions (ordered partitions) of $n = 1, 2, \ldots$. Any composition structure can be associated with an exchangeable random composition of the set of natural numbers. Following Donnelly and Joyce, we study the problem of characterizing a generic composition structure as a convex mixture of the "extreme" ones. We topologize the family $\mathscr{U}$ of open subsets of $[0, 1]$ so that $\mathscr{U}$ becomes compact and show that $\mathscr{U}$ is homeomorphic to the set of extreme composition structures. The general composition structure is related to a random element of $\mathscr{U}$ via a construction introduced by J. Pitman.

1. Introduction. The problem of describing all random exchangeable partitions of $\mathbb{N} := \{1, 2, \ldots\}$ was solved by Kingman [8, 9]. Let π be such a partition thought of as a sequence of its restrictions (π_n) onto $[n] := \{1, 2, \ldots, n\}$, so that the probability distribution of each π_n is invariant under permutations of $[n]$. Suppose the classes of π_n are arranged in order of decreasing size and let $f_{i,n}$ be the size of the ith largest class, with the convenience $f_{i,n} = 0$ if the number of classes represented in $[n]$ is less than i. The sequence of probability distributions for the vectors of counts $(f_{1,n}, f_{2,n}, \ldots)$, $n = 1, 2, \ldots$ is what Kingman called a *partition structure*. By exchangeability, the distribution of π is determined uniquely by the partition structure. A now classical representation result of Kingman says that, for each $i \geq 1$, there exists a limit $f_{i,n}/n \to f_i$, with $f_1 \geq f_2 \geq \cdots \geq 0$ and $\sum f_i \leq 1$. The partition structure can be recovered from the frequencies $(f_1, f_2, \ldots)$ via a "paintbox process" construction. Aldous [1] suggested an elegant proof based on de Finetti's theorem.

It is implicit in Kingman's representation that any class of π is either a singleton or infinite. Each infinite class has a positive limiting frequency. The singleton classes are present if and only if their total limiting proportion $1 - \sum f_i$ is positive.

Ordering the classes of π_n by size is not consistent for different values of n. Pitman [11] studied the representation problem with a more informative labelling of partition classes of π_n by the order of their minimal elements. That is to say, the first class contains 1, the second class contains the minimal natural number which is not in the first class, and so on. Pitman showed that the relative class sizes converge and described all possible distributions for the limiting frequencies $(g_1, g_2, \ldots)$. If $\sum g_i = 1$ (this is equivalent to $\sum f_i = 1$), then $(g_1, g_2, \ldots)$ is a size-biased permutation of $(f_1, f_2, \ldots)$. According to

Received February 1996; revised October 1996.

AMS 1991 *subject classifications*. Primary 60G09; secondary 60C05, 60J50.

Key words and phrases. Composition structure, partition structure, exchangeability, paintbox process, random set.

Pitman's representation theorem, each partition π endowed with this labelling can be derived from $(g_1, g_2, \ldots)$ via a sequential procedure generalizing the Dubins–Pitman "Chinese restaurant process" [1, 11].

The situation changes radically for random *exchangeable compositions* of $\mathbb{N}$, that is, partitions with a total order on the collection of classes. Suppose the classes represented on $[n]$ are arranged consistently with their ordering in the whole of the composition and let $(h_{1,n}, h_{2,n}, \ldots)$ be the class sizes counts. Donnelly and Joyce [4] showed that the limiting frequencies, $h_{i,n}/n \to h_i$, exist also in this case, but in general the composition cannot be reconstructed from $(h_1, h_2, \ldots)$, unless $\sum h_i = 1$: a condition which holds iff the collection of classes is *simply ordered*, that is, has the same order structure as a finite or infinite set of natural numbers. Motivated primarily by applications in population genetics, Donnelly and Joyce concentrated on the simply ordered case to which the models of biological interest belong. To describe the most general exchangeable composition, Donnelly and Joyce suggested an ordered version of the "paintbox process" involving a random probability measure on [0, 1]. However, they did not give an explicit limiting procedure to derive the representing measure from $(h_{1,n}, h_{2,n}, \ldots)$, nor did they suggest a unique parametrization of extremes.

Pitman [10] studied recently compositions which are very different from simply ordered: they satisfy $h_i = 0,\ i = 1, 2, \ldots$, and the collection of classes is isomorphic, as an ordered set, to the set of rational numbers. Pitman's compositions are derived from excursion intervals of recurrent random processes like Brownian motion by a different version of ordered paintbox construction.

In this paper we refine the general characterization result of [4]. We prove that an exchangeable composition of $\mathbb{N}$ can be derived from a random open subset of [0, 1] via the construction used in [10]. Our basic observation is that this open set arises as a limit of the interval partitions encoding $(h_{1,n}, h_{2,n}, \ldots)$. We show that the set of extreme composition structures is homeomorphic to the family $\mathscr{U}$ of open subsets of [0, 1], provided $\mathscr{U}$ is properly topologized. It follows that using a special class of *uniformized* measures suffices to make the representation of Donnelly and Joyce unique.

