Gibbs distributions for random partitions generated by a fragmentation process

Nathanaël Berestycki and Jim Pitman*

University of British Columbia and E.N.S. Paris,
and Department of Statistics, U.C. Berkeley

December 14^{th}, 2005

Abstract

In this paper we study random partitions of $\{1, \ldots, n\}$ where every cluster of size j can be in any of w_j possible internal states. The Gibbs (n, k, w) distribution is obtained by sampling uniformly among such partitions with k clusters. Gibbs distributions arise naturally as equilibrium distributions of reversible coagulation - fragmentation processes. The goal of this work is to study random processes where at step k the process has the Gibbs (n, k, w) distribution, so that this microscopical equilibrium is subject to irreversible fragmentation as time evolves. It is not always possible to combine those two features, and in our main result we identify those weight sequences w_j for which such a process exists subject to some simplifying assumptions. In this case the time-reversed process turns out to be the discrete Marcus-Lushnikov coalescent process with affine collision rate $K(x, y) = a+b(x+y)$ for some real numbers a and b.

Keywords Fragmentation processes, Gibbs distributions, Marcus-Lushnikov processes, Gould convolution identities.

*Research supported in part by N.S.F. Grant DMS-0405779

1 Introduction

Gibbs models for random partitions generated by random processes of coagulation and fragmentation have been widely studied ([26], [27], [28], [17]). They typically arise as equilibrium distributions of time-reversible processes of coagulation and fragmentation (see for instance [7], and [3] for general results about exchangeable fragmentation-coalescence processes in equilibrium). There is a much smaller literature in which Gibbs models are derived from an irreversible Markovian coagulation process [14]. This paper presents a Gibbs model for an irreversible Markovian fragmentation process. While Gibbs models for physical processes of fragmentation have been treated before, such models typically involve the evolution of an equilibrium distribution depending on a time parameter t, with the equilibrium moving towards a more fragmented state as t increases. At the microscopic level such an evolution still involves both fragmentation and coagulation. The point here is to provide a rigorous Markovian model of irreversible fragmentation at the microscopic level with no possibility of coagulation allowed.

The simplest way to describe the process treated here is to specify its time reversal. This turns out to be to the Marcus-Lushnikov coalescent process with collision rate kernel $K_{x,y} = a + b(x+y)$ for some constants a and b, where $K_{x,y}$ represents rate of collisions between clusters of x particles and clusters of y particles. This model was solved by Hendriks et al. [14], who showed that the distribution at time t in such a coalescent process started from a monodisperse initial condition is a mixture of microcanonical Gibbs distributions with mixing coefficients depending on t. Here, what is essentially the same model, modulo time reversal and a formulation in discrete rather than continuous time, is derived from a different set of assumptions describing the evolution of the process with time running in the direction of fragmentation. The probabilistic link between the two sets of assumptions is a time reversal calculation using Bayes rule. The most interesting feature of this calculation is that starting from a natural recursive assumption for the fragmentation process in terms of Gibbs distributions, there is only one-parameter family of possible solutions to the problem, with the parameter corresponding to the ratio of the two parameters a and b in the collision rate kernel of the reversed time process.

1.1 Canonical and microcanonical Gibbs distribution

Typically, the state of a coagulation/fragmentation process is represented by a random partition of n, that is a random variable with values in the set $\mathcal{P}_{[n]}$ of all partitions of n. In later sections of this paper the state of the process will be represented rather as a random partition of the set $\{1, 2, \ldots, n\}$, as this device simplifies a number of calculations. But the rest of this introduction follows the more common convention of working with partitions of n. Let

$$\lambda = 1^{c_1} 2^{c_2} \cdots n^{c_n} \tag{1}$$

denote a typical partition of n. Regarding the state of the system as a partition of n particles into clusters of various sizes, the state λ in (1) indicates that there are c_j clusters of size j for each $1 \leq j \leq n$. Note that $\sum_j jc_j = n$, the total number of particles. The total number of clusters is $k := \sum_j c_j$. The numbers $c_1, c_2, \ldots, c_j \ldots$ may be called numbers of *monomers, dimers,* $\ldots$ *j-mers*, or numbers of *singletons, doubletons,* $\ldots$ *j-tons*. The Gibbs model most commonly derived from equilibrium considerations is the *canonical Gibbs distribution on partitions of* n *with weight sequence* (w_j) defined by

$$P(\lambda \mid n, ; w_1, w_2, \ldots) = \frac{n!}{Y_n} \prod_{i=1}^{k} \frac{1}{c_i!} \left(\frac{w_i}{i!}\right)^{c_i} \tag{2}$$

where

$$Y_n = Y_n(w_1, w_2, \ldots) \tag{3}$$

is a normalization constant. This polynomial in $w_1, w_2, \ldots$ is known in the combinatorics literature as the *complete Bell* (or *exponential*) *polynomial* [5]. In the physics literature the Gibbs formula (2) is commonly written in terms of $x_i = w_i/i!$ instead of w_i, and the polynomial

$$Z_n(x_1, x_2, \ldots) := n! Y_n(1!x_1, 2!x_2, \ldots) \tag{4}$$

is called the *canonical partition function*. For textbook treatments of such models, and references to earlier work see [24]. Typically, the canonical Gibbs distribution (2) is derived either from thermodynamic considerations, or from a set of detailed balance equations corresponding to a reversible equilibrium between processes of fragmentation and coagulation. In the latter case the canonical Gibbs distribution is represented as the equilibrium distribution of a time-reversible Markov chain with state space $\mathcal{P}_{[n]}$.

Conditioning a canonical Gibbs distribution on the number of clusters k yields a corresponding *microcanonical Gibbs distribution* for each $1 \leq k \leq n$. This distribution assigns to the partition λ displayed in (1) the probability

$$P(\lambda \mid n, k; w_1, w_2, \ldots) = \frac{n!}{B_{n,k}} \prod_{i=1}^{k} \frac{1}{c_i!} \left(\frac{w_i}{i!}\right)^{c_i} \tag{5}$$

where $B_{n,k} = B_{n,k}(w_1, w_2, \ldots)$ is a *partial Bell (or exponential) polynomial*, and

$$Z_{n,k}(x_1, x_2, \ldots) := n! B_{n,k}(1!x_1, 2!x_2, \ldots) \tag{6}$$

is known as a *microcanonical partition function*. A great many expressions, representations and recursions for these polynomials $B_{n,k}$ and $Z_{n,k}$ are known [5, 24]. These formulae are useful whenever the weight sequence (w_j) is such that the associated polynomials admit an explicit formula as functions of n and k, or can be suitably approximated (see e.g. [16]). Some of these results are reviewed in [24]. The class of weight sequences (w_j) for which the microcanonical Gibbs model is "solvable", meaning there is an explicit formula for the $B_{n,k}$, is quite large.

In connection with the study of irreversible partition-valued processes, it has been found in several cases (see below) that the distribution of the process at time t is a probabilistic mixture over k of microcanonical Gibbs distributions, that is to say a probability distribution of the form

$$P(\lambda) = \sum_{k=1}^{n} p_{nk} P(\lambda \mid n, k; w_1, w_2, \ldots) \tag{7}$$

where p_{nk} represents the probability that λ has k components, so $p_{n,k} \geq 0$ and $\sum_{k=1}^{n} p_{n,k} = 1$, and both $p_{n,k}$ and the weight sequence w_i may be functions of t. There seems to be no standard term in the literature for a distribution on partitions of this form, which will be called here a *modified Gibbs* distribution. For example, Lushnikov [20] showed that the coalescent model with monodisperse initial condition and collision rates $K_{x,y} = xf(y) + yf(x)$ leads to such modified Gibbs distributions. Hendriks et. al [14] showed that this is also the case for $K_{x,y} = a + b(x+y)$ for constants a and b.

1.2 Organization and summary of the paper.

The rest of this paper is organized as follows. Section 2 presents some background material, and introduces the formalism of Gibbs distributions over partitions of the set $\{1, 2, \ldots, n\}$. Results for irreversible fragmentation processes are presented in Section 3. Section 4 presents the main result of the paper. This result states that, under an additional set of assumption (most notably, the *linear selection rule*), it is possible to construct a Gibbs fragmentation process with weight sequence (w_j) if and only if $w_j = \prod_{m=2}^{j}(mc + jb)$ for some constants b and c: in this case it is shown that the time-reversal of the fragmentation process is the discrete Marcus-Lushnikov coalesent with affine coalescent rate: $K(x, y) = a + b(x + y)$. This theorem leaves out an important case, which is that of the sequence $w_j = (j-1)!$. In section 5, we approach this problem from the angle of Kingman's coalescent process and the Ewens sampling formula. In particular we construct a continuous analogue of the desired process. However, curiously, we show that the existence of this process with discrete time cannot be obtained by taking the discrete skeleton of the continuous process, so that its existence remains an open question.

2 Preliminaries

Let $[n]$ denote the set $\{1, \ldots, n\}$. A *partition of* n is an unordered collection of non-empty disjoint subsets of $[n]$ whose union is $[n]$. A generic partition of $[n]$ into k sets will be denoted

$$\pi_k = \{A_1, \ldots, A_k\} \quad (1 \leq k \leq n) \tag{8}$$

Let $\mathcal{P}_{[n]}$ denote the set of all partitions of $[n]$, and let $\mathcal{P}_{[n,k]}$ be the subset of $\mathcal{P}_{[n]}$ comprising all partitions of $[n]$ into k components. Given a sequence of weights $(w_j, j =$

$1, 2, \cdots$) define the *microcanonical Gibbs distribution on* $\mathcal{P}_{[n,k]}$ *with weights* $(w_1, w_2, \ldots)$ to be the probability distribution on $\mathcal{P}_{[n,k]}$ which assigns to each partition π_k as in (8) the probability

$$p(n_1, \ldots, n_k \mid k) = \frac{1}{B_{n,k}} \prod_{i=1}^{n} w_{n_i} \tag{9}$$

where $n_i := \#A_i$ denotes the number of elements of A_i and

$$B_{n,k} := B_{n,k}(w_1, w_2, \ldots) := \sum_{\pi_k} \prod_{i=1}^{k} w_{n_i(\pi_k)} \tag{10}$$

where for the sum is over all $\pi_k \in \mathcal{P}_{[n,k]}$ and the $n_i(\pi_k)$ for $1 \leq i \leq k$ are sizes of the components of π_k in some arbitrary order. Given a partition π of $[n]$, the *corresponding partition* λ *of* n is $\lambda := 1^{c_1} 2^{c_2} \cdots n^{c_n}$ as in (1) where c_i is the number of components of π of size i. For each vector of non-negative integer counts $(c_1, \ldots, c_n)$ with $\sum_i i c_i = n$ the number of partitions π of $[n]$ corresponding to the partition $1^{c_1} 2^{c_2} \cdots n^{c_n}$ of n is well known to be

$$\frac{n!}{\prod_i c_i! (i!)^{c_i}} \tag{11}$$

The probability distribution on partitions of n induced by the microcanonical Gibbs distribution on $\mathcal{P}_{[n,k]}$ with weights $(w_1, w_2, \ldots)$ therefore identical to the microcanonical Gibbs distribution on $\mathcal{P}_n$ with the same weights $(w_1, w_2, \ldots)$ as defined in (5), and there is the following standard expression for B_{nk} [5]

$$B_{n,k} = n! \sum_{\lambda_k} \prod_{i=1}^{k} \frac{1}{c_i!} \left(\frac{w_i}{i!} \right)^{c_i} \tag{12}$$

where the sum is over all partitions λ_k of n into k components, and $c_i = c_i(\lambda_k)$ is the number of components of λ_k of size i. Thus transferring from Gibbs distributions on partitions of n into k components to Gibbs distributions on partitions of the set $[n]$ into k components is just a matter of keeping track the universal combinatorial factor (11).

