Random Combinatorial Structures and Prime Factorizations

Richard Arratia, A. D. Barbour, and Simon Tavaré

Introduction

Many combinatorial structures decompose into components, with the list of component sizes carrying substantial information. An integer factors into primes—this is a similar situation, but different in that the list of sizes of factors carries *all* the information for identifying the integer. The combinatorial structures to keep in mind include permutations, mappings from a finite set into itself, polynomials over finite fields, partitions of an integer, partitions of a set, and graphs.

The similar behavior of prime factorization and cycle decompositions of permutations was observed by Knuth and Trabb Pardo [24]. We attempt to explain *why* such systems are similar.

We are interested in probability models which pick "a random combinatorial structure of size n", meaning that each of the objects of that size is equally likely. We also consider the model which picks an integer uniformly from 1 to n. Such models lead to stochastic processes that count the number of components of each conceivable size. What are the common features of these processes?

There are two broad areas of commonality. The first and most basic is essentially an algebraic property. It involves representing the distribution of the combinatorial process as that of a sequence of independent but not identically distributed random variables, conditioned on a weighted sum; see (5). All our combinatorial examples satisfy this exactly. On the other hand, prime factorizations of a uniformly chosen integer cannot be described in terms of conditioning a process of independent random variables on the value of a weighted sum because the value of the weighted sum in this case tells us the value of the random integer. However, by considering conditioning as a special case of the more general construction of "biasing" a distribution, we can view prime factorization as having a very close relative of the conditioning property. Conditioning independent random variables on various weighted sums has a long history; for combinatorial examples we refer the reader to Shepp and Lloyd [29], Holst [20], Kolchin [25], Diaconis and Pitman [13], and Arratia and Tavaré [9].

The second broad area of commonality, shared by some but not all of the examples listed above, is an analytic property. The number of components of size at most x has, for fixed x, a limit in distribution as $n \to \infty$, and the expected value of this limit is asymptotic to $\theta \log x$ as $x \to \infty$, where $\theta > 0$ is a constant. We call combinatorial structures that have this property "logarithmic". For the main examples in this paper the logarithmic structures are permutations, polynomials, mappings, the Ewens sampling formula, and prime factorizations, and the nonlogarithmic structure is that of integer partitions.

Richard Arratia is professor of mathematics at the University of Southern California. His e-mail address is rarratia@math.usc.edu.

Andrew Barbour is professor of mathematics at the University of Zürich. His e-mail address is adb@amath.unizh.ch.

Simon Tavaré is professor of mathematics and professor of biological sciences at the University of Southern California. His e-mail address is stavare@gnome.usc.edu.

Based in part on a lecture presented by Simon Tavaré in November 1995 at the University of Utah and on lectures presented by Richard Arratia in December 1996 at the Institute for Advanced Study, Lucent Technologies, and IBM Watson Research Labs.

Work supported in part by NSF grant No. DMS 96-26412 and Schweizerischer Nationalfonds grant Nr. 20-43453.95.

Combinatorial Examples

Consider a combinatorial structure which decomposes into components. Let $p(n)$ be the number of instances of size n. Given an instance of size n, the most basic description reports only the number k of components. We are interested in a fuller description, the component structure, specifying how many of these k components are of sizes one, two, three, and so on.

For a given combinatorial structure, the usual approach is to assume that n is fixed and to *count* how many of the $p(n)$ instances of size n have each particular component structure. Equivalently, one can think of drawing an instance at random from the uniform distribution over all $p(n)$ possibilities and ask for the *probability* of each particular component structure. In this random formulation, the counts of components of each size become random variables. We write C_i or $C_i(n)$ for the number of components of size i. Thus $(C_1(n), C_2(n), \ldots, C_n(n))$ specifies the entire component size counting process, and $K(n) := C_1(n) + C_2(n) + \cdots + C_n(n)$ is the total number of components. The random variables $C_1(n), C_2(n), \ldots$ are dependent, and in fact a certain weighted sum of them is constant:

$$(1) \qquad C_1(n) + 2C_2(n) + \cdots + nC_n(n) = n.$$

We illustrate the above concepts with four examples from combinatorics and one family of distributions from genetics that plays a unifying role. For each example we provide one instance, with $n = 10$. Some references are listed after the name of the example; these are not intended to be exhaustive, but rather pointers to the literature.