The present study fits the general framework of questions around exchangeability surveyed in [1]. From another point of view, this is a particular instance of the problem about the limiting shape of a random combinatorial object as outlined in [12].

2. Compositions of numbers and sets. Let $\mathscr{N}_n$ be the set of compositions of the natural number n; each $\eta \in \mathscr{N}_n$ is an ordered collection of natural numbers $(n_1, \ldots, n_k)$ with sum n. The number of parts k may vary. We associate with η an interval partition of (0, 1):

$$\gamma(\eta) := \left(0, \frac{n_1}{n}\right) \cup \left(\frac{n_1}{n}, \frac{n_1+n_2}{n}\right) \cup \cdots \cup \left(\frac{n_1+\cdots+n_{k-1}}{n}, 1\right) \tag{1}$$

This can be regarded as a normalized version of the graphical representation of compositions described in [2]. Clearly, $\gamma(\eta)$ encodes the same information

as the vector of class sizes. There are 2^{n-1} compositions of n, because any of the points $1/n, \ldots, (n-1)/n$ may be the left endpoint of one of the partition intervals or not.

Suppose n identical objects are partitioned into classes with sizes given by η, in the sense that, for each $i \geq 1$, there are n_i objects in the ith class. Draw a random sample without replacement of size $m < n$ and put the classes represented in the sample in the order inherited from η. Let $q(\eta, \xi)$, for $\xi \in \mathcal{N}_m$, be the probability that the resulting composition of m is ξ.

Let Π_n be the class of probability distributions on $\mathcal{N}_n$. A *composition structure* is a sequence (p_n) which satisfies $p_n \in \Pi_n$ and

$$p_m(\xi) = \sum_{\eta \in \mathcal{N}_n} q(\eta, \xi) p_n(\eta) \tag{2}$$

for all $m < n,\ n = 1, 2, \ldots$ and $\xi \in \mathcal{N}_m$. The set Π of all composition structures is a projective limit of finite dimensional simplexes Π_n, with projections given by (2), and as such is a Choquet simplex. By Choquet theory, any element of Π can be uniquely represented as a mixture of the extreme points of Π. It is therefore important to describe the set of extreme points.

Drawing a sample of m objects from the set of size n amounts to deleting, one-by-one at random, $n - m$ objects. Thus, the values of q can be computed recursively from those for the case $m = n - 1$. In this special case, there are two possible kinds of changes from η to ξ.

Case 1. ξ is derived from η by decrementing some n_i by 1, where $n_i > 1$.

Case 2. ξ is derived from η by deleting a 1 in a subsequence of 1's (such as $1, 1, \ldots, 1$) of length j.

It is easy to see that

$$q(\eta, \xi) = \begin{cases} n_i/n, & \text{in Case 1,} \\ j/n, & \text{in Case 2,} \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

An alternative approach to composition structures involves sets with distinct elements.

A composition of the set $[n] := \{1, \ldots, n\}$ is an ordered collection of disjoint nonempty subsets, say $s = (A_1, \ldots, A_k)$, with $\cup A_i = [n]$. Let $\mathcal{S}_n$ be the collection of all compositions of $[n]$. For $s \in \mathcal{S}_n$, the vector of class sizes,

$$l(s) := (\#A_1, \ldots, \#A_k),$$

is a composition of the natural number n; we call $l(s)$ the *configuration* of s. For $\eta = (n_1, \ldots, n_k)$, there are

$$D(\eta) := \frac{n!}{n_1! \cdots n_k!} \tag{4}$$

compositions $s \in \mathcal{S}_n$ with configuration $l(s) = \eta$.

Let ρ_n: $\mathcal{S}_n \to \mathcal{S}_{n-1}$ be the restriction mapping which respects the order of classes. For $s \in \mathcal{S}_n$ and $r = \rho_n(s) \in \mathcal{S}_{n-1}$, r is derived from s either by removing the element n from one of classes of s with more than one element

or by deleting a singleton class $\{n\}$. For $\xi \in \mathscr{N}_{n-1}$, $\eta \in \mathscr{N}_n$ and $r \in \mathscr{S}_{n-1}$ with $l(r) = \xi$, let

$$D(\xi, \eta) := \#\{s \in \mathscr{S}_n \colon l(s) = \eta,\ \rho_n(s) = r\}$$

be the number of extensions of r to a composition with configuration ξ. It is easily seen that this number depends on r only through ξ and is given by

$$D(\xi, \eta) = \begin{cases} 1, & \text{in Case 1,} \\ j, & \text{in Case 2,} \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

Let $\mathscr{S}$ be the projective limit of $(\mathscr{S}_n, \rho_n)$. Each sequence $(s_n) \in \mathscr{S}$ corresponds to a composition of $\mathbb{N}$ such that s_n is the restriction on $[n]$. Each $\mathscr{S}_n$ is finite, so that $\mathscr{S}$ endowed with product topology becomes a compact, totally disconnected space. A random composition of $\mathbb{N}$ is a random element of $\mathscr{S}$, that is, a sequence (S_n) of random variables defined on a common probability space such that the equality $\rho(S_n) = S_{n-1}$ holds with probability 1 for all $n \geq 2$. Invoke the Kolmogorov extension theorem to see that the probability distribution P of (S_n) is uniquely determined by the marginal distributions P_n of S_n, which must be consistent with respect to restrictions for different values of n.