2.1 Combinatorial and Physical Interpretations.

The following well-known interpretations provide both motivation and intuition for the study of Gibbs distributions and Bell polynomials. Suppose that n particles labelled by elements of the set $[n]$ are partitioned into *clusters* in such a way that each particle belongs to a unique cluster. Formally, the collection of clusters is represented by a partition of $[n]$. Suppose further that each cluster of size j can be in any one of w_j different *internal states* for some sequence of non-negative integers (w_j). Let the *configuration* of the system of n particles be the partition of the set of n particles into clusters, together with the assignment of an internal state to each cluster. For each partition π of $[n]$ with k components of sizes $n_1, \ldots, n_k$, there are $\prod_{i=1}^{k} w_{n_i}$ different configurations with that partition π. So $B_{n,k}(w_1, w_2, \ldots)$ defined by (10) gives *the number of configurations with k clusters*; the Gibbs distribution (9) with weight sequence (w_j) is *the distribution of the random partition of $[n]$ if all configurations with k clusters are equally likely*, and formula (5) describes the corresponding Gibbs distribution on partitions of n induced by the same hypothesis.

The term "internal state" suggests a physical interpretation which makes sense for an arbitrary sequence of numbers (w_j). But many particular choices of (w_j) have natural interpretations, both combinatorial and physical. In particular, the following four examples have been extensively studied. Many more combinatorial examples are known where Gibbs distributions arise naturally from an assumption of equally likely outcomes on a suitable configuration space. Related problems of enumeration and asymptotic distributions have been extensively studied [16, 19, 11, 12, 25].

2.2 Some important examples

We introduce here a few natural examples of Gibbs distributions and their combinatorial interpretations (for particular sequences of weights (w_j)) that motivate much of the subsequent work in the following sections.

Example 1. *Uniform distribution on partitions.* Take $w_j = 1$ for all j. Then the configuration is just a partition of $[n]$ so $B_{n,k}(1, 1, \ldots)$ is the number of partitions of $[n]$ into k components, known as a *Stirling number of the second kind*. The microcanonical Gibbs model corresponds to assuming all partitions of $[n]$ into k components are equally likely.

Example 2. *Uniform distribution on permutations.* Suppose that the internal state of a cluster C of size j is one of the $(j-1)!$ cyclic permutations of C. Then each configuration corresponds to a permutation of $[n]$. Therefore $B_{n,k}(0!, 1!, 2! \ldots)$ is the number of permutations of $[n]$ with k cycles, known as an *unsigned Stirling number of the first kind*. The microcanonical Gibbs model corresponds to assuming all permutations of n with k cycles are equally likely.

Example 3. Cutting a rooted random segment. Suppose that the internal state of a cluster C of size j is one of $j!$ linear orderings of the set C. Identify each cluster as a directed graph in which there is a directed edge from a to b if and only if a is the immediate predecessor of b in the linear ordering. Call such a graph a *rooted segment*. Then $B_{n,k}(1!, 2!, 3! \ldots)$ is the number of directed graphs labelled by $[n]$ with k such segments as its components. In the previous two examples, explicit formulae for the $B_{n,k}$ are fairly complicated. But this time there is a simple formula:

$$B_{n,k}(1!, 2!, 3! \ldots) = \binom{n-1}{k-1} \frac{n!}{k!} \tag{13}$$

is known as a *Lah number* [5, p. 135]. The Gibbs model in this instance is a variation of Flory's model for a linear polymerization process [10]. Another interpretation is

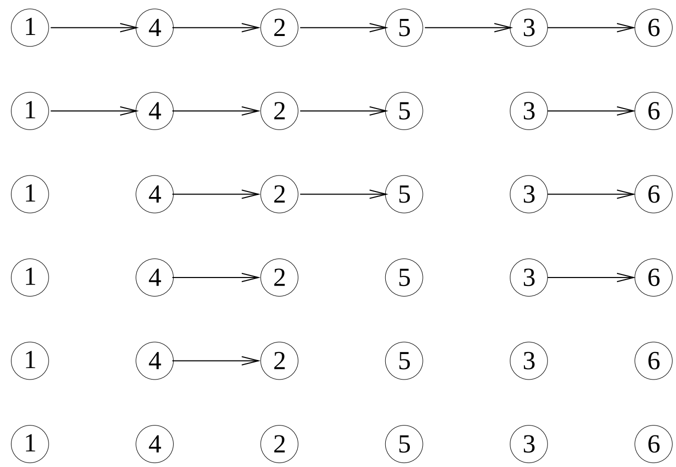

Figure 1: Cutting a rooted random segment

provided by Kingman's coalescent [1, 18]. It is easily shown in this case that a sequence of random partitions $(\Pi_k, 1 \leq k \leq n)$ such the Π_k has the Gibbs(n, k) distribution is obtained as follows. Let G_1 be a uniformly distributed random rooted segment labelled by $[n]$, and let G_k be derived from G_1 by deletion of a set of $k-1$ edges picked uniformly at random from the set of $n-1$ edges of G_1, and let Π_k be the partition induced by the components of G_k. If the $n-1$ edges of G_1 are deleted sequentially, one by one, the random sequence $(\Pi_1, \Pi_2, \ldots, \Pi_n)$ is a refining sequence of random partitions such that Π_k has the Gibbs$(n|k)$ distribution. This is illustrated in figure 1. The time-reversed sequence $(\Pi_n, \Pi_{n-1}, \ldots, \Pi_1)$ is then governed by the rules of *Kingman's coalescent*: conditionally given Π_k with k components, Π_{k-1} is equally likely to be any one of the $\binom{k}{2}$ different partitions of $[n]$ obtained by merging two of the components of Π_k. Equivalently, the sequence $(\Pi_1, \Pi_2, \ldots, \Pi_n)$ has uniform distribution over the set $\mathcal{R}_n$ of all refining sequences of partitions of $[n]$ such that the kth term of the sequence has k components. The consequent enumeration $\#\mathcal{R}_n = n!(n-1)!/2^{n-1}$ was obtained by Erdös et al [8]. The fact that Π_k determined by this model has the Gibbs(n, k) distribution with weight sequence $w_j = j!$ was obtained by Bayewitz et. al. [2] and Kingman [18].

Example 4. *Cutting a rooted random tree.* Suppose the internal state of a cluster C of size j is one of the j^{j-1} rooted trees labelled by C. Then $B_{n,k}(1^{1-1}, 2^{2-1}, 3^{3-1}, \ldots)$ is the number of forests of k rooted trees labelled $[n]$. This time again there is a simple formula for $B_{n,k}$. As a consequence of Cayley's enumeration of random forests [22, 5]

$$B_{n,k}(1^{1-1}, 2^{2-1}, 3^{3-1}, \ldots) = \binom{n-1}{k-1} n^{n-k} \tag{14}$$

The Gibbs model in this instance corresponds to assuming that all forests of k rooted

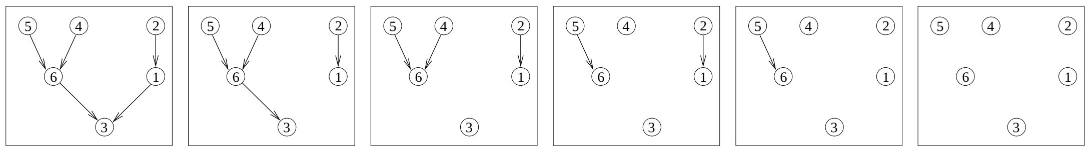

Figure 2: Cutting a rooted random tree with 5 edges

trees are equally likely. This model turns up naturally in the theory of random graphs and has been studied and applied in several other contexts. The coalescent obtained by reversing the process of deleting the edges at random is the *additive coalescent* as discussed in [23]. This is what is illustrated in figure 2.

3 Fragmentation Processes

Definition 5. Call a sequence of $\mathcal{P}_{[n]}$-valued random variables $(\Pi_1, \Pi_2, \ldots, \Pi_n)$ a *fragmentation process* or a *refining process* if with probability one the sequence (Π_k) is a refining sequence of partitions as k increases, and Π_k has k components. The time-reversed sequence $(\Pi_n, \Pi_{n-1}, \ldots, \Pi_1)$ will then be called a *coalescent.*

Definition 6. Call a $(\Pi_1, \Pi_2, \ldots, \Pi_n)$ a *Gibbs fragmentation process*, if there is some sequence of weights $(w_1, \ldots, w_{n-1})$ such that Π_k has the microcanonical Gibbs distribution (5) on $\mathcal{P}_{[n,k]}$ for every $1 \leq k \leq n$.

Similarly, call a $\mathcal{P}_{[n]}$-valued random process $(\Pi_t, t \in I)$, with index set I a subset of real numbers, a *fragmentation process* if with probability one both

(i) for every pair of times s and t in I with $s < t$ the partition Π_t is a refinement of Π_s, and

(ii) for each $1 \leq k \leq n$ there is some $t \in I$ such that Π_t has k components.

Condition (ii) implies that whenever a split occurs, the split is a binary split in which one and only one block of the partition splits in two, thereby incrementing the number of components by 1. This condition also forces Π_t to be the partition of $[n]$ with one component of size n for all sufficiently small t, and to to be the partition of $[n]$ into n singletons for all sufficiently large t.

Definition 7. Call a $(\Pi_t, t \in I)$ a *Gibbs fragmentation process with weights* $(w_1, \ldots, w_{n-1})$ if for every $t \in I$ and $1 \leq k \leq n$, the conditional distribution of Π_t given that Π_t has k components is is the microcanonical Gibbs distribution (5) on $\mathcal{P}_{[n,k]}$. Note that if $(\Pi_k, k \in [n])$ is such a Gibbs fragmentation process, then the unconditional distribution of Π_k is the microcanonical Gibbs distribution on $\mathcal{P}_{[n,k]}$, because condition condition (ii) implies that Π_k has k components with probability 1.