Example 1. Integer partitions *[17, 27]. Partition the integer n as $n = l_1 + l_2 + \cdots + l_k$ with $l_1 \geq l_2 \geq \cdots \geq l_k \geq 1$. For integer partitions, $p(n)$ is the traditional notation for the number of such partitions, and $\sum p(n)x^n = \prod_{i\geq 1}(1 - x^i)^{-1}$. We write $C_i(n)$ for the number of parts which are i, and the component counting structure $(C_1(n), \ldots, C_n(n))$ is an encoding of the partition. An instance for $n = 10$ is*

$$10 = 5 + 3 + 1 + 1.$$

In this instance $C_1(10) = 2, C_3(10) = C_5(10) = 1$, the other $C_i(n)$ being zero.

Example 2. Permutations *[29, 7]. Consider the cycle decomposition of a permutation of the set $\{1, 2, \ldots, n\}$, with $C_i(n)$ being the number of cycles of length i. The total number of instances of size n is $p(n) = n!$, and $C_1(n)$ is the number of fixed points. An instance for $n = 10$ is the function π with $\pi(1) = 9$, $\pi(2) = 1$, $\pi(3) = 7$, $\pi(4) = 4$, $\pi(5) = 3$, $\pi(6) = 2$, $\pi(7) = 5$, $\pi(8) = 8$, $\pi(9) = 10$, $\pi(10) = 6$, whose cycle decomposition is*

$$\pi = (1\ 9\ 10\ 6\ 2\ )\ (3\ 7\ 5\ )\ (4)\ (8).$$

In this instance $C_1(10) = 2, C_3(10) = C_5(10) = 1$.

Example 3. Mappings *[15, 25]. Consider all mappings from the set $\{1, 2, \ldots, n\}$ to itself, so that there are $p(n) = n^n$ possibilities. A mapping f corresponds to a directed graph with edges $(i, f(i))$ for $i = 1$ to n, and the "components" of f are precisely the connected components of the underlying undirected graph. An instance for $n = 10$ is the function f with $f(1) = 9$, $f(2) = 6$, $f(3) = 5$, $\pi(4) = 4$, $f(5) = 3$, $f(6) = 6$, $f(7) = 3$, $f(8) = 8$, $f(9) = 2$, $f(10) = 2$. In this instance $C_1(10) = 2, C_3(10) = C_5(10) = 1$. Note that the number of fixed points, 3 in this instance, is not $C_1(10)$.*

Example 4. Polynomials over GF(q) *[16, 3]. Consider monic polynomials of degree n over the finite field $\mathrm{GF}(q)$. Writing $f(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0$, we see that there are $p(n) = q^n$ possibilities. These polynomials can be uniquely factored into a product of monic irreducible polynomials, and $C_i(n)$ reports the number of irreducible factors of degree i. For the case $q = 2$, an instance with $n = 10$ is*

$$f(x) = x^{10} + x^8 + x^5 + x^4 + x + 1 =$$
$$(x + 1)^2(x^3 + x^2 + 1)(x^5 + x^4 + x^3 + x + 1).$$

In this instance $C_1(10) = 2, C_3(10) = C_5(10) = 1$.

Example 5. The Ewens Sampling Formula *[14]. The Ewens sampling formula (ESF) with parameter $\theta > 0$ is not in general a combinatorial model, but it does play a central role in our story. The model arose originally in population genetics, where the parameter θ is a mutation rate. See also Chapter 41 of [21].*

For each $n = 1, 2, \ldots$ and $\theta > 0$, the ESF is a distribution for $(C_1(n), C_2(n), \ldots, C_n(n))$. It gives the distribution of the cycle structure of a random permutation of n objects, choosing a permutation with probability biased by $\theta^{K(n)}$, where $K(n)$ is the number of cycles. For irrational θ these are certainly not models in combinatorics, but for $\theta = 1, 2, 3, 4, \ldots$ the ESF is the distribution of a "random permutation with colored cycles" in which there are θ colors available.

Independent Random Variables, Conditioned on a Weighted Sum

One unifying feature of our combinatorial examples is that each has a component structure that can be described in terms of a process of independent random variables $Z_1, Z_2, \ldots$, conditioned on the value of a weighted sum. We illustrate this in the example of random permutations.