Permutations of $[n]$ act naturally on the subsets of $[n]$ and thus also on $\mathscr{S}_n$. We call a random composition S_n of $[n]$ *exchangeable* if the distribution of S_n is invariant under all permutations of $[n]$. We call a random composition (S_n) of $\mathbb{N}$ exchangeable if each S_n is exchangeable. In other words, exchangeability of (S_n) amounts to the invariance under all one-to-one mappings $\mathbb{N} \to \mathbb{N}$ which move only finitely many points.

Note that two compositions of $[n]$ can be transformed into each other by a permutation of $[n]$ iff they have the same configuration. It follows that S_n is exchangeable iff, for each $\eta \in \mathscr{N}_n$, the conditional distribution of S_n given $l(S_n) = \eta$ is uniform over all $D(\eta)$ compositions with configuration η. This means, of course, that $l(S_n)$ is a sufficient statistic for the class of exchangeable probability distributions on $\mathscr{S}_n$.

The next proposition says that composition structures and exchangeable compositions of $\mathbb{N}$ are essentially the same objects.

PROPOSITION 1. *The formula*

$$p_n(\eta) = P(l(S_n) = \eta), \qquad \eta \in \mathscr{N}_n,\ n = 1, 2, \ldots \tag{6}$$

sets up a one-to-one correspondence between distributions of exchangeable random compositions of $\mathbb{N}$ *and composition structures.*

PROOF. Suppose (S_n) is exchangeable and let P_n be the distribution of S_n. Consistency means that the equality

$$P_{n-1}(S_{n-1} = r) = \sum_{s \in \mathscr{S}_n \colon\ \rho_n(s) = r} P_n(S_n = s)$$

holds for all $n \geq 2$ and $r \in \mathscr{S}_{n-1}$. By exchangeability, this is equivalent to

$$p_{n-1}(\xi) = \sum_{\eta \in \mathscr{N}_n} \frac{D(\xi, \eta) D(\xi)}{D(\eta)} p_n(\eta),$$

where (p_n) is defined by (6) and $\xi = l(S_{n-1})$. Computing the coefficient from (5) and (4), we see that (p_n) is a composition structure.

Conversely, suppose (p_n) is a composition structure. Using (6) to define exchangeable distributions P_n and reversing things, we see that the sequence (P_n) is consistent and therefore determines a probability measure on the Borel sigma algebra of $\mathscr{S}$. □

COROLLARY 2. *Equation (6) yields an affine homeomorphism between* Π *and the set of probability distributions for exchangeable random compositions of* $\mathbb{N}$.

We use in the sequel the notation $L_n := l(S_n)$ for the configurations induced by an exchangeable composition (S_n). The sequence $L_1, L_2, \ldots$ is a Markov chain, with cotransition probabilities given by q. The problem of describing composition structures can be seen as the problem of describing all Markov chains on $\cup \mathscr{N}_n$ with given cotransition probabilities. Another interpretation of the problem relies on a combinatorial construction of the underlying probability space in the spirit of "Brattelli diagrams" (see [5, 12]).

CONSTRUCTION 3. Consider a directed graph on $\mathscr{N}^* := \bigcup_{n \geq 0} \mathscr{N}_n$, where $\mathscr{N}_0 := \varnothing$. Think of $\mathscr{N}_n$ as the nth floor of the graph. If vertices ξ and η lie on adjacent floors and are as in Case 1 or 2, we join them by $D(\xi, \eta)$ edges [for convenience, set $D(\varnothing, 1) = 1$] and label the edges by integers from 1 to $D(\xi, \eta)$.

There is a one-to-one correspondence between $\mathscr{S}_n$ and the set of directed paths in the graph joining $\varnothing$ with a vertex on the nth floor. Indeed, fix $\eta \in \mathscr{N}_n$ and argue inductively. Suppose a path of length $n-1$ with endpoint ξ corresponds to $r \in \mathscr{S}_{n-1}$. Because the number of extensions of r to some $s \in \mathscr{N}_n$ with $l(s) = \eta$ is equal to $D(\xi, \eta)$, we can identify each extension with an edge connecting ξ and η; joining such an edge to the path we get an extended path of length n corresponding to some s. Note that in the case $D(\xi, \eta) > 1$, the edges joining ξ and η encode the position where the singleton class $\{n\}$ is to be inserted into r. In this terms, the restriction mapping ρ_n amounts to cutting the last edge.

Infinite directed paths in $\mathscr{N}^*$ which start at $\varnothing$ encode compositions of $\mathbb{N}$. The space of paths is compact in product topology. A random exchangeable composition corresponds to a probability measure on the space of infinite paths with the following property: given that η is the nth node of the path, all paths of length n which join $\varnothing$ with η are equiprobable.