A basic problem, only partially solved in this paper, is the following:

Problem 8. *For which weight sequences* $(w_1, \ldots, w_{n-1})$ *does there exist a* $\mathcal{P}_{[n]}$*-valued Gibbs fragmentation process?*

The above definitions were made in terms of $\mathcal{P}_{[n]}$-valued processes, as this formalism that seems most convenient for computations with Gibbs distributions. Parallel definitions can be made in terms of $\mathcal{P}_n$-valued processes, using the partial ordering of refinement on $\mathcal{P}_n$ defined as follows: for partitions λ and μ of n, λ is a refinement of μ if and only if there exist corresponding partitions λ' and μ' of $[n]$ such that λ' is a refinement of μ'. Less formally, the parts of λ can be partitioned to form the parts of μ. The notions of Gibbs distributions and refining sequences transfer between $\mathcal{P}_n$ and $\mathcal{P}_{[n]}$ in such a way that the following results can be formulated with either state space. The many-to-one correspondence between partitions of the set $[n]$ and partitions of the integer n lifts easily to define a many-to-one correspondence between $\mathcal{P}_n$-valued and $\mathcal{P}_{[n]}$-valued processes, in such a way that the partial ordering of refinement is preserved. Thus for a $\mathcal{P}_{[n]}$-valued Gibbs fragmentation process there is a corresponding $\mathcal{P}_n$-valued fragmentation process, and vice-versa.

To see that Problem 8 is of some interest, note that Example 3 (cutting a random rooted segment) provides a $\mathcal{P}_{[n]}$-valued Gibbs fragmentation process for each n corresponding to the sequence of weights $w_j = j!$. Example 4 (cutting a random rooted tree) does the same thing for the weights $w_j = j^{j-1}$. What about for the sequence $w_j = 1$ of Example 1 (uniform random partitions) or the sequence $w_j = (j-1)!$ of Example 2 (uniform random permutations)? In these examples it does not seem at all obvious how to construct a Gibbs fragmentation process for $n \geq 4$. (Note that for $n \leq 3$ there exists a $\mathcal{P}_{[n]}$-valued Gibbs fragmentation process for arbitrary positive weights w_1 and w_2, for trivial reasons.) The question for $w_j = 1$ is largely settled by the following proposition. See section 5 regarding $w_j = (j-1)!$.

Proposition 9. *There is an* $n_0 < \infty$ *such that for all* $n \geq n_0$ *there does not exist a* $\mathcal{P}_{[n]}$*-valued Gibbs fragmentation process* $(\Pi_k, k \in [n])$ *with equal weights* $w_1 = w_2 = \cdots = w_{n-1}$.

Proof. Let $\Pi_{[n,k]}$ denote a random partition with the Gibbs distribution on $\mathcal{P}_{[n,k]}$ with equal weights $w_1 = w_2 = \cdots = w_{n-1}$, meaning that $\Pi_{[n,k]}$ has uniform distribution on $\mathcal{P}_{[n,k]}$. Let

$$X_{(n,k,1)} \geq X_{(n,k,2)} \geq \cdots \geq X_{(n,k,k)} \tag{15}$$

denote the sizes of components of $\Pi_{[n,k]}$ arranged in decreasing order. Then for each fixed i and k with $1 \leq i \leq k$ the ith largest component of $\Pi_{[n,k]}$ has relative size $X_{(n,k,i)}/n$ which converges in probability to $1/k$ as $n \to \infty$. This follows easily from the law of large numbers, and the elementary fact that $\Pi_{[n,k]}$ has the same distribution as $\Pi^*_{n,k}$ given that $\Pi^*_{n,k}$ has k components, where $\Pi^*_{n,k}$ is the random partition of $[n]$ generated by n independent random variables $U_1, \ldots, U_n$ each with uniform distribution on $[k]$.

(So i and j are in the same component of $\Pi^*_{n,k}$ if and only if $U_i = U_j$.) In particular, there is an $n_0 < \infty$ such that for all $n \geq n_0$ both

$$P(\text{ size of smallest component of } \Pi_{[n,2]} > (5/12)n) > 1/2 \tag{16}$$

and also

$$P(\text{ size of largest component of } \Pi_{[n,3]} > (5/12)n) < 1/2 \tag{17}$$

But if $(\Pi_{[n,k]}, 1 \leq k \leq n)$ were a fragmentation process, then $\Pi_{[n,3]}$ would be derived from $\Pi_{[n,2]}$ by by splitting one of the two components of $\Pi_{[n,2]}$. Since the component of $\Pi_{[n,2]}$ that is not split is among the three components of $\Pi_{[n,3]}$, if the smallest component of $\Pi_{[n,2]}$ contains more than $(5/12)n$ elements, then so does the largest component of $\Pi_{[n,3]}$. Thus for a fragmentation process, 16 implies the reverse of the inequality 17, and this contradiction yields the result. □

The above argument proves the non-existence for large n of a Gibbs fragmentation process for any weight sequence (w_j) such that for $k = 2$ or $k = 3$ the components in the Gibbs partition of $[n]$ into k components are approximately equal in size with high probability. For the weight sequence $w_j = j!$ of Example 3, what happens instead is that the sequence of ranked sizes 15, normalized by n, has a non-degenerate limit distribution for each k. As observed in [18, §5], this limit distribution on $[0,1]^k$ is the distribution of the ranked lengths of k subintervals of $[0,1]$ obtained by cutting $[0,1]$ at $k-1$ points picked independently and uniformly at random from $[0,1]$. This asymptotic distribution has been extensively studied [15]. For the weight sequence $w_j = j^{j-1}$ of Example 4, the behavior is different again. What happens is that for each fixed k the sequence of ranked sizes 15, when normalized by n, converges in probability to $(1, 0, \dots 0)$. That is to say, for any fixed k, for sufficiently large n, after k steps in the fragmentation process, there is with high probability one big component of relative mass nearly 1, and $k-1$ small components with combined relative mass nearly zero. To be more precise, it is easily shown that the $k-1$ small components, when kept in the order they are broken off the big component, have unnormalized sizes $X_{n,1}, \dots, X_{n,k-1}$ that are approximately independent for large n with asymptotic distribution

$$\lim_{n\to\infty} P(X_{n,i} = j) = \frac{j^{j-1}}{j!} e^{-j} \quad (i, j = 1, 2, \dots) \tag{18}$$

which is the *Borel distribution* of the total progeny of a critical Poisson-Galton-Watson process with Poisson(1) offspring distribution started with one individual. See e.g. [23, §4.1] and [6] for proofs and various generalizations. As a consequence of (18) and the asymptotic independence of the $X_{n,1}, \dots, X_{n,k-1}$, the asymptotic distribution of the combined size $X_{n,1} + \dots + X_{n,k-1}$ of all but the largest component of the partition of $[n]$ into k components is the distribution of the total progeny of the Poisson-Galton-Watson process with Poisson(1) offspring distribution started with k individuals, which is the

Borel-Tanner distribution [6]

$$\lim_{n\to\infty} P(X_{n,1}+\ldots+X_{n,k-1}=m)=\frac{k-1}{m}\frac{m^{m-k+1}e^{-m}}{(m-k+1)!} \tag{19}$$

4 Existence of Gibbs fragmentation processes

What follows is the derivation of a particular family of sequences (w_j), subject to certain additional assumptions, such that there does exist such a Gibbs fragmentation process for every n. This family of Gibbs fragmentation processes provides a generalization of the two combinatorial examples provided by Example 3 (cutting a random rooted segment) and Example 4 (cutting a random rooted tree). The existence of this family is implied by the work of Hendriks et al. [14], who studied a corresponding family of Markov processes governed by the Marcus-Lushnikov coalescent model [1, 21, 20] with collision kernel $K_{ij}=a+b(i+j)$, call it the *affine coalescent*. The discrete-time skeleton of the affine coalescent turns out to just the time reversal of the process derived here. The segment splitting model corresponds to $b=0$ and the tree splitting example to $a=0$.

However, it not obvious without careful calculation why the assumptions imposed here on a fragmentation process end up characterizing the time reversal of a affine coalescent.

First a superficial remark about Problem 8. If there exists any $\mathcal{P}_{[n]}$-valued Gibbs fragmentation process governed by $(w_1,\ldots,w_{n-1})$, then there exists one that is a Markov chain. For given a non-Markovian process, one can always create a Markov chain with the same one-step transition probabilities and the same marginal distributions. So Problem 8 reduces to:

Problem 10. *For which weight sequences $(w_1,\ldots,w_{n-1})$ does there exist a transition matrix $P(\pi,\nu)$ indexed by $\mathcal{P}_{[n]}$ such that $P(\pi,\nu)>0$ only if ν is a refinement of π, and*

$$\sum_{\nu\in\mathcal{P}_{[n]}} p(\pi|k-1)P(\pi,\nu)=p(\nu\mid k)\quad (1\le k\le n-1) \tag{20}$$

where $p(\nu\mid k)$ is given by the microcanonical Gibbs formula (5) if λ is a partition of $[n]$ into k components of sizes $n_1,\ldots,n_k$, and $p(\lambda\mid k)=0$ otherwise.

Less formally, such a transition matrix $P(\pi,\nu)$ corresponds to a *splitting rule*. Such a rule describes for each $1\le k\le n-1$ and each partition π of $[n]$ into $k-1$ components, the probability that π splits into a partition ν of $[n]$ into k components. Given that $\Pi_{k-1}=\pi_{k-1}$ say with $\pi_{k-1}=\{A'_1,\ldots,A'_{k-1}\}$ say, the only possible values π_k of Π_k are those $\pi_k=\{A_1,\ldots,A_k\}$ such that two of the A_j, say A_1 and A_2, form a partition of one of the A'_i, and the remaining A_j are identical to the remaining A'_i. The initial splitting rule starting with $\pi_1=\{1,\ldots,n\}$ is described by the assumption that the distribution Π_2 is governed by the Gibbs formula $p(n_1,n_2\mid 2)$ determined by the weight sequence

$(w_1, \ldots, w_{n-1})$ for n_1 and n_2 with $n_1 + n_2 = n$. The simplest way to continue is to use the following

Recursive Gibbs Rule: whenever a component is split, given that the component currently has size m, it is split according to the Gibbs formula $p(n_1, n_2 \,|\, 2)$ for n_1 and n_2 with $n_1 + n_2 = n$.

To complete the description of a splitting rule, it is also necessary to specify for each partition $\pi_{k-1} = \{A'_1, \ldots, A'_{k-1}\}$ the probability that the next component to be split is A'_i, for each $1 \leq i \leq k-1$. Here the simplest possible assumption seems to be the following:

Linear Selection Rule: Given $\pi_{k-1} = \{A'_1, \ldots, A'_{k-1}\}$, split A'_i with probability proportional to $\#A_i - 1$.