Cauchy's formula says that for nonnegative integers $a_1, a_2, \ldots, a_n$ with $a_1 + 2a_2 + \cdots + na_n = n$, the number of permutations having a_i cycles of length i, for $i = 1$ to n, is $n!/\prod(a_i!\, i^{a_i})$. Pick-

ing a random permutation of n objects and choosing uniformly over the $n!$ possibilities, one can say that, for any $\mathbf{a} = (a_1, \ldots, a_n) \in \mathbb{Z}_+^n$,

$$\mathbb{P}((C_1(n), \ldots, C_n(n)) = \mathbf{a}) = \mathbb{1}(\sum_{j=1}^{n} j a_j = n) \prod_{j=1}^{n} \left(\frac{1}{j}\right)^{a_j} \frac{1}{a_j!}. \tag{2}$$

The formula above uses an indicator function, $\mathbb{1}(A) = 1$ if A is true, and 0 if not.

Now suppose that $Z_1, Z_2, \ldots$ are independent Poisson random variables with $\mathbb{E}Z_j = 1/j$. In contrast to (2),

$$\begin{aligned} \mathbb{P}((Z_1, \ldots, Z_n) = \mathbf{a} &= \prod_{j=1}^{n} \mathbb{P}(Z_j = a_j)) \\ &= \prod_{j=1}^{n} e^{-1/j} \left(\frac{1}{j}\right)^{a_j} \frac{1}{a_j!} \\ &= e^{-\sum_1^n 1/j} \prod_{j=1}^{n} \left(\frac{1}{j}\right)^{a_j} \frac{1}{a_j!}. \end{aligned} \tag{3}$$

Let T_n be the following weighted sum of the independent random variables:

$$T_n = Z_1 + 2Z_2 + \cdots + nZ_n.$$

It follows from (2) and (3) that

$$\mathbb{P}(T_n = n) = \sum_{\mathbf{a}:\ \sum j a_j = n} \mathbb{P}((Z_1, \ldots, Z_n) = \mathbf{a}) = e^{-\sum_1^n 1/j}. \tag{4}$$

Now

$$\mathbb{P}((Z_1, \ldots, Z_n) = \mathbf{a} \mid T_n = n) = \frac{\mathbb{1}(\sum j a_j = n)\mathbb{P}((Z_1, \ldots, Z_n) = \mathbf{a})}{\mathbb{P}(T_n = n)}.$$

Using (4), we see that this ratio simplifies to the expression in (2). This proves that the distribution of $(C_1(n), \ldots, C_n(n))$ equals the distribution of $(Z_1, \ldots, Z_n)$ conditional on the event $\{T_n = n\}$. All of our combinatorial processes satisfy an identity of this form; that is

$$\mathcal{L}((C_1(n), \ldots, C_n(n))) = \mathcal{L}((Z_1, Z_2, \ldots, Z_n) \mid T_n = n). \tag{5}$$

For a general treatment of (5) for combinatorial structures, see [9].

Logarithmic Combinatorial Structures

Some of our examples have a limit in distribution:

$$(C_1(n), C_2(n), \ldots, C_n(n), 0, 0, \cdots) \Rightarrow (Z_1, Z_2, \ldots). \tag{6}$$

In those examples where the limit exists, such as random permutations, random mappings, random polynomials over $GF(q)$, and the Ewens sampling formula, it turns out that the limit process $(Z_1, Z_2, \ldots)$ *has* independent coordinates which satisfy (5). For partitions of an integer and for partitions of a set, (6) is not satisfied, and each coordinate $C_i(n)$ goes off to infinity as n grows.

There are combinatorial examples, such as random forests, which satisfy both (5) and (6) but which still do not satisfy all our requirements for being "logarithmic". The condition that best characterizes the property of being a "logarithmic combinatorial structure" is that both (5) and (6) hold and, for some constant $\theta \in (0, \infty)$,

$$i\,\mathbb{E}Z_i \to \theta, \quad i\,\mathbb{P}(Z_i = 1) \to \theta \quad \text{as } i \to \infty. \tag{7}$$

The terminology "logarithmic" comes from the relation

$$\sum_{i \le x} \mathbb{E}Z_i \sim \theta \log x \tag{8}$$

as $x \to \infty$. A logarithmic combinatorial structure of size n tends to have around $\theta \log n$ components.

Continuum Limits for Logarithmic Combinatorial Structures

Scale invariant Poisson processes on $(0, \infty)$

What happens if we rescale the limit process $(Z_1, Z_2, \ldots)$ for a logarithmic combinatorial structure? Formally, consider the random measure $\mathcal{M}_n$ with mass Z_i at the point i/n, for $i \ge 1$. The independence of the Z_i means that for any system of nonoverlapping subintervals of $(0, \infty)$ the random measure $\mathcal{M}_n$ assigns independent masses. The logarithmic property implies that the expected mass assigned to an interval (a, b) is $\mathbb{E}\sum_{i/n \in (a,b)} Z_i \sim \sum_{na < i < nb} \theta/i \sim \theta \log(b/a) = \int_a^b \theta dx\,/x$.