REMARKS. (1) Terminology is difficult here. The notion of composition appears in combinatorics (see [2]). Donnelly and Joyce [4] speak of classes in-

duced by "ranking" of $\mathbb{N}$ and call composition structures "consistent ordered sampling distributions" (they miss Case 2 in the description of cotransition probabilities on page 241). Kerov and Vershik [7] called a measure on the path space of a graded graph "central" if it satisfies the equiprobability conditions as in Construction 3.

(2) Pitman's [11] scheme of labelling the classes by their minimal elements leads also to a graded graph of compositions. The basic difference is that in Pitman's scheme, when n grows, the newly appearing singleton classes are always posed at the end of the composition, while in our model they can be inserted elsewhere. Kerov [6] investigated the central measures on Pitman's graph by combinatorial methods.

3. Uniformized measures. Sufficiency considerations of the preceeding section show that the interval partition $\gamma(L_n)$ contains all essential information about exchangeable composition S_n. To discuss the convergence issue, as n grows, one needs to topologize the set of interval partitions. We will do this by representing an open subset of $[0,1]$ as a probability measure.

Let $\mathscr{U}$ be the class of open subsets of $[0,1]$. Each $u \in \mathscr{U}$ is represented canonically as a union of disjoint open *component intervals*.

DEFINITION 4. For $u \in \mathscr{U}$, the probability measure $\hat{u}$ on $[0,1]$ defined by the following conditions is called the *uniformized* measure associated with u.

(a) $\hat{u}([0,t]) = x - \delta$, if $(x-\delta, x) \subset u$ is a component interval and $t \in (x-\delta, x)$.

(b) $\hat{u}([0,t]) = t$ for $t \in [0,1]\backslash u$.

A generic uniformized measure is partly atomic and partly continuous, and either of these parts is sufficient to reconstruct the measure. It is easy to see that $\hat{u}$ puts zero mass on u and coincides with Lebesgue measure on the set remaining after cutting out all component intervals together with their endpoints. If $(x-\delta, x) \subset u$ is a component interval, then its right endpoint is an atom of mass δ; thus the total atomic part of $\hat{u}$ coincides with the Lebesgue measure of u. Clearly, $u([0,t]) \le t$ for $t \in [0,1]$. Two extremes are: $\widehat{\varnothing}$ =Lebesgue measure, and $\widehat{(0,1)}$ =unit mass at 1.

Let $\widehat{\mathscr{U}}$ be the space of uniformized measures endowed with the weak topology.

PROPOSITION 5. *$\widehat{\mathscr{U}}$ is closed and is therefore a metrizable compactum.*

PROOF. We wish to show that any sequence $(\hat{u}_n)$ of uniformized measures contains a subsequence converging weakly to a uniformized measure.

For any uniformized measure and positive ε there are finitely many atoms with masses greater than ε. Selecting a subsequence we can assume without loss of generality that, for any rational ε, each $\hat{u}_n$ has the same number, $k(\varepsilon)$ say, of atoms with masses greater than ε. Selecting a smaller subsequence we

can achieve that all these $k(\varepsilon)$ atoms, enumerated from the left to the right, have converging masses and locations, as $n \to \infty$. Setting $n \to \infty$ and then $\varepsilon \to 0$, use the limiting masses and locations to define the atomic part of a new uniformized measure, $\hat{u}$ say, associated with some $u \in \mathscr{U}$.

We will prove that convergence $\hat{u}_n([0,t]) \to \hat{u}([0,t])$ holds for any t not an endpoint of a component interval of u. Indeed, assume first that $t \notin u$ and fix ε. Then t is not covered by an interval of u_n with size greater than ε for all sufficiently large n. Therefore, condition (a) in Definition 4 implies $\hat{u}_n([0,t]) \geq t - \varepsilon$. Along with $\hat{u}_n([0,t]) \leq t$ and condition (a) this yields

$$\hat{u}_n([0,t]) \to t = \hat{u}([0,t]).$$

Another case: t is covered by a component interval $(x - \delta, x) \subset u$. For n large, t is covered by a component interval $(x_n - \delta_n, x_n) \subset u_n$. Denoting by x and δ the respective limits, we have

$$\hat{u}_n([0,t]) = x_n - \delta_n \to x - \delta = \hat{u}([0,t]).$$

Weak convergence of $\hat{u}_n$ to $\hat{u}$ becomes obvious by noting that convergence of distribution functions may fail for at most countably many values of t. □

Call a distribution function *uniformized* if it corresponds to a uniformized measure. Uniformized distribution functions arise as a result of probability integral transform as follows. Let X be a real random variable with right continuous distribution function F. Consider the random variable $F(X)$ and set

$$F^*(t) := P(F(X) \leq t).$$

Observe that F^* is uniformized, and the relevant open set u is found from

$$\hat{u}([0,t]) = F^*(t), \qquad t \in (0,1). \tag{7}$$

Geometrically, u is the union of flats of F^* (note that F^* has the same flats in $(0,1)$ as the inverse F^{-1}). For $t \in (0,1)$, we can write $F^*(t) = F_- \circ F^{-1}(t)$, where F_- is the left continuous version of F and F^{-1} the right continuous generalized inverse. Obviously, $F^{**} = F^*$. The equality $F^* = F$ holds iff F is uniformized.