While this selection rule is somewhat arbitrary, it is natural to investigate its implications for the following reasons. Firstly, components of size 1 cannot be split, so the probability of picking a component to split must depend on size. This probability must be 0 for a component of size 1, and 1 for a component of size $n-k+2$. The simplest way to achieve this is by linear interpolation. Secondly, both the segment splitting model and the tree splitting model described in Examples 3 and 4 follow this rule. In each of these examples a component of size m is derived from a graph component with $m-1$ edges, so the linear selection rule corresponds to picking an edge uniformly at random from the set of all edges in the random graph whose components define Π_{k-1}. Given two natural combinatorial examples with the same selection rule, it is natural ask what other models there might be following the same rule. At the level of random partitions of $[n]$, this question is answered by the following proposition. It also seems reasonable to expect that the conclusions of the proposition will remain valid under weaker assumptions on the selection rule. See the end of Section 4 for further discussion.

4.1 Statement of the main result

Recall the definition of the discrete Marcus-Lushnikov coalescent process on $\mathcal{P}_{[n]}$ with affine kernel: this is the unique Markov chain on $\mathcal{P}_{[n]}$ such that π_1 is the partition consisting of singletons and π_k is obtained from π_{k-1} by merging a block of size i and j with probability proportional to $K_{i,j} = a + b(i+j)$ for some constants a and b. In the case $a = 1$ and $b = 0$ this is the finite version of Kingman's coalescent of Example 3 (blocks coalesce at rate 1) while if $a = 0$ and $b = 1$ this is the so-called additive coalescent already mentioned above, which is deeply related to Example 4.

Fix $n \geq 4$, let $(w_j, 1 \leq j \leq n-1)$ be a sequence of positive weights with with $w_1 = 1$, and let $(\Pi_k, 1 \leq k \leq n)$ be a $\mathcal{P}_{[n]}$-valued fragmentation process defined by the recursive Gibbs splitting rule derived from these weights, with the linear selection rule.

Theorem 11. *The three following statements are equivalent:*

(i) For each $1 \leq k \leq n$ the random partition Π_k has the Gibbs distribution on $\mathcal{P}_{[n,k]}$ derived from $(w_j, 1 \leq j \leq n-1)$;

(ii) There are two real numbers b and c such that the time-reversal of $(\Pi_k, 1 \leq k \leq n)$ is the discrete Marcus-Lushnikov coalescent with affine kernel $K_{i,j} = 2c + b(i+j)$.

(iii) The weight sequence w_j is of the form

$$w_j = \prod_{m=2}^{j} (mc + jb) \ , \ (j = 1, 2, \ldots, n-1) \tag{21}$$

The constants c and b appearing in (ii) and (iii) are only unique up to constant factors. To be more precise, if either of conditions (ii) or (iii) holds for some (b, c), then so does the other condition with the same (b, c). The process (ii) appears in a related work of Hendriks et al. [14], who showed that the distribution at time t of the continuous time $\mathcal{P}_n$-valued Markovian coalescent process, with collision kernel $K(x, y) := 2c + b(x + y)$, is given by a mixture over k, with mixing weights depending on t, of the Gibbs distribution (9) on partitions of n into k components induced by the weight sequence (21).

Note the implication of Theorem 11 that if Π' is a coalescent process constructed as in (ii), then the reversed process $(\Pi_1, \Pi_2, \ldots, \Pi_n)$ is a Gibbs fragmentation process governed by the recursive Gibbs splitting rule with weights (w_j) as in (iii) and linear selection probabilities.

Lying behind Theorem 11 is the following evaluation of the associated Bell polynomial: for $w_j = \prod_{m=2}^{j}(mc + jb), j = 2, 3, \ldots$

$$B_{n,k}(1, w_2, w_3, \ldots) = \binom{n-1}{k-1} \prod_{m=k+1}^{n} (mc + nb) \qquad (1 \leq k \leq n) \tag{22}$$

This evaluation, which can be read from [14, (19)-(21)], is shown to be a consequence of a famous convolution identity due to Gould [13].

4.2 Proof of Theorem 11, first half

Let $(\Pi_k, 1 \leq k \leq n)$ be a $\mathcal{P}_{[n]}$-valued fragmentation process defined by the recursive Gibbs splitting rule derived from these weights, with the linear selection rule, as assumed in the statement of the proposition.

Observe first that given any weight sequence $(w_1, \ldots, w_{n-1})$ the above two rules determine the transition probabilities of a $\mathcal{P}_{[n]}$-valued Markov process, hence a distribution of of a refining sequence of random partitions of $[n]$, say $\Pi_1, \ldots, \Pi_n$ such that Π_k has k components. What is not obvious without calculation, and turns out to depend on the weight sequence (w_j), is whether the distribution of Π_k is governed by the Gibbs

EPPF (9) for every $1 \le k \le n$. This is trivially true for $k = 1$ and $k = n$, and true by construction for $k = 2$. So the only interesting case to consider is $3 \le k \le n-1$. Assume this now, and make also the *inductive assumption* that the distribution of Π_{k-1} is governed by the Gibbs EPPF with weights $w_1, w_2, \ldots$.

Let π_k now denote any particular partition of $[n]$ into k components, say $\{A_1, \ldots, A_k\}$ with $\# A_i = n_i, 1 \le i \le k$, and for $1 \le i < j \le n$ let $\pi_{k-1}^{i,j}$ be the partition of $[n]$ into $k-1$ components derived from $\{A_1, \ldots, A_k\}$ by merging of A_i and A_j. Then the above splitting rule implies that

$$P(\Pi_{k-1} = \pi_{k-1}^{i,j}, \Pi_k = \pi_k) = P(\Pi_{k-1} = \pi_{k-1}^{i,j}) \frac{(n_i + n_j - 1)}{(n-k+1)} \frac{w_{n_i} w_{n_j}}{B_{n_i+n_j,2}} \tag{23}$$

From 9, the inductive assumption that Π_{k-1} has the Gibbs distribution with weights (w_j) can be conveniently written as

$$P(\Pi_{k-1} = \pi_{k-1}^{i,j}) = \frac{w_{n_i+n_j} \prod_{i=1}^{k} w_{n_i}}{B_{n,k-1}\, w_{n_i} w_{n_j}} \tag{24}$$

After substituting (24) into (23), and using (9), it is clear that

$$P(\Pi_{k-1} = \pi_{k-1}^{i,j}, \Pi_k = \pi_k) = p(n_1, \ldots, n_k \mid k) \frac{B_{nk}}{B_{n,k-1}} \frac{f(n_i + n_j)}{(n-k+1)} \tag{25}$$

where

$$f(m) := \frac{(m-1) w_m}{B_{m,2}} \qquad (2 \le m \le n-1) \tag{26}$$

Summing the probability (25) over all possible choices of (i,j) with $1 \le i < j \le k$ yields $P(\Pi_k = \pi_k)$. The random partition Π_k is governed by the Gibbs EPPF (9) if and only if this sum equals $p(n_1, \ldots, n_k \mid k)$. Clearly, this is the case if and only if the sequence $(f(m), 2 \le m \le n-1)$ defined by (26) is such that

$$\sum_{1 \le i < j \le k} f(n_i + n_j) = g(n,k) \text{ whenever } \sum_{i=1}^{k} n_i = n \tag{27}$$

for some function $g(n,k)$, and then

$$B_{n,k} = \frac{(n-k+1)}{g(n,k)} B_{n,k-1} \qquad (3 \le k \le n-1) \tag{28}$$

Here n is regarded as fixed, and for each k with $3 \le k \le n-1$ condition (27) must hold for every sequence of k positive integers $(n_1, \ldots, n_k)$ with sum n. According to the following lemma, condition (27) for $k = 3$ forces

$$f(m) = \frac{m-1}{B_{m,2}} = 2c + mb \quad (2 \le m \le n-1) \tag{29}$$

for some constants c and b. Then easily $g(n,k) = (k-1)(kc+nb)$ and (28) becomes

$$B_{n,k} = \frac{(n-k+1)}{(k-1)(kc+nb)} B_{n,k-1} \tag{30}$$

Clearly too, it is necessary that $kc + nb > 0$ for all $3 \le k \le n-1$.

Lemma 12. *Fix $n \ge 3$ and let $(f(m), 2 \le m \le n-1)$ be a sequence such that for every triple of positive integers (n_1, n_2, n_3) with $n_1 + n_2 + n_3 = n$*

$$f(n_1 + n_2) + f(n_2 + n_3) + f(n_1 + n_3) = C \tag{31}$$

for some constant C. Then there exist constants a and b such that $f(m) = am + b$ for every $2 \le m \le n-1$, and $C = 2an + 3b$.

Proof. For $n = 3$ or $n = 4$ the conclusion is trivial, so assume $n \ge 5$. Since $f(m)$ is defined only for $2 \le m \le n-1$, it is enough to show that

$$f(l) - f(l-1) = f(l-1) - f(l-2) \text{ for all } 4 \le l \le n-1 \tag{32}$$

Let i be the integer part of $l/2$ and $j = l - i$. Then $i \ge 2$ and either $j = i$ or $j = i+1$, so $j \ge 2$ too. Write $EQ(n_1, n_2, n_3)$ for the equation (31) determined by a particular choice of (n_1, n_2, n_3). Keeping in mind that $l = i + j$, we have

$$EQ(i-1, j-1, n-l+2): \quad f(l-2) + f(n-i+1) + f(n-j+1) = C \tag{33}$$

$$EQ(i-1, j, n-l+1): \quad f(l-1) + f(n-i+1) + f(n-j) = C \tag{34}$$

$$EQ(i, j-1, n-l+1): \quad f(l-1) + f(n-i) + f(n-j+1) = C \tag{35}$$

$$EQ(i, j, n-l): \quad f(l) + f(n-i) + f(n-j) = C \tag{36}$$

Subtract (33) from (34) to obtain

$$f(l-1) - f(l-2) = f(n-j+1) - f(n-j) \tag{37}$$

and subtract 35 from 36 to obtain

$$f(l) - f(l-1) = f(n-j+1) - f(n-j) \tag{38}$$

and Lemma 12 follows. □

At this point it is easy to see that $(i) \Rightarrow (ii)$. Indeed, assuming (i), we may apply the previous calculations to see that (25) holds with f again an affine function such as in

(29). By assumption $P(\Pi_k = \pi_k) = p(n_1, \ldots, n_k|k)$ so after division by $p(n_1, \ldots, n_k|k)$, (25) turns into

$$P(\Pi_{k-1} = \pi_{k-1}^{i,j} | \Pi_k = \pi_k) = \frac{B_{n,k}}{B_{n,k-1}(n-k+1)} f(n_i + n_j) \propto 2c + b(n_i + n_j)$$

which exactly says that the time-reversal of $(\Pi_k, 1 \le k \le n)$ has the transitions of the discrete Marcus-Lushnikov process with affine rate.