The "scale invariant" Poisson process $\mathcal{M}$ on $(0, \infty)$ with intensity $\theta\,dx\,/x$ has exactly $\theta \log(b/a)$ for the expected number of points in any interval (a, b). Like the $\mathcal{M}_n$, the Poisson process assigns independent masses to nonoverlapping subintervals. It is not hard to show that for any logarithmic combinatorial structure satisfying (8) the random measures $\mathcal{M}_n$ converge in distribution to $\mathcal{M}$:

$$\mathcal{M}_n \Rightarrow \mathcal{M}. \tag{9}$$

The convergence above is characterized by integrating against continuous functions with com-

pact support. Observe that for the random variable T_n that appears in the conditioning (5) we have

$$\frac{T_n}{n} = \sum_{i \le n} \frac{i}{n} Z_i = \int_{(0,1]} x \mathcal{M}_n(dx).$$

Thus it is natural to anticipate from (9) that

$$T_n/n \Rightarrow T \tag{10}$$

where

$$T := \int_{(0,1]} x \mathcal{M}(dx). \tag{11}$$

The limit process $\mathcal{M}$ is simple. Since $\mathcal{M}$ has an *intensity* measure which is continuous with respect to Lebesgue measure, with probability one $\mathcal{M}$ has no double points. Thus we can identify $\mathcal{M}$ with a random discrete subset of $(0, \infty)$. In particular the points of $\mathcal{M}$ in $(0, 1]$ can be labeled X_i for $i = 1, 2, \dots$ with

$$0 < \cdots < X_2 < X_1 \le 1. \tag{12}$$

With this labeling, the integral in (11) is expressed as the sum of locations of all points of the Poisson process $\mathcal{M}$ in (0,1):

$$T = X_1 + X_2 + \cdots . \tag{13}$$

Computation with Laplace transforms shows that the density g_θ of T, with $g_\theta(x) = 0$ if $x < 0$, satisfies

$$x g_\theta(x) = \theta \int_{x-1}^{x} g_\theta(u) du, x > 0, \tag{14}$$

so that for $x > 0$,

$$x g'_\theta(x) + (1 - \theta) g_\theta(x) + \theta g_\theta(x - 1) = 0. \tag{15}$$

Equation (15) shows why $\theta = 1$ is special. See Vervaat [32], p. 90, and Watterson [33]. For the case $\theta = 1$, the density g_1 of T is $g_1(t) = e^{-\gamma} \rho(t)$, where γ is Euler's constant and ρ is Dickman's function [31].

The rescaled limit of the large components is the Poisson-Dirichlet process

The limit in (6) inherently focuses on the small components, in that convergence in distribution for infinite-dimensional random vectors is equivalent to convergence for the restrictions to the first b coordinates, for each fixed b. How can we discuss the limit distribution for the large components?

One way is to let $L_i(n)$ be the size of the i^{th} largest component of a combinatorial structure of size n, with $L_i(n) = 0$ whenever $i > K(n)$, the number of components. Note that the vectors $(C_1(n), C_2(n), \dots, C_n(n))$ and $(L_1(n), L_2(n), \dots, L_n(n))$ carry exactly the same information; each can be expressed as a function of the other.

Typically, for any fixed i and k, $\mathbb{P}(L_i(n) > k) \to 1$, so it would not be useful to ask for the limit of $(L_1(n), L_2(n), \dots)$. However, all the examples which satisfy (7) have a limit for the process of large components [18, 5]; rescaling all sizes by n,

$$\left(\frac{L_1(n)}{n}, \frac{L_2(n)}{n}, \dots \right) \Rightarrow (V_1, V_2, \dots). \tag{16}$$

The distribution of the limit is called the Poisson-Dirichlet distribution with parameter θ, after Kingman [22, 23]. It is most directly characterized by the density functions for its finite-dimensional distributions, which involve the density g_θ of T described in (13)–(15). The joint density of $(V_1, V_2, \cdots, V_k)$ is supported by points $(x_1, \dots, x_k)$ satisfying $x_1 > x_2 > \cdots > x_k > 0$ and $x_1 + \cdots + x_k < 1$, and at such points has value, in the special case $\theta = 1$,

$$\rho\left(\frac{1 - x_1 - x_2 - \cdots - x_k}{x_k} \right) \frac{1}{x_1 x_2 \cdots x_k}, \tag{17}$$

where ρ is Dickman's function. For the case of general $\theta > 0$ the expression for the joint density function is [33]

$$g_\theta\left(\frac{1 - x_1 - \cdots - x_k}{x_k} \right) \frac{e^{\gamma\theta} \theta^k \Gamma(\theta) x_k^{\theta - 1}}{x_1 x_2 \cdots x_k}. \tag{18}$$