Uniformized distributions are useful to characterize the order properties of random reals. To be precise, consider the sigma algebra $\mathscr{F}$ on $\mathbb{R}^\infty$ generated by the events $\{x_i \leq x_j\}; i, j \in \mathbb{N}$, where $\{x_1, x_2, \ldots\}$ is a generic element of $\mathbb{R}^\infty$. Each F induces in a natural way a measure on $\mathscr{F}$.

PROPOSITION 6. *Let F and G be distribution functions. The following are equivalent:*

(a) $F^* = G^*$;

(b) *F and G induce the same measure on $\mathscr{F}$;*

(c) *F and G can be obtained from each other by a monotone change of variable.*

PROOF. The proof of (a) $\Longleftrightarrow$ (c) is straightforward using inverses. (c) $\Rightarrow$ (b) follows by noting that the events $\{x_i \le x_j\}$ and $\{F(x_i) \le F(x_j)\}$ coincide. A more delicate assertion (b) $\Rightarrow$ (a) follows from Proposition 10. $\square$

This proposition says that the order properties of an iid sample are determined by masses and arrangement of atoms of F. This is plain for F continuous, because ties appear in an iid sample with probability 0 and the sample can be uniquely ranked.

REMARKS. (1) Weak convergence $F_n \Rightarrow F$ does not imply convergence $F_n^* \Rightarrow F^*$. Indeed, for F_n continuous and F atomic, all F_n^* are uniform, while F^* is atomic.

(2) Elementary analysis shows that the topology on $\mathscr{U}$ introduced here coincides with the topology induced by the Hausdorff distance for the complement closed sets.

4. Ordered paintbox process. The following ordered version of Kingman's paintbox process was introduced in [10].

CONSTRUCTION 7. Fix $u \in \mathscr{U}$ and let $R_1, R_2, \ldots$ be iid uniform $[0,1]$ random variables. Denote by $R_{1:n} \ldots, R_{n:n}$ the increasing sequence of order statistics of the first n variables. With probability 1 there is a unique permutation σ_n: $[n] \to [n]$ satisfying $R_{\sigma_n(1)} = R_{1:n}, \ldots, R_{\sigma_n(n)} = R_{n:n}$. The classes of S_n are defined inductively for $i = 1, \ldots, n$ by the rule: $\sigma_n(1)$ is in the first class of S_n; $\sigma_n(i)$ is attached to the same class as $\sigma_n(i-1)$ iff both $R_{i-1:n}$ and $R_{i:n}$ fall in the same component interval, otherwise $\sigma_n(i)$ starts the next class of S_n not represented by the integers $\sigma_n(1), \ldots, \sigma_n(i-1)$.

With probability 1, each class of (S_n) in this construction is either a singleton or infinite. The integers i with $R_i \notin u$ constitute singleton classes. Two integers i and j are in one composition class iff R_i and R_j fall in the same component interval. Exchangeability of (S_n) follows at once from exchangeability of $R_1, R_2, \ldots$. Note that the order of classes is induced by the standard ordering of $[0,1]$, so that the minimum of $R_1, \ldots, R_n$ is always in the first class of S_n.

We will use in the sequel the notation P^u for the distribution of (S_n) corresponding to u and (p_n^u) for the related composition structure.

Donnelly and Joyce [4] suggested the following model.

CONSTRUCTION 8. Let $X_1, X_2, \ldots$ be iid random variables and $X_{1:n}, \ldots, X_{n:n}$ be the order statistics of the first n. Define σ_n to be a permutation of $[n]$ with $X_{\sigma_n(1)} = X_{1:n}, \ldots, X_{\sigma_n(n)} = X_{n:n}$. Use the rule: $\sigma_n(1)$ is in the first class of S_n; $\sigma_n(i)$ is attached to the same class as $\sigma_n(i-1)$ iff $X_{\sigma_n(i)} = X_{\sigma_n(i-1)}$; otherwise $\sigma_n(i)$ starts the next class.

Of course, Construction 8 yields a fine composition of $\mathbb{N}$, such that all classes are singletons, iff the X's are continuously distributed. In general, the permutation σ_n is not unique, but this does not matter.

Both constructions describe the same class of compositions. Denote by F the distribution function of the X's. Let F^* be the uniformization of F and u the union of flats of F^*, so that (7) holds. To establish the equivalence, recall the well-known fact that $F^{-1}(R_i)$ and X_i have the same distributions and note that the following events coincide:

$$\{F^{-1}(R_i) = F^{-1}(R_j)\} = \{F^*(R_i) = F^*(R_j)\}$$
$$= \{R_i \text{ and } R_j \text{ fall in one interval}\}.$$

Next is the law of large numbers for the Donnelly–Joyce–Pitman construction. The role of "empirical distribution" is played by the random uniformized measure corresponding to the interval partition $U_n := \gamma(L_n)$.