4.3 Proof of Theorem 11, second half

To review the argument so far, it has been shown that for $n \ge 4$, given a sequence of weights $(w_1, \ldots, w_{n-1})$, if the $\mathcal{P}_{[n]}$-valued fragmentation process $(\Pi_k, 1 \le k \le n)$ governed by the recursive Gibbs splitting rule with these weights, and selection of components for splitting with probability proportional to (size - 1), is such that Π_k has the Gibbs $(n, k, w_1, \ldots, w_{n-1})$ distribution for every $1 \le k \le n$, then the sequence of weights is such that

$$B_{m,2}(w_1, \ldots, w_{n-1}) = \frac{(m-1)}{(2c + mb)} w_m \qquad (3 \le m \le n-1) \tag{39}$$

for some constants c and b, and moreover the $B_{n,j} = B_{n,j}(w_1, \ldots, w_{n-1})$ are such that

$$B_{n,j-1} = B_{n,j} \frac{(j-1)(jc + nb)}{(n-j+1)} \qquad (3 \le j \le n-1)$$

Since the Gibbs $(n, k, w_1, \ldots, w_{n-1})$ distribution is unaffected by a common scaling of all the w_i, there is no loss of generality in supposing that $w_1 = 1$. Then $B_{n,n} = 1$ and the recursion (28) yields the formula

$$B_{n,k} = \binom{n-1}{k-1} \prod_{m=k+1}^{n} (mc + nb) \qquad (2 \le k \le n) \tag{40}$$

To complete the discussion, it remains to identify the weight sequence $(1, w_2, \ldots, w_{n-1})$ induced by b and c via the preceding formulae. Formula (39) gives easily

$$w_2 = \binom{n}{2}^{-1} B_{n,n-1} = \binom{n}{2}^{-1} \binom{n-1}{n-2} (nc + nb) = 2c + 2b \tag{41}$$

and for small j one can use (39) to check the formula

$$w_j = \prod_{m=2}^{j} (mc + jb) \qquad (j = 1, 2 \ldots)$$

However, it is not at all obvious from the recursion (30) why the w_j should admit such a factorization for all j, and it seems hopeless to try to prove (21) by induction on j using this recursion. Formally, the expression (21) for $j = n$ is obtained by simply taking $k = 1$ in (40) and using the fact that $B_{n,1}(w_1, w_2, \ldots) = w_n$. However, while the conclusion turns out to be correct, this argument is entirely spurious for the following reason. There is no possible justification of the substitution $k = 1$ in (40), as this formula was derived entirely from consideration of a $\mathcal{P}_{[n]}$-valued fragmentation process $(\Pi_k, 1 \leq k \leq n)$ whose entire distribution is determined by the sequence $(w_1, \ldots, w_{n-1})$ independently of the value of the value of $B_{nn}(1, w_2, \ldots) = w_n$. Another similar idea is to take $k = 2$ instead in (40), and deduce a formula for $B_{n,2}$. Since $f(m) = (m-1)w_m/B_{m,2}$ taking $m = n$ we derive an expression for w_n. But of course, the objection remains, and this has to do with the fact that the expression $f(m) = (m-1)w_m/B_{m,2}$ is only valid for $m \leq n-1$.

We now provide a rigorous proof that equation (21)

$$w_j = \prod_{m=2}^{j} (mc + jb) \qquad (j = 1, 2 \ldots, n-1)$$

holds.

Lemma 13. *Suppose $w_j = \prod_{m=2}^{j}(mc + jb)$ is defined by the above equation, where b and c are such that $mc + jb > 0$ for $1 \leq m$. Then for $n \geq 1$*

$$B_{n,k}(w_1, w_2, \ldots) = \binom{n-1}{k-1} \prod_{m=k+1}^{n} (mc + nb) \qquad (1 \leq k \leq n) \tag{42}$$

Proof. The idea is to use the calculations already done so far and show that a recursion similar to (28) and (30) holds. In fact, the idea is to define a fragmentation process $(\Pi_k, 1 \leq k \leq n)$ with values in $\mathcal{P}_{[n]}$ such that Π_k has k blocks for all $1 \leq k \leq n$, which evolves according to the linear selection rule and then splits a block of size m according to the Gibbs $(m, 2, w_1, w_2, \ldots)$ distribution. We first claim that $(\Pi_k, 1 \leq k \leq n)$ is a Gibbs fragmentation, i.e. for each k Π_k has the Gibbs $(n, k, w_1, \ldots)$ distribution. Indeed we can proceed by recursion and use earlier calculations. As before, and using the same notations, we have:

$$P(\Pi_{k-1} = \pi_{k-1}^{i,j}, \Pi_k = \pi_k) = p(n_1, \ldots, n_k \mid k) \frac{B_{nk}}{B_{n,k-1}} \frac{f(n_i + n_j)}{(n-k+1)} \tag{43}$$

with $f(m) = (m-1)w_m/B_{m,2}$ for $1 \leq m \leq m-1$. But we claim that it is still true that

$$f(m) = 2c + mb \tag{44}$$

To see this, note that

$$B_{m,2} = \frac{1}{2} \sum_{k=1}^{m-1} \binom{m}{k} w_k w_{m-k}$$

So it is enough to prove that

$$\frac{2c+mb}{2(m-1)}\sum_{k=1}^{m-1}\frac{w_k}{k!}\frac{w_{m-k}}{(m-k)!}=\frac{w_m}{m!}$$

Equivalently, if $v_k = w_k/k!$,

$$\frac{2c+mb}{2(m-1)}\sum_{k=1}^{m-1} v_k v_{m-k} = v_m \tag{45}$$

We derive (45) from a well-known convolution identity due to Gould [13], generalizing the Chu-Vandermonde identity. We start by a remark that it is enough to prove (45) when $c = -1$. (If $c \neq -1$ a simple multiplication by c^{m-1} on both sides and a change of variable $b = bc$ yields the desired conclusion).

Let $A_k(a,b) = \frac{a}{a-bk}\binom{a-bk}{k}$ for all real numbers a and b, i.e.

$$A_k(a,b) = \frac{a(a-bk-1)\dots(a-bk-k+1)}{k!}$$

when $k \geq 1$ and $A_0(a,b) = 1$. A main result of [13] is the identity valid for all a, b and any integer m

$$\sum_{k=0}^{m} A_k(a,b)A_{m-k}(c,b) = A_m(a+c,b) \tag{46}$$

In (46) we plug in $a = c = -1$. Note that then $A_k(-1,-b) = v_k$ for $k \geq 1$ (with the c defining the v_k also equal to -1) and $A_m(-2,-b) = -2\frac{bm-m-1}{bm-2}v_m$. Isolating the terms with $k = 0$ and $k = m$ in the sum we obtain

$$-2v_m + \sum_{k=1}^{m-1} v_k v_{m-k} = \frac{-2(bm-m-1)}{bm-2}v_m$$

which transforms into

$$\frac{bm-2}{2}\sum_{k=1}^{m-1} v_k v_{m-k} = (m-1)v_m$$

which is exactly what (45) states in the reduced case $c = -1$. Hence f is an affine function (44).

From (44) it follows by direct calculation

$$\sum_{i<j} f(n_i+n_j) = g(n,k) = (k-1)(kc+nb)$$

whenever $\sum_i n_i = n$. Therefore, by summing over all values

$$B_{n,k} = \frac{(n-k+1)}{g(n,k)}B_{n,k-1} \qquad (3 \leq k \leq n-1)$$

So it follows that

$$B_{n,k} = \frac{(n-k+1)}{(k-1)(kc+nb)} B_{n,k-1}$$

On the other hand since $w_1 = 1$ we have that $B_{n,n}(w_1, w_2 \ldots) = 1$ so the same recursion holds for the $B_{n,k}$:

$$B_{n,k} = \binom{n-1}{k-1} \prod_{m=k+1}^{n} (mc+nb) \qquad (2 \le k \le n)$$

which is what the lemma claimed. □

In particular, this shows that the fragmentation process $(\Pi_k, 1 \le k \le n)$ with Gibbs splitting rule and linear selection is a Gibbs fragmentation.

To complete the proof of Theorem 11, we give a short proof to the well-known fact that Bell polynomials uniquely determine the underlying weight sequence w_j. This will end the proof of Theorem 11 since it has now been shown that any Gibbs fragmentation has the same Bell polynomial as when the weight sequence has the form (21), and that furthermore the time-reversal is the discrete Marcus-Lushnikov affine coalescent.

Lemma 14. *Suppose $w_1 = 1$. Then the sequence of numbers $(B_{n,k}(w_1, w_2, \ldots), 1 \le k \le n)$ uniquely determine the numbers $(w_1 = 1, w_2, \ldots, w_{n-1}, w_n)$.*

Proof. First remark that $B_{n,n-1} = \binom{n}{2} w_2$. Indeed to evaluate $B_{n,n-1}$ we must see what are the different partitions of $\mathcal{P}_{[n]}$ with $(n-1)$ blocks. Of course the only way to do this is to have $(n-2)$ singletons and one block i, j of size 2. The singletons contribute $w_1 = 1$ to the product, and to each choice of $i \neq j$ i, j corresponds to a mass w_2. So, $B_{n,n-1} = \binom{n}{2} w_2$. In particular, w_2 is determined by $B_{n,n-1}$.

When we evaluate $B_{n,n-2}$ there are two types of partitions with $n-2$ blocks: those with $n-3$ singletons and one block of size 3, and those with 2 blocks of size 2, the rest being singletons. Therefore if $N_{(2,2)}$ is the number of ways to choose two blocks of size 2 among n elements,

$$B_{n,n-2} = \binom{n}{3} w_3 + N_{(2,2)} w_2^2$$

so w_3 is also uniquely determined by the $B_{n,k}$. We can proceed further by recursion:

$$B_{n,n-k} = \binom{n}{k+1} w_{k+1} + \sum_{\mathbf{n}=(n_1,\ldots,n_r)} N_{(n_1,\ldots,n_r)} \prod_{q=1}^{r} w_{n_q} \tag{47}$$

where $N_{(n_1,\ldots,n_r)}$ is the number of partitions of $[n]$ with 1 block of size n_1,1 block of size $n_2, \ldots$, and one block of size n_r (we allow repetitions of n_i and do not write the singleton indices). In this equation the sum is on all multi-indices $\mathbf{n}$ such that $n_i \le k$ for all i and the resulting partition has exactly $n-k$ blocks. Using the induction assumption, the w_i for $i \le k$ are all determined by the $B_{n,j}$, so (47) shows that w_{k+1} is also uniquely determined by the $(B_{n,j}, 1 \le j \le n)$. □

This also concludes the proof of Theorem 11. □

While the above argument provides a recursive formula for $(w_i, 1 \le i \le j)$ in terms of the $B_{n,k}(w_1, w_2, \ldots)$ for $n - j + 1 \le k \le n$, this formula is quite unwieldy. For example, given a sequence of positive numbers $b_{n,k}$ for $1 \le k \le n$, there exists a unique sequence of weights $(w_i, 1 \le i \le n)$ with $w_i > 0$ such that $B_{n,k}(w_1, \ldots w_n) = b_{n,k}$ for every $1 \le k \le n$. But given a formula for the $b_{n,k}$ for $1 \le k \le n$ the general expression for the w_i is so complicated that it may be difficult to determine whether or not the w_i are positive.