The Poisson-Dirichlet process arises from the scale invariant Poisson process by conditioning

For all the combinatorial systems in this paper the discrete dependent process $(C_1(n), \dots, C_n(n))$ comes from the independent process $(Z_1, Z_2, \dots)$ by conditioning on $T_n = n$, as in (5). Restricting to the logarithmic class, each of the discrete ingredients in this, $(C_1(n), \dots, C_n(n))$ and $(Z_1, Z_2, \dots)$, can be rescaled to get a continuum limit in which only the parameter θ appears. For the dependent system $(C_1(n), \dots, C_n(n))$, the continuum limit is the Poisson-Dirichlet distribution of $(V_1, V_2, \dots)$, a dependent process. For the independent system $(Z_1, Z_2, \dots)$ the continuum limit is the scale invariant Poisson process $\mathcal{M}$, an "independent process". It is most natural to expect to fill in the fourth edge of this square diagram, and relate the dependent and independent continuum systems to each other by conditioning.

Theorem 1 *[5].* For any $\theta > 0$, let the scale invariant Poisson process on $(0, \infty)$, with intensity $\theta\, dx/x$, have its points falling in (0,1] labeled so that (12) holds. Let $(V_1, V_2, \dots)$ have the Poisson-Dirichlet distribution with parameter θ. Then

$$\mathcal{L}((V_1, V_2, \dots)) = \mathcal{L}(\,(X_1, X_2, \dots)\,|T = 1). \tag{19}$$

Prime Factorizations of Uniformly Chosen Integers

Our primary common theme has been that natural random combinatorial processes arise from independent random variables by conditioning on the value of a weighted sum, as in (5). Our secondary common theme is that many of these processes are "logarithmic", with the limit of the component counting process, viewed from the small end, being a process of independent random variables whose rescaled limit is the scale invariant Poisson process. Also, the logarithmic combinatorial structures, viewed from the large end, have a rescaled dependent process limit, the Poisson-Dirichlet process. How does the prime factorization of a random integer, chosen uniformly from 1 to n, fit into this picture? In brief, the factorization into primes fits the picture perfectly, although there are superficial changes.

As is usual in number theory, the dummy variable p is understood to denote an arbitrary prime, and $\pi(n)$ is the number of primes $\le n$. A positive integer i has a prime factorization of the form $i = \prod p^{a_p}$, which can be viewed as a decomposition

$$\log i = \sum a_p \log p. \tag{20}$$

Our probability model is to pick a random integer $N = N(n)$ uniformly from the n possibilities $1, 2, \dots, n$:

$$\mathbb{P}(N = i) = \frac{1}{n}, \quad i = 1, 2, \dots, n\,.$$

The prime factorization of this random integer,

$$N(n) = \prod p^{C_p(n)} \tag{21}$$

defines a process of random variables $C_p(n)$. The coordinates here are mutually dependent, and we write $C_p(n)$ to emphasize that we focus on the *distribution* rather than on the factorization of a particular integer. Taking logarithms in (21) gives

$$\log N = \sum C_p(n) \log p \ \le \log n, \tag{22}$$

which is similar to (1).

The superficial differences

The first difference between prime factorizations and our decomposable combinatorial structures is that the coordinates are indexed by primes p rather than by positive integers i. More importantly, the possible component sizes, which show up as the *weights* on the left-hand side in (1) and (22), are not $1, 2, 3, \dots$, but rather $\log 2, \log 3, \log 5, \dots$.

The second difference is that the overall system size, which shows up on the right-hand sides of (1) and (22), is for prime factorizations not the parameter n, but rather $\log n$. Thus, for example, the Hardy-Ramanujan theorem, that a "normal" integer n has around $\log\log n$ prime divisors, is like the statement that a random permutation of n objects typically has about $\log n$ cycles. Both statements say that a system of size s has typically about $\log s$ components.