PROPOSITION 9. *Let (S_n) be an exchangeable random composition associated with some $u \in \mathscr{U}$, as in Construction 7. As $n \to \infty$, $\widehat{U}_n$ converges weakly to $\hat{u}$ a.s. Henceforth, the composition structure derived via Construction 8 can be parametrized uniquely by uniformized distribution functions.*

PROOF. Use Construction 8 with some F satisfying (7). One can see that

$$\widehat{U}_n([0,t]) = \frac{1}{n}\sum_{i=1}^{n} 1_{[0,t]}(F_n(X_i)),$$

where F_n is the empirical d.f. of $X_1, \ldots, X_n$. By the law of large numbers for independent Bernoulli trials,

$$\frac{1}{n}\sum_{i=1}^{n} 1_{[0,t]}(F(X_i)) \to F^*(t) \quad \forall\, t \in (0,1).$$

Weak convergence follows from these formulas and the Glivenko–Cantelli theorem.

The second assertion follows from the first and Proposition 6. □

For open sets of the form

$$u = (0, x_1) \cup (x_1, x_1 + x_2) \cup (x_1 + x_2, x_1 + x_2 + x_3) \cup \cdots,$$

with $x_i \geq 0$, $\sum x_i = 1$, we have

$$p_n^u(n_1, \ldots, n_k) = \sum_{j_1 < \cdots < j_k} D(n_1, \ldots, n_k)(x_{j_1})^{n_1} \cdots (x_{j_k})^{n_k}.$$

This family of open sets is dense in $\mathscr{U}$.

Keep in mind permanently that we have topologized the collection of open sets through correspondence $\mathscr{U} \leftrightarrow \widehat{\mathscr{U}}$.

PROPOSITION 10. *The correspondence* $u \leftrightarrow (p_n^u)$ *is continuous and one-to-one.*

PROOF. Continuity of composition structure means continutity of each p_n^u. Let u_ε be the open set derived from u by removing all subintervals of length smaller than ε. Suppose another open set v is sufficiently close to u, so that v_ε and u_ε have the same number, say $k(\varepsilon)$, of intervals whose respective $2k(\varepsilon)$ endpoints are close to each other. The probability $p_n^{u_\varepsilon}$ is easily seen to be a polynomial in these $2k(\varepsilon)$ variables; therefore $p_n^{v_\varepsilon}$ is close to $p_n^{u_\varepsilon}$. Those sample points $R_1, \dots, R_n$ which fall into the complement to u_ε make up singleton classes of the composition associated with u_ε. The probability that a uniform $[0,1]$ sample hits twice any of the intervals of u with length less than ε is of the order $O(\varepsilon)$ uniformly in u, that is, the sample points falling into the difference set $u \backslash u_\varepsilon$ are likely to form singleton classes. It follows that

$$p_n^u(\eta) = p_n^{u_\varepsilon}(\eta) + O(\varepsilon).$$

The same holds also for $p_n^v(\eta)$, therefore this probability is close to $p_n^u(\eta)$ and the continuity follows.

We shall prove now that u can be recovered from (p_n^u). Indeed, set

$$\alpha_n = n \int_0^1 t^{n-1}\, d\hat{u}(t).$$

The sequence (α_n) determines measure $\hat{u}$ uniquely, as in the Hausdorff moment problem. On the other hand, α_n can be expressed through composition structure as the expected size of the last part of $\eta = (n_1, \dots, n_k)$:

$$\alpha_n = \sum_{\eta \in \mathcal{N}_n} n_k p_n^u(\eta) \tag{8}$$

To see this, denote by A the last class of S_n and use exchangeability of the R's in Construction 7 to obtain

$$\alpha_n = nP^u(1 \in A) = nP(1 \in A,\ R_1 \in u^c) + nP(1 \in A,\ R_1 \in u).$$

Assume first that R_1 falls into u^c, then 1 is in the last class iff it is the sample maximum, whence

$$P(1 \in A,\ R_1 \in u^c) = P(R_1 = R_{n:n},\ R_1 \in u^c) = \int_{u^c} t^{n-1}\, dt = \int_{u^c} t^{n-1}\, d\hat{u}(t).$$

Alternatively, R_1 may fall into a component interval $(x-\delta, x) \subset u$; then 1 is in the last class iff other points $R_2, \dots, R_n$ fall into $[0, x]$: an event of probability δx^{n-1}. Consequently,

$$P(1 \in A,\ R_1 \in u) = \int_u t^{n-1}\, d\hat{u}(t),$$

whence (8) and the moments can be derived from (p_n^u). □

5. The representation. Let $\mathscr{U}_n$ be the subclass of $\mathscr{U}$ consisting of open sets such that the endpoints of component intervals are integer multiples of $1/n$, and the total length of the intervals is 1. Obviously, (1) sets up a one-to-one correspondence $\mathscr{N}_n \leftrightarrow \mathscr{U}_n$. With a slight abuse, we will denote by p_n both a generic distribution on $\mathscr{N}_n$ and its image on $\mathscr{U}$.