Finally, we comment on the linear selection rule and possible extensions of our method to other selection rules. Crucial to our arguments was the fact that given $\pi_{k-1} = \{A'_1, \ldots, A'_{k-1}\}$, the probability of splitting A'_i should equal $p_{nk}(\#A'_i)$ for some non-negative function $p_{nk}(m)$ defined for $1 \le m \le n$. Such a function p_{nk} must obviously satisfy the following conditions:

$$p_{nk}(m) = 0 \text{ if } m = 1 \text{ or } m \ge n - k + 1 \tag{48}$$

$$\sum_{i=1}^{k-1} p_{nk}(m_i) = 1 \text{ if } \sum_{i=1}^{k-1} m_i = n \tag{49}$$

where the condition (49) must hold for all sequences of positive integers $(m_1, \ldots, m_{k-1})$ with sum n. Also, if π_{k-1} happens to be composed of $k-2$ singleton components and one component of size $n-k+2$, obviously the component of size $n-k+2$ must be the next to split, so that

$$p_{nk}(n - k + 2) = 1 \tag{50}$$

Clearly, the simplest such p_{nk}, corresponding to the above selection rule is

$$p_{nk}(m) = \frac{m-1}{n-k+1} \quad (1 \le m \le n - k + 2) \tag{51}$$

and this is the only function which is affine over the range $1 \le m \le n-k+2$ and satisfies the a priori constraints (48) and (49). For $k = 3$ the constraint (49) just amounts to

$$p_{n3}(m) = 1 - p_{n3}(n - m) \tag{52}$$

which does not force p_{n3} to be affine. However, it appears that for $k \ge 4$ it may be possible to deduce (51) from the assumptions (48) - (49) by an argument similar to the proof of Lemma 12 above.

More complex splitting rules can certainly be imagined, but it is not obvious how to proceed with their analysis. Even the splitting rule proposed above requires some careful discussion.

5 Gibbs fragmentations for random permutations in continuous time

Given a symmetric non-negative *collision rate function* $K_{i,j}$ defined for positive integers i and j, call the $\mathcal{P}_{[n]}$-valued continuous time parameter Markovian coalescent process $(\Pi_t, t \geq 0)$, in which each pair of clusters of sizes i and j is merging at rate $K_{i,j}$, the *Marcus-Lushnikov coalescent with collision kernel* $K_{i,j}$. See [1] for background. It is assumed throughout this section, in keeping with the definition of a coalescent process given in the previous section, that such a coalescent process is started with the *monodisperse initial condition*. That is to say Π_0 is the partition of $[n]$ into n singletons. Both Marcus and Lushnikov worked with the corresponding $\mathcal{P}_n$-valued process rather than a $\mathcal{P}_{[n]}$-valued process, but there is no difficulty in translating results from one state-space to the other, by application of the standard criterion for a function of a Markov process to be Markov. Lushnikov [20] found the remarkable result that for a collision kernel of the form $K_{i,j} = if(j) + jf(i)$ for each $t > 0$ the distribution of Π_t is of the form

$$P(\Pi_t = \pi) = \sum_{k=1}^{n} p_{n,k}(t) p(\pi \mid n, k; w_j(t), j = 1, 2, \ldots) \tag{53}$$

where $p(\pi \mid n, k; w_j, j = 1, 2, \ldots)$ denotes the microcanonical Gibbs distribution on $\mathcal{P}_{[n,k]}$ with weights w_j, and the functions $p_{n,k}(t) = P(\#\Pi_t = k)$ and the weights $w_j(t)$ are determined by a system of differential equations. As mentioned earlier, Hendriks et al. [14] showed that for $K_{i,j} = a + b(i + j)$ for constants a and b the $w_j(t)$ can be chosen independently of t as $w_j(t) = w_j$ where w_j is determined by a and b via formula (21) for $c = a/2$. The $\mathcal{P}_{[n]}$-valued Marcus-Lushnikov coalescent $(\Pi_t, t \geq 0)$ with $K_{i,j} = a+b(i+j)$ is thus a Gibbs coalescent with these weights (w_j). Given a continuous time $\mathcal{P}_{[n]}$-valued coalescent or fragmentation process $(\Pi_t, t \in I)$, define the *discrete skeleton* of $(\Pi_t, t \in I)$ to be the $\mathcal{P}_{[n]}$-valued process $(\Pi^*_k, 1 \leq k \leq n)$ where Π^*_k is the common value of Π_t for all $t \in I$ such that $\#\Pi_t = k$. As observed by Hendriks et al. the Marcus-Lushnikov coalescent $(\Pi_t, t \geq 0)$ with $K_{i,j} = a + b(i + j)$ is such that the process $(\#\Pi_t, t \geq 0)$ is independent of the discrete skeleton of (Π_t); the fact that $(\Pi_t, t \geq 0)$ is a Gibbs coalescent with a particular sequence of weights w_j is therefore equivalent to the fact that the discrete skeleton is a Gibbs coalescent with the same weights. Note however that this equivalence relies on the independence of the process $(\#\Pi_t, t \geq 0)$ and the discrete skeleton. In general, it is not necessarily true that the discrete skeleton of a continuous time Gibbs coalescent is a discrete time Gibbs coalescent. This phenomenon is illustrated by an example later in this section.

Kingman [18] studied the particular case of the Marcus-Lushnikov coalescent with $a = 1$ and $b = 0$. In this process, at any given time t, given that $\#\Pi_t = k$, each of the $k(k-1)/2$ pairs clusters in existence at time t is merging at rate 1. Call this process with state space $\mathcal{P}_{[n]}$ *Kingman's n-coalescent*, Motivated by applications to genetics, Kingman [18] proposed the following construction. Given a coalescent process $(\Pi_t, t \geq 0)$, suppose that each cluster of Π_t is subject to mutation at rate $\theta/2$ for some $\theta > 0$. Now define a random partition $\tilde{\Pi}_\theta$ of $[n]$ by declaring that i and j are in the same block of $\tilde{\Pi}_\theta$ if and only if no mutation affects the clusters containing i and j in the interval $(0, \tau_{ij})$ where τ_{ij} is the *collision time* of i and j in the coalescent process $(\Pi_t, t \geq 0)$, that is the first time t that i and j are in the same cluster of Π_t. Kingman obtained the following remarkable result:

Proposition 15. (Kingman [18]) *Suppose that $(\Pi_t, t \geq 0)$ is Kingman's n-coalescent. Then*

$$P(\tilde{\Pi}_\theta = \pi) = \frac{\theta^{k-1}}{[\theta+1]_{n-1}} \prod_{i=1}^{k} (n_i - 1)! \tag{54}$$

for each partition π of $[n]$ into k components of sizes $n_1, \ldots, n_k$

The distribution of $\tilde{\Pi}_\theta$ defined by (54) first appears in [9] and is known as Ewens' Sampling Formula with parameter θ. Note that it is a canonical Gibbs distribution on $\mathcal{P}_{[n]}$ with weight sequence $((j-1)!, j \geq 1)$. This distribution is a particular mixture over k, with mixing coefficients depending on θ, of the microcanonical Gibbs distributions on $\mathcal{P}_{[n,k]}$ with weights $(j-1)!$, as interpreted in Example 2. Compare with the distribution of Π_t for each fixed t, which is a mixture over k, with mixing coefficients depending on t, of the microcanonical Gibbs distributions on $\mathcal{P}_{[n,k]}, 1 \leq k \leq n$, with the different weight sequence $(j!, j \geq 1)$ as interpreted in Example 3. It does not seem all obvious from a combinatorial perspective why there should be such a connection between the Gibbs models with these two weight sequences. Neither does there seem to be any straightforward extension of Proposition 15 to a Marcus-Lushnikov coalescent (Π_t) with e.g. $K_{ij} = a + b(i+j)$ for $b \neq 0$. The random partition $\tilde{\Pi}_\theta$, and more generally the fragmentation process $(\tilde{\Pi}_\theta, \theta \geq 0)$ discussed below, can be defined starting from any coalescent (Π_t), but there seems to be a manageable formula for the distribution of $\tilde{\Pi}_\theta$ only for Kingman's coalescent. As a development of Kingman's result, there is the following proposition:

Theorem 16. *There exists a Gibbs fragmentation process $(\tilde{\Pi}_\theta, \theta \geq 0)$ with weight sequence $((j-1)!, j \geq 1)$ such that for each $\theta > 0$ the distribution of $\tilde{\Pi}_\theta$ is the canonical Gibbs distribution on $\mathcal{P}_{[n]}$ with these weights and parameter θ, as displayed in (54).*

Proof. Given the path of a Kingman coalescent process $(\Pi_t, t \geq 0)$, construct a random tree $\mathcal{T}$ as follows. Let the vertices of the tree $\mathcal{T}$ be labelled by the random collection $\mathbf{V}$ of all subsets of $[n]$ which appear as clusters in the coalescent at some time in its evolution. Because the coalescent develops via binary mergers, starting with n singletons and terminating $\Pi_t = [n]$ for all sufficiently large t, the set $\mathbf{V}$ comprises the collection of all n singleton subsets of $[n]$, which are the *leaves* of the tree, the whole set $[n]$ which is the *root* of the tree, and $n-2$ further subsets of $[n]$, whose identities depend on how the coalescent evolves, which are the *internal vertices* of the tree. The tree $\mathcal{T}$ has $n + 1 + (n-2) = 2n - 1$ vertices all together. Associate with each subset v of $[n]$ that is a vertex of the tree the time $t(v)$ at which the coalescent forms the cluster v. Thus $t(v) = 0$ if and only of v is one of the n singleton leaf vertices, $t([n]) = \inf\{t : \#\Pi_t = 1\}$, and the collection of times $t(v)$ as v ranges over the $n-1$ non-leaf vertices of the tree is the set of times t at which the process $(\#\Pi_t, t \geq 0)$ experiences a downward jump. For each non-leaf vertex v in $\mathcal{T}$, let there be exactly two edges of $\mathcal{T}$ directed from v to v_1 and v_2, where v_1 and v_2 are the two clusters which merged to form v. Let each vertex v of $\mathcal{T}$ be placed at height $t(v)$ equal to the time of its formation, and for $i = 1, 2$ regard the directed edge from v to v_i as a segment of length $t(v) - t(v_i)$. Now Kingman's construction of $\tilde{\Pi}_\theta$ amounts to supposing that there is a Poisson process of cut points on the edges of this tree, with rate $\theta/2$ per unit length, and identifying the blocks of $\tilde{\Pi}_\theta$ with the restrictions to the set of n leaves of $\mathcal{T}$ (identified with $[n]$) of the components of the random forest obtained by cutting segments of $\mathcal{T}$ at the Poisson cut points. Now conditionally given the tree $\mathcal{T}$, construct the Poisson cut points simultaneously for all $\theta > 0$ so that for each edge of the tree of of length ℓ the moments of cuts of that edge form a homogeneous Poisson process of rate $\ell\theta/2$, and these processes are independent for different edges. Then $\tilde{\Pi}_\theta$ has been constructed simultaneously for each $\theta > 0$ in such a way that $\tilde{\Pi}_\theta$ is obviously a refinement of $\tilde{\Pi}_\phi$ for $\theta > \phi$. Since with probability one there are no ties between the times of cuts on different segments, it is clear that the process $(\tilde{\Pi}_\theta, \theta \geq 0)$ develops by binary splits. Thus $(\tilde{\Pi}_\theta, \theta \geq 0)$ is a Gibbs fragmentation process. □

While the one-dimensional distributions of this process $(\tilde{\Pi}_\theta, \theta \geq 0)$ are given by Ewens' sampling formula (54), the two and higher dimensional distributions seem difficult to describe explicitly. In particular, a calculation of the simplest transition rate associated with the process $(\tilde{\Pi}_\theta, \theta \geq 0)$, provided below, shows that this rate is depends on θ. It seems quite difficult to give a full account of all transition rates of $(\tilde{\Pi}_\theta, \theta \geq 0)$, though their general form can be described and a method for their computation for small n will be indicated. For $n \geq 2$ the process $(\tilde{\Pi}_\theta, \theta \geq 0)$ turns out to be non-Markovian, so its distribution is not determined by its transition rates.