The third difference is that for prime factorizations, the size of a particular random choice within the system of size n is not constant, but rather

$$\log N = \sum C_p(n) \log p,$$

which is uniformly distributed over the set $\{0, \log 2, \log 3, \log 4, \dots, \log n\}$. In particular, the analog of (5) cannot be something involving conditioning on the exact value of a weighted sum of independent random variables.

The similarities with other logarithmic combinatorial structures

The first similarity is that prime factorizations satisfy an analog of (6). For primes the independent random variables Z_p that arise in the limit are geometrically distributed:

$$\mathbb{P}(Z_p \ge k) = p^{-k},$$
$$\mathbb{P}(Z_p = k) = (1 - 1/p)p^{-k}, \qquad k = 0, 1, 2, \dots\,.$$

Take the natural enumeration of primes in order of size, $p_1 = 2, p_2 = 3, p_3 = 5, \dots$. In the analog of (6), which is

$$(C_p(n))_p \Rightarrow (Z_p)_p \quad \text{as } n \to \infty, \tag{23}$$

convergence in distribution means simply that for each fixed k, $(C_{p_1}(n), \dots, C_{p_k}(n)) \Rightarrow (Z_{p_1}, \dots, Z_{p_k})$ as $n \to \infty$. This can be easily verified in terms of cumulative distribution functions, as follows. Let $a_1, \dots, a_k \in \mathbb{Z}_+$ be given, and let $r = 2^{a_1} 3^{a_2} \cdots p_k^{a_k}$. Then as events, $\{C_{p_i} \ge a_i$ for $i = 1$ to $k\} = \{\, r|N\,\}$, with probability $(1/n)\lfloor n/r \rfloor$. This converges to $1/r$, and $1/r = \prod_1^k p_i^{-a_i} = \prod_1^k \mathbb{P}(Z_{p_i} \ge a_i) = \mathbb{P}(Z_{p_i} \ge a_i$ for $i = 1$ to $k)$.

A second similarity is that the Z_p satisfy the analog of (8). Recalling that the "system size" is not n but rather $\log n$, the analog of (8) with $\theta = 1$ is that

$$\sum_{\log p \le \log n} \mathbb{E}Z_p \ = \sum_{p \le n} \frac{1}{p - 1} \ \sim \log(\log n).$$

Note that the analog of (7) would be that $(\log p)\mathbb{E}Z_p \to 1$, which is not true.

A third similarity is a Poisson-Dirichlet ($\theta = 1$) limit for the rescaled sizes of the large components. List the prime factors of our random integer N as

$P_1 \geq P_2 \geq \cdots \geq P_K$ where $K = \sum C_p$ is the number of prime factors of N, including multiplicities. To have enough coordinates to fill out the process, define $P_i = 1$ whenever $i > K$. The process giving the sizes of the components of our random integer is $(\log P_1, \log P_2, \ldots)$; this is the analog of $(L_1, L_2, \ldots)$. Rescaling by the system size, which is $\log n$, and writing $P_i(n)$ to emphasize that we are considering a joint distribution which depends on the parameter n, we find that the analogy suggests the result proved by Billingsley [12]:

$$\left(\frac{\log P_1(n)}{\log n}, \frac{\log P_2(n)}{\log n}, \ldots \right) \Rightarrow (V_1, V_2, \ldots). \tag{24}$$

A fourth similarity is that the limit process (Z_p) rescaled converges in distribution to the scale invariant Poisson process, as in (9). The random measure $\mathcal{M}_n$ for primes is $\mathcal{M}_n := \sum Z_p \, \delta(\log p / \log n)$ which puts down mass Z_p at the location $\log p / \log n$. With this definition, (9) holds. Corresponding to (10) is the statement that $T_n / \log n \Rightarrow T$, where

$$T_n = \sum_{p \leq n} Z_p \log p, \tag{25}$$

and T is from (13) with $\theta = 1$. These results for the independent process are easy but may have been written down first in [1].

Not exactly conditioning, but the closest possible analog

A direct attempt to start with (5) and find an analog for the prime factorization of a uniformly distributed random integer is doomed to fail, since the value of $T_n = \sum_{p \leq n} Z_p \log p$ completely determines the vector $(Z_p)_{p \leq n}$, thanks to unique prime factorization.