A composition structure (p_n) can be considered as a sequence of distributions on $\mathscr{U}$. Similarly, an exchangeable composition (S_n) induces a random sequence $U_n := \gamma(L_n)$. A fundamental fact is that U_n converges.

THEOREM 11. *Suppose (S_n) is an exchangeable composition of $\mathbb{N}$. Then the sequence (U_n) converges almost surely to a random element $U \in \mathscr{U}$. The conditional distribution of (S_n) given $U = u$ is P^u, as in Construction 7.*

PROOF. Pick n, $\xi \in \mathscr{N}_n$ and an arbitrary sequence $\eta_N \in \mathscr{N}_N$, $N = 1, 2, \ldots$. Set $u_N = \gamma(\eta_N)$, and let $p_n^{u_N}$ be the distribution on $\mathscr{N}_N$ associated with u_N, as in Construction 7. Obviously, $u_N \in \mathscr{U}_N$.

The probability that a random point from the uniform $[0, 1]$ distribution discovers a component interval of u_N is the same as the probability of picking this interval by the following scheme: select j uniformly at random from $\{1, 2, \ldots, N\}$ and mark the component interval covering $((j-1)/N, j/N)$. This means that, for this special case (i.e., when the open set is in $\mathscr{U}_N$), the construction yields ξ with the same probability as via repeated sampling from the discrete uniform distribution. On the other hand, if the sampling is *without* replacement, the probability of ξ is $q(\eta_N, \xi)$. Estimating the probability of ties in the sampling with replacement we get for large N,

$$q(\eta_N, \xi) = p_n^{u_N}(\xi) + O(n^2/N). \tag{9}$$

Now let Z_n be the size of the last class of S_n. Using (9) and arguing as in the proof of Proposition 10 we obtain

$$E(Z_n|L_N) = n \int_0^1 t^{n-1}\, d\widehat{U}_N + O(1/N), \tag{10}$$

where $\widehat{U}_n$ is the random uniformized measure associated with U_n. The Markov property of (L_N) implies that the conditional expectation in the LHS of (10) is a reversed martingale and as such must converge as $N \to \infty$. Therefore, the RHS of (10) converges as well. Recall that convergence of all moments in the Hausdorff moment problem implies the convergence of measures. Therefore, there is a weak limit for $\widehat{U}_n$, which by Proposition 5 must be also a uniformized measure, say $\widehat{U}$.

Apply the martingale convergence once again to conclude that $E(Z_n|L_N) \to E(Z_n|U)$. The conditional distribution of (L_n) given U is uniquely determined by the conditional expectations of $E(Z_n|U)$, as in Proposition 10. □

Next is our main result. In words: any composition structure is a unique mixture of ordered paintbox processes.

COROLLARY 12. *Let (p_n) be a composition structure. There exists a unique probability measure μ on $\mathscr{U}$ such that the representation*

$$p_n(\eta) = \int_{\mathscr{U}} p_n^u(\eta)\, d\mu(u) \tag{11}$$

holds for all n and $\eta \in \mathscr{N}_n$. As measures on $\mathscr{U}$, the distributions p_n converge to μ weakly.

Along with Propositions 5 and 9, this implies that the parametrization of extremes by open sets is also nice topologically.

COROLLARY 13. *The set of extreme composition structures is homeomorphic to $\mathscr{U}$.*

REMARKS.

1. We get also some information about asymptotic combinatorics of compositions of large integers. For $\xi \in \mathscr{N}_n$, $\eta_N \in \mathscr{N}_N$ and $N > n$ define the *dimension function* $D(\xi, \eta_N)$ to be the number of directed paths in the compositions graph (see Construction 3) with endpoints ξ and η_N. Equivalently, this is the number of extensions of ξ to η_N. For the cotransition probability we have

$$q(\eta_N, \xi) = P(L_n = \xi | L_N = \eta_N) = \frac{D(\xi, \eta_N)}{D(\eta_N)}.$$

   As in the proof of Theorem 11, the limit of the RHS exists for all n and ξ iff $\gamma(\eta_N)$ converge. A similar fact for Pitman's graph was established by Kerov [6] by direct analysis of the dimension function.
2. Note that Proposition 9 is a particular case of Corollary 12. The representation (11) could be derived from this proposition and Theorem 8 of [4].
3. The method of moments used here suggests also a new proof of Kingman's representation theorem (see [8]). Indeed, given that the decreasing class sizes of an exchangeable partition of $[N]$ are $(f_{1,N}, f_{2,N}, \ldots)$, the probability that the partition of $[n]$ is the single-class partition equals $\sum_i (f_{i,N}/N)^n + O(1/N)$. Along with martingale convergence, this implies the existence of limits for the relative frequencies $f_{i,N}/N$.