In connection with Proposition 16 and such calculations, the following problem arises:

Problem 17. *Does there exist for each n a $\mathcal{P}_{[n]}$-valued Gibbs fragmentation process $(\Pi_k, 1 \leq k \leq n)$ with weight sequence $((j-1)!, j \geq 1)$?*

The following calculations show that for n large enough, the discrete skeleton of $(\tilde{\Pi}_\theta, \theta \geq 0)$ does not provide such a process.

5.1 Calculations with the tree derived from Kingman's coalescent.

Let $\mathcal{T}_n$ denote the random tree derived as in the proof of Theorem 16 from Kingman's n-coalescent $(\Pi_t, t \geq 0)$. For $1 \leq k \leq n$ let $T_k = \inf\{t : \#\Pi_t = k\}$. Note that almost surely

$$0 = T_n < T_{n-1} < \cdots < T_1 \tag{55}$$

The tree $\mathcal{T}_n$ has root vertex at level T_1 and $n-2$ internal vertices at levels $T_2 > \cdots > T_2$ and n leaves at level $T_n = 0$. Regard the tree $\mathcal{T}$ as the union of a collection of line segments, with appropriate identification of endpoints. Each line segment in the tree joins a vertex at some level T_i to a vertex at level T_j for some $i < j$. For $1 \le i \le n-1$ define the ith *stratum* of the tree to be the portion of the tree between levels T_i and T_{i+1}. So the ith stratum consists of $i+1$ disjoint line segments, each of length $T_i - T_{i+1}$. Each segment in the ith stratum is a subsegment of some segment of $\mathcal{T}$ which joins some vertex at level T_i or higher to some vertex at level T_{i+1} or lower. Each segment in the ith stratum corresponds to one of the $i+1$ clusters of Π_t in existence throughout the time interval (T_{i+1}, T_i) during which $\#\Pi_t = i+1$. The entire tree $\mathcal{T}$ is thus composed as a disjoint union of line segments L_{ij} say, as i with $1 \le i \le n-1$ labels the strata and j with $1 \le j \le i+1$ labels clusters within strata. Since the length of each segment L_{ij} is $T_{i+1} - T_i$, the total length of all segments in the tree $\mathcal{T}$ is

$$L_n := \sum_{i=1}^{n-1} (i+1)(T_{i+1} - T_i) \tag{56}$$

From the definition of the underlying coalescent process $(\Pi_t, t \ge 0)$, the random variable $T_{i+1} - T_i$ has exponential distribution with rate $i(i+1)/2$, and these random variables are independent for $1 \le i \le n-1$. If T_λ has exponential distribution with rate λ then for $\alpha > 0$ $E\exp(-\alpha T_\lambda) = \lambda/(\alpha+\lambda)$, so (56) gives

$$E\exp(-\frac{\theta}{2}L_n) = \prod_{i=1}^{n-1} \frac{i(i+1)/2}{\theta(i+1)/2 + i(i+1)/2} \tag{57}$$

which reduces to

$$E\exp(-\frac{\theta}{2}L_n) = \frac{(n-1)!}{[\theta+1]_{n-1}} \tag{58}$$

where $[\theta+1]_{n-1} = \prod_{i=1}^{n-1}(\theta+i)$. On the other hand, given L_n, the Poisson process with rate $\theta/2$ per unit segment length in the tree has no points with probability $\exp(-(\theta/2L_n)$. So the expectation calculated in (58) is just the probability that $\tilde{\Pi}_\theta$ is the partition of $[n]$ with one component, in agreement with Ewens'sampling fomrula (54) for $k=1$ and $n_1 = n$.

Consider now the problem of calculating the joint distribution of Θ and $\tilde{\Pi}_\Theta$ where Θ is the time of the first split in the fragmentation process $(\tilde{\Pi}_\theta, \theta \ge 0)$. Since the event that $(\Theta > \theta)$ is identical to the event that $\tilde{\Pi}_\theta$ has one component, the distribution of Θ is given by the same formula:

$$P(\Theta > \theta) = \frac{(n-1)!}{[\theta+1]_{n-1}} \tag{59}$$

which implies by differentiation that Θ has density

$$P(\Theta \in d\theta)/d\theta = \frac{(n-1)!}{[\theta+1]_{n-1}} \sum_{i=1}^{n-1} \frac{1}{i+\theta} \tag{60}$$

Let

$$S_i := (i+1)(T_{i+1} - T_i) \tag{61}$$

so that S_i is the total length of segments in the ith stratum. Let I denote the index of the stratum in which the first cut point falls at time Θ. Then it follows from the representation of L_n as the sum of independent exponential variables $L_n = \sum_{i=1}^{n-1} S_i$ that the sum over i in 60 corresponds to summing over the possible values i of I. That is, for $1 \le i \le n-1$,

$$P(\Theta \in \theta, I = i)/d\theta = \frac{(n-1)!}{[\theta+1]_{n-1}} \frac{1}{i+\theta} \tag{62}$$

and hence

$$P(I = i \mid \Theta = \theta) = \frac{(i+\theta)^{-1}}{\sum_{j=1}^{n-1} (j+\theta)^{-1}} \tag{63}$$

Observe now that given $\Theta = \theta$ and $I = i$, the partition $\tilde{\Pi}_\Theta$ consists of two components, obtained as the restriction to $[n]$, identified as the set of leaves of the tree $\mathcal{T}$, of the two components of $\mathcal{T}$ separated by the cut at time Θ in stratum i of $\mathcal{T}$. To be precise, $\tilde{\Pi}_\Theta = \{C, [n] - C\}$ where C is the cluster of Π_t in existence during the time interval (T_{i+1}, T_i) corresponding to the segment of $\mathcal{T}$ which is cut at time Θ. This C is one of the clusters of Π^*_{i+1}, where $(\Pi^*_k, 1 \le k \le n)$ is the discrete skeleton of $(\Pi_t, t \ge 0)$. Now by construction of the Poisson cutting process, and the fact that the discrete skeleton $(\Pi^*_k, 1 \le k \le n)$ of $(\Pi_t, t \ge 0)$ is independent of $(\#\Pi_t, t \ge 0)$, it is clear that the conditional distribution of Π^*_{i+1} given $\Theta = \theta$ and $I = i$ is identical to its unconditional distribution, that is the Gibbs distribution on $\mathcal{P}_{[n,i+1]}$ with weights $(j!, j \ge 1)$, and that C is one of the $i+1$ components of Π^*_{i+1} picked by a mechanism independent of the sizes of these components. Therefore,

$$P(\#C = n_1 \mid \Theta = \theta, I = i) = P(\#C_{i+1} = n_1) \tag{64}$$

where C_{i+1} is a random component of Π^*_{i+1}. But the Gibbs distribution of Π^*_{i+1}, or rather the combinatorial interpretation developed in Example 3 now yields the formula

$$P(\#C = n_1 \mid \Theta = \theta, I = i) = \binom{n-1-n_1}{i-1} \binom{n-1}{i}^{-1} \tag{65}$$

Let π now denote any particular partition of $[n]$ into two components of sizes n_1 and n_2 where $n_1 + n_2 = n$. It follows easily from (65) by exchangeability that the conditional distribution of $\tilde{\Pi}_\Theta$ given $\Theta = \theta, I = i)$ is given by

$$P(\tilde{\Pi}_\Theta = \pi \mid \Theta = \theta, I = i) = \binom{n}{n_1}^{-1} \binom{n-1}{i}^{-1} \left[\binom{n-1-n_1}{i-1} + \binom{n-1-n_2}{i-1} \right] \tag{66}$$

Combining this expression with (63) shows that the conditional distribution of $\tilde{\Pi}_\Theta$ given $\Theta = \theta$ is given by

$$P(\tilde{\Pi}_\Theta = \pi \mid \Theta = \theta) \frac{\sum_{i=1}^{n-1} (i+\theta)^{-1} \binom{n-1}{i}^{-1} \left[\binom{n-1-n_1}{i-1} + \binom{n-1-n_2}{i-1} \right]}{\binom{n}{n_1} \sum_{j=1}^{n-1} (j+\theta)^{-1}} \tag{67}$$

In particular, the time Θ of the first split of the process process $(\tilde{\Pi}_\theta, \theta \geq 0)$ and the state $\tilde{\Pi}_\Theta$ of this process at the time of the first split are *not* independent. So, in contrast to the situation for the the underlying coalescent $(\Pi_t, t \geq 0)$, the discrete skeleton of the fragmentation process $(\tilde{\Pi}_\theta, \theta \geq 0)$ is not independent of the process $(\#\tilde{\Pi}_\theta, \theta \geq 0)$. A formula for the unconditional distribution of $\tilde{\Pi}_\Theta$ is obtained from (67) and (60) by integration. Thus for π with two components of sizes n_1 and n_2

$$P(\tilde{\Pi}_\Theta = \pi) = \frac{(n-1)!}{\binom{n}{n_1}} \int_0^\infty \frac{d\theta}{[\theta+1]_{n-1}} \sum_{i=1}^{n-1} \frac{\left[\binom{n-1-n_1}{i-1} + \binom{n-1-n_2}{i-1} \right]}{\binom{n-1}{i}(i+\theta)} \tag{68}$$

Of course from this formula one might guess that there is no reason for this to be same as the probability that a Gibbs $(n, 2, (j-1)!)$ random partition is equal to π. To see this more rigorously, let J_Θ denote the size of a component of Π_Θ picked by the toss of a fair coin independent of Π_Θ. Then $P(J_\Theta = n_1)$ equals $\binom{n}{n_1}/2$ times the common value of $P(\tilde{\Pi}_\Theta = \pi)$ for all π with two components of sizes n_1 and $n - n_1$.