Just as the random integer $N(n)$ encodes the process $(C_p(n))_p$, the independent process $(Z_p)_{p \leq n}$ can be encoded in a random integer. We define

$$M \equiv M(n) := \exp(T_n) = \prod_{p \leq n} p^{Z_p}. \tag{26}$$

The procedures for defining N and M are in sharp contrast. For N pick a number at random, and consider the distribution of the multiplicities of primes in its factorization. For M pick those multiplicities independently at random, and look at the resulting distribution of integers. As n varies, the $M(n)$ all use the same multiplicities (Z_p); the dependence on n is only through the range of the product. The random integer M may be larger than n, but it is always free of primes $p > n$.

For $i \leq n$, $\mathbb{P}(N = i) = 1/n$ does not vary with i, while for an integer i free of primes larger than n, say $i = \prod_{p \leq n} p^{a_p}$, we have

$$\begin{aligned} \mathbb{P}(M(n) = i) &= \prod_{p \leq n} \mathbb{P}(Z_p = a_p) \\ &= \prod_{p \leq n} (1 - 1/p) p^{-a_p} = c(n)/i; \end{aligned}$$

with a normalizing constant

$$c(n) = \prod_{p \leq n} (1 - 1/p).$$

Thus, to convert from the distribution of the independent process, encoded as the values of $\mathbb{P}(M(n) = i)$, into the distribution of the dependent process, encoded as the values $\mathbb{P}(N = i) = (1/n)\mathbb{1}(i \leq n)$, not only do we condition on $i \leq n$, which corresponds to conditioning on the event $\{T_n \leq \log n\}$, but we also bias with a factor proportional to i. In summary, for all positive integers i,

$$\begin{aligned} &\mathbb{P}(N(n) = i) \\ &\quad = \mathbb{P}(M(n) = i) \left(\prod_{p \leq n} (1 - 1/p)^{-1} \right) \mathbb{1}(i \leq n) \, \frac{i}{n} \end{aligned} \tag{27}$$

We can view biasing and conditioning in a unified framework as follows. In the context of random elements A, B of a discrete space $\mathcal{X}$, one says that "the distribution of B is the h-bias of the distribution of A" if for all $\alpha \in \mathcal{X}$, $\mathbb{P}(B = \alpha) = c_h \, h(\alpha) \mathbb{P}(A = \alpha)$, where the normalizing constant c_h may be expressed as $c_h = (\mathbb{E}h(A))^{-1}$. Starting from the given distribution for A, one can form this h-biased distribution if and only if $h(\alpha) \geq 0$ for all α such that $\mathbb{P}(A = \alpha) > 0$ and $0 < \mathbb{E}h(A) < \infty$. Conditioning on an event of the form $\{A \in S\}$, where $S \subset \mathcal{X}$, is exactly the case of biasing where h is an indicator function, $h(\alpha) = \mathbb{1}(\alpha \in S)$, and the normalizing constant is $c_h = 1/\mathbb{P}(A \in S)$. For our examples, let A be the independent process either $A = (Z_1, Z_2, \ldots, Z_n)$ for the combinatorial processes, or $A = (Z_p)_{p \leq n}$, which can be encoded as $M(n)$, for the prime factorizations. Similarly, let B be the dependent process: either $B = (C_1(n), \ldots, C_n(n))$ for the combinatorial processes, or for the prime factorizations, $B = (C_p(n))_{p \leq n}$, which can be encoded as $N(n)$.

The conditioning relation (5) can be viewed as the statement that the distribution of B is the h-bias of the distribution of A, where $h(A)$ is the indicator function of the event $\{T_n = n\}$. The relation (27) also says that distribution of B is the h-bias of the distribution of A, but now $h(A) = \mathbb{1}(T_n \leq \log n) \exp(T_n - \log n)$, corresponding to the last two factors of (27).

The close similarity of these two versions of biasing shows in the asymptotics of the normalizing factor. For logarithmic combinatorial structures having $\theta = 1$, in particular for our examples 2 and 4, the constant is

$$c_h = \mathbb{P}(T_n = n)^{-1} \sim e^{\gamma} n,$$

i.e. the exponential of Euler's constant, times the system size. For prime factorizations, the constant is the first factor on the right side of (27), $c_h = \prod_{p \le n}(1 - 1/p)^{-1}$, with $c_h \sim e^{\gamma} \log n$ by Mertens's theorem. Reading this as e^{γ} times the system size, one can see that prime factorizations have exactly the same asymptotics as examples 2 and 4.

Exploiting the similarity of primes and permutations

The similarity of logarithmic combinatorial structures having $\theta = 1$ (such as permutations) and prime factorizations leads us to try to use one system to study the other. In particular, results for permutations lead to new conjectures in number theory. We describe two examples from [1], one still open, the other recently worked out. We also quote a third conjecture, as (28), without including the somewhat long story of how it relates to permutations.