6. The simply ordered case. For exchangeable composition (S_n), let $(h_{1,n}, h_{2,n}, \ldots)$ be the vector of class sizes of S_n complemented by zeros, as in the Introduction. The following assertion is a reformulation of Theorem 5 from [4].

PROPOSITION 14. *For $i = 1, 2, \ldots$, the limit $h_i := h_{i,n}/n$ exists with probability 1. Furthermore, if $h_i = 0$ for some i then also $h_j = 0$, for $j \geq i$.*

PROOF. It is sufficient to consider only the extreme compositions, because the general case is a mixture. Let P^u be the distribution of (S_n), as in Construction 7.

Assume first that 0 is not an isolated point of u^c. Fix i and ε. The law of large numbers guarantees that the set $\{j\colon R_j < \varepsilon\}$ contains the first i classes of S_n for all sufficiently large n. Hence $(h_{1,n}+\cdots+h_{i,n})/n < \varepsilon$ and the relative frequencies converge to 0.

If 0 is an isolated point of u^c then u has a component interval starting at 0. Letting h_1 be the length of this interval, we have $h_{1,n}/n \to h_1$. Argue inductively: if h_1 is a concentration point of u^c then $(h_{2,n}+h_{3,n}+\cdots)/n \to 0$; otherwise the point h_1 starts the second most left interval and $h_{2,n}/n$ converges, and so on. □

Recall that a random composition of $\mathbb{N}$ is *simply ordered* if, with probability 1, there is a well-defined first, second, and so on, class, as defined in the Introduction.

PROPOSITION 15. *An exchangeable composition of* $\mathbb{N}$ *is simply ordered iff the associated composition structure satisfies the condition*

$$\lim_{j\to\infty} \sum \left(\frac{n_1+\cdots+n_j}{n} \right) p_n(\eta) = 1,$$

where the sum is extended over all $\eta = (n_1, \ldots, n_k) \in \mathcal{N}_n$, *and* $n_i = 0$ *for* $i > k$.

PROOF. In the simply ordered case, the complement to the directing open set cannot have concentration points different from 1. This holds iff $\sum h_i = 1$ or, equivalently, $E\sum h_i = 1$. □

It would be interesting to know more about the limiting properties of $(h_{1,n}, h_{2,n}, \ldots)$. It follows from the results of [3] that the asymptotic behavior of $h_{1,n}$ can be rather irregular. Another question of interest is a characterization of uniformized measures in terms of moments.

Acknowledgments. I am indebted to Erich Berger for the idea of direct proof of Proposition 9 and to a referee for helpful comments. My special thanks go to Jim Pitman for the suggestion to clarify a topological point (Remark 2 in Section 3).

REFERENCES

[1] ALDOUS, D. J. (1985). Exchangeability and related topics. *Lecture Notes in Math.* 1117 1–198. Springer, Berlin.

[2] ANDREWS, G. E. (1976). *The Theory of Partitions*. Addison-Wesley, Reading, MA.

[3] BARYSHNIKOV, Y., EISENBERG, B. and STENGLE, G. (1995). A necessary and sufficient condition for the existence of the limiting probability of a tie for the first place. *Statist. Probab. Lett.* 23 211–220.

[4] DONNELLY, P. and JOYCE, P. (1991). Consistent ordered sampling distributions: characterization and convergence. *Adv. in Appl. Probab.* 23 229–258.

[5] KEROV, S. V. (1989). Combinatorial examples in the theory of AF-algebras. In *Zapiski Nauchnych Seminarov LOMI* 172 55–67. Nauka, Leningrad. (In Russian.)
[6] KEROV, S. V. (1995). Subordinators and permutation actions with a quasi-invariant measure. In *Zapiski Nauchnych Seminarov POMI* 223 181–218. Nauka, Sankt Petersburg. (In Russian.)
[7] KEROV, S. V. and VERSHIK, A. M. (1981). Asymptotic theory of characters of the symmetric group. *Functional Anal. Appl.* 15 246–255.
[8] KINGMAN, J. F. C. (1978). The representation of partition structures. *J. London Math. Soc.* 18 374–380.
[9] KINGMAN, J. F. C. (1982). The coalescent. *Stochastic Process. Appl.* 13 235–248.
[10] PITMAN, J. (1995). Partition structures derived from Brownian motion and stable subordinators. *Bernoulli*. To appear.
[11] PITMAN, J. (1995). Exchangeable and partially exchangeable random partitions. *Probab. Theory Related Fields* 102 145–158.
[12] VERSHIK, A. (1995). Asymptotic combinatorics and algebraic analysis. In *Proceedings of the International Congress on Mathematics, 1994, August 3–11, Zürich* 1384–1394. Birkhäuser, Basel.

INSTITUTE OF MATHEMATICAL STOCHASTICS
UNIVERSITY OF GÖTTINGEN
LOTZESTRASSE 13
37083 GÖTTINGEN
GERMANY
E-MAIL: gnedin@math.uni-goettingen.de