In particular when $n_1 = 1$, note that

$$\sum_{i=1}^{n-1} \frac{\left[\binom{n-1-n_1}{i-1} + \binom{n-1-n_2}{i-1} \right]}{\binom{n-1}{i}(i+\theta)} \quad \geq \quad \sum_{i=1}^{n} \frac{i}{(n-1)(i+\theta)} \geq 1/2$$

for n large enough if $\theta \leq 1$. We will restrict the domain of integration not only to $[0, 1]$ but actually to $[0, a/\log n]$. If $\theta \leq a/\log n$, remark that

$$\begin{aligned} [\theta+1]_{n-1} &= \exp(\sum_{i=1}^{n-1} \ln(\theta+i)) \\ &\leq (n-1)! \exp\left(\sum_{i=1}^{n-1} \ln\left(1 + \frac{a}{i \log n}\right)\right) \\ &\leq c(n-1)! \end{aligned}$$

for some $c > 0$. Therefore

$$\begin{aligned} P(J_\Theta = 1) &= \frac{1}{2}(n-1)! \int_0^\infty \frac{d\theta}{[\theta+1]_{n-1}} \sum_{i=1}^{n-1} \frac{\left[\binom{n-1-n_1}{i-1} + \binom{n-1-n_2}{i-1}\right]}{\binom{n-1}{i}(i+\theta)} \\ &\geq \frac{1}{2}(n-1)! \int_0^{a/\log n} \frac{d\theta}{c(n-1)!} \frac{1}{2} \\ &\leq c_1 (\log n)^{-1} \end{aligned}$$

for some $c_1 > 0$ if n is large enough.

On the other hand, we will show that if Π has the Gibbs $(n, 2, w)$ distribution with $w_j = (j-1)!$, and J is the size of a randomly picked component, then

$$P(J = 1) \asymp \frac{\ln n}{n}$$

To see this we start by remarking that $B_{n,2}(w) = (1/2) \sum_{j=1}^{n-1} \binom{n}{j} (j-1)!(n-1-j)!$ so

$$B_{n,2}(w) = \frac{n!}{2} \sum_{j=1}^{n-1} \frac{1}{j(n-j)} \asymp n! (\ln n)^{-1}$$

Since the number of permutations with exactly two cycles one of which has size 1 is $n(n-2)!$, we conclude that

$$P(J = 1) = (1/2) \frac{n(n-2)!}{B_{n,2}(w)} \asymp \frac{\ln(n)}{n}$$

as claimed. Since the two quantities have different asymptotics this shows that for n large enough,the distribution of the partition of $[n]$ into two parts obtained at the time of the first split is not the common distribution of $\tilde{\Pi}_\theta$ given $\#\tilde{\Pi}_\theta = 2$ for all $\theta > 0$. In fact, for $n = 4$ the two quantities are already not the same: it is possible to do the integration explicitly to obtain

$$P(J_\Theta = 1) = P(J_\Theta = 3) = \frac{1}{2} + \log 2 - \frac{3}{4} \log 3 \tag{69}$$

and

$$P(J_\Theta = 2) = -2 \log 2 + \frac{3}{2} \log 3 \tag{70}$$

while:

$$P(J = 1) = P(J = 3) = 4/11 \tag{71}$$

$$P(J = 2) = 3/11 \tag{72}$$

We conclude that the discrete skeleton of the process $(\tilde{\Pi}_\theta)_{\theta \geq 0}$ does not give a discrete Gibbs fragmentation associated with $w_j = (j-1)!$.

5.2 A reformulation with walks on the symmetric group

In view of the combinatorial interpretation of Example 2, problem 17 can be restated as:

Problem 18. *Does there exist for each n a sequence of random permutations $(\sigma_k, 1 \leq k \leq n)$ such that σ_k has uniform distribution on the set permutations of $[n]$ with k cycles, and for $k \leq \ell$ the partition generated by the cycles of σ_ℓ is a refinement of σ_k?*

This problem may be partially reformulated in terms of random walks on the symmetric group. Suppose we consider the Cayley graph G_n of the symmetric group induced by the set of generators $S = \{\text{all transpositions}\}$, that is, we put an edge between two permutations σ and π if and only if σ may be written as $\sigma = \tau \cdot \pi$ for some transposition τ. It is well-known that multiplying a permutation by a transposition can only result in a coagulation or a fragmentation in the cycle structure. More precisely, suppose $C = (x_1, \ldots, x_k)$ is a cycle of the permutation π. If we multiply by the transposition $\tau = (x_i, x_j)$ then the resulting cycle structure in σ is the same as that of π except that C breaks into $(x_1, \ldots, x_{i-1}, x_j, x_{j+1}, \ldots, x_k)$ on the one hand and $(x_i, \ldots, x_{j-1})$ on the other hand. Conversely, suppose $C = (x_1, \ldots, x_k)$ and $C' = (y_1, \ldots, y_l)$ are two cycles of π and we multiply π by the transposition (x_i, y_j). In the resulting permutation, C and C' will be replaced by a unique cycle $C'' = (x_1, \ldots, x_{i-1}, y_j, y_{j+1}, \ldots, y_{j-1}, x_i, \ldots, x_k)$. In particular, any (random) walk on G_n may be viewed as a coagulation and fragmentation process on the cycle structure of the permutation. Moreover, a well-known result due to Cayley states that if σ is a permutation then the *graph distance* between σ and the identity permutation I (i.e., the minimum number of edges one must cross to go from I to σ on G_n) is simply $n - \#\text{cycles of } \sigma$. As a consequence, by considering a time-reversal of the process, to solve Problem 18, it is enough to construct a random process $(\sigma_k)_{0 \leq k \leq n-1}$ on G_n which has the following property:

1. σ_k is a random walk on G_n, in the sense that if $\sigma_k = \sigma$ at the next stage σ can only jump to one the neighbors of σ.

2. σ_k has the following marginal distribution: at stage k σ_k is uniformly distributed on the sphere of radius k about the identity, that is the set of all permutations whose distance to the identity is k.

Property 1. ensures that the cycles of σ_k perform a coagulation fragmentation process. In conjunction with property 2., since σ_k must be at distance k, it must be the case that all jumps of σ are produced by some fragmentation. Moreover, by Cayley's result for the distance of a permutation to the identity, if σ_k is uniform on the sphere of radius k then its cycle structure is a realization of the Gibbs distribution (9) with weight sequence $w_j = (j-1)!$ (Note however this is not strictly equivalent to Problem 18 since not all fragmentations in the Gibbs sense are allowed by moving along edges of G_n).

In this context, a very natural process to consider is the simple random walk on G_n, conditioned to never backtrack. In other words, starting from the identity, at each step choose uniformly among all edges that lead from distance k to distance $k+1$. Although this may seem a very natural candidate for properties 1. and 2., this is far from being the case. Much is known about this process, and in particular it has been shown in [4] that the distribution of this process at time $k = \lfloor an \rfloor$ for any $0 < a < 1$ is asymptotically singular with respect to the uniform distribution on the sphere of radius k.

References

[1] D.J. Aldous. Deterministic and stochastic models for coalescence (aggregation and coagulation): a review of the mean-field theory for probabilists. *Bernoulli*, 5: 3–48, 1999.

[2] M.H. Bayewitz, J. Yerushalmi, S. Katz, and R. Shinnar. The extent of correlations in a stochastic coalescence process. *J. Atmos. Sci.*, 31:1604–1614, 1974.

[3] J. Berestycki. Exchangeable fragmentation-coalescence processes and their equilibrium distribution. *Electr. J. Prob.*, 9:770-824, 2004.

[4] N. Berestycki. The hyperbolic geometry of random transpositions. *Ann. Probab.*, to appear.

[5] L. Comtet. *Advanced Combinatorics.* D. Reidel Pub. Co., Boston, 1974. (translated from French).

[6] L. Devroye. The branching process method in the Lagrange random variate generation, `cgm.cs.mcgill.ca/~luc/branchingpaper.ps` (1992).

[7] R. Durrett, B. L. Granovsky, and S. Gueron. The equilibrium behavior of reversible coagulation-fragmentation processes. *J. Theoret. Probab.*, 12(2):447-474, 1999.

[8] P. Erdős, R.K. Guy, and J.W. Moon. On refining partitions. *J. London Math. Soc. (2)*, 9:565–570, 1970.

[9] W.J. Ewens. the sampling theory of selectively neutral alleles. *Theor. Popul. Biol.*, 3:87–112, 1972.

[10] P. J. Flory. *Princinples of polymer chemistry.* Ithaca, London: Cornell University Press 1953

[11] B. Friedst. The structure of partitions of large integers. *Trans. Amer. Math. Soc.*, 337:703-735, 1993.

[12] W. M. Y. Goh and E. Schmutz. Random set partitions *SIAM J. Discrete Math.*, 7:419-436, 1994.

[13] H.W. Gould. A series of transformations leading to convolution identities. *Duke Math. J.*, 28:193-202, 1961.

[14] E.M. Hendriks, J.L. Spouge, M. Eibl, and M. Shreckenberg. Exact solutions for random coagulation processes. *Z. Phys. B - Condensed Matter*, 58:219–227, 1985.

[15] L. Holst. On the lengths of the pieces of a stick broken at random. *J. Appl. Probab.*, 17:623 – 634, 1980.

[16] L. Holst. On number related to objects of unlike partitions and occupancy problems. *European J. Combin.* 2(3):231-237, 1981.

[17] F. Kelly. *Reversibility in stochastic networks*, Wiley, 1979

[18] J. F. C. Kingman. The coalescent. *Stochastic Processes and their Applications*, 13:235–248, 1982.

[19] V. F. Kolchin. *Random mappings.* Translation series in Mathematics and Engineering. Optimization Software Inc. Publications Division, New York, 1986. MR88:a60022.

[20] A.A. Lushnikov. Coagulation in finite systems. *J. Colloid and Interface Science*, 65:276–285, 1978.

[21] A.H. Marcus. Stochastic coalescence. *Technometrics*, 10:133–143, 1968.

[22] J. Pitman. Enumerations of trees and forests related to branching processes and random walks. *Microsurveys in Discrete Probability*, D. Aldous and J. Propp editors. DIMACS Ser. Discrete Math. Theoret. Comp. Sci no.41 163-180. Amer. Math. Soc. Providence RI, 1998.

[23] J. Pitman. Coalescent random forests. *J. Combin. Theory A.*, 85:165-193, 1999.

[24] J. Pitman. Combinatorial stochastic processes. *Lecture notes for the Saint-Flour course*, to appear (2001).

[25] A. Rényi. Probabilistic methods in combinatorial mathematics. In R.C. Bose and T.A. Dowlilng, eds. *Combinatorial mathematics and its applications*, page 1-13. Univ. of North Carolina Press, Chapel Hill 1969.

[26] P. Whittle. The equilibrium statistics of a clustering process in uncondensed phase. *Proc. Roy. Lond. Soc. A,* 285:501–519, 1965.

[27] P. Whittle. Statistical processes of aggregation and polymerisation. *Proc. Camb. Phil. Soc.*, 61:475–495, 1965.

[28] P. Whittle. *Systems in Stochastic Equilibrium.* Wiley, 1986.