First, we describe an open problem. Thanks to the "Feller coupling" in [2], we know for permutations that the independent system $(Z_1, \ldots, Z_n)$ can be converted to the dependent system $(C_1(n), \ldots, C_n(n))$ using, on average, $2 + o(1)$ changes. More specifically, in the Feller coupling, to convert to $(C_1(n), \ldots, C_n(n))$, one of the coordinates Z_j is increased by one, and a random selection of coordinates are decreased, with the number of decrements having mean $1 + o(1)$. This leads to two versions of a conjecture about prime factorizations according to whether or not some small exceptional probability is allowed.

The first version of the conjecture is that couplings of $M = M(n)$ and $N = N(n)$, defined in (26) and (21), exist in which to change from M to N, one prime factor is inserted into M, and a random number of prime factors are deleted from M, the number of deletions having mean $1 + o(1)$. With probability approaching zero, additional prime factors may also be inserted into M.

The second version of the conjecture is a direct translation from the situation for permutations and is simpler to state: simply remove the *possibility* of inserting *more* than one prime factor. Thus, the conjecture is that there is a coupling using a single insertion and a random number of deletions. In other words, we conjecture that one can construct, on a single probability space, random integers $M(n)$ and $N(n)$ and a random prime P_0 such that N always divides $P_0 M$. The first version is a moral certainty to be true and should be provable using analysis. For the second version we have no strong convictions about whether or not the conjecture is true, but it might actually be neater to prove this stronger conjecture: since it does not involve any $o(1)$ error bound, conceivably there is a purely algebraic or combinatorial proof.

Second, we tell the story of a conjecture that lasted only half a year before being proved. In the 1950s Kubilius [26] proved his "fundamental lemma", which may be stated as follows. Let $u = u(n)$ and $\beta = 1/u$. Let $A = A(n)$ be the independent process, observing all primes with $\log p \le \beta \log n$, i.e $A = (Z_p)_{\log p \le \beta \log n}$. Similarly, let $B = B(n)$ be the dependent process, observing all small prime factors of an integer chosen uniformly from 1 to n, where the small primes p are those with $\log p \le \beta \log n$. Kubilius proved that the total variation distance $d_{TV}(A, B)$ tends to zero if $\beta \to 0$, together with an upper bound of the form $d_{TV}(A, B) = O(\exp(-(u/8)\log u) + n^{-1/15})$. What happens if we do not have $\beta(n) \to 0$? In particular, what happens if β is constant?

The natural conjecture, from [1], is that prime factorizations have the same behavior as random permutations. In [7] an explicit strictly monotone function $H : [0, 1] \to [0, 1]$ was described, with $H(0) = 0, H(1) = 1$, together with a heuristic argument that for permutations, for the case β constant $\in [0, 1]$ while $n \to \infty$, looking at $A = (Z_i)_{i \le \beta n}$ versus $B = (C_i(n))_{i \le \beta n}$, $d_{TV}(A, B) \to H(\beta)$. A proof of this was given in [30]. A simpler characterization of H, which follows easily from (19), is that $H(\beta)$ *is* the total variation distance between the restrictions to $[0, \beta]$ of the Poisson-Dirichlet process with $\theta = 1$ and the corresponding scale invariant Poisson process $\mathcal{M}$. Finally, it is proved in [10] that for prime factorizations, for any constant $\beta \in [0, 1]$, $d_{TV}(A, B) \to H(\beta)$. In terms of Buchstab's function ω and Dickman's function ρ [31], H can be described as follows: for $0 < \beta < 1$, with $u = 1/\beta$,

$$\begin{aligned} 2H(\beta) &= e^{\gamma}\mathbb{E}|\omega(u - T) - e^{-\gamma}| + \rho(u) \\ &= \int_{t>0} |\omega(u - t) - e^{-\gamma}|\rho(t)dt + \rho(u). \end{aligned}$$

Discussion

For random combinatorial structures, and for prime factorizations, identifying the limit processes is a first step. The next step involves giving bounds on the rates of convergence, under various metrics on the spaces involved. Estimates for the logarithmic class are studied in [5, 6], Hansen and Schmutz [19], and Stark [30], and applications of such estimates are given in [1, 8, 11].

For comparing the discrete dependent process with its independent discrete limit, it is very effective to consider the total variation distance, as Kubilius did for primes. Even for combinatorial structures which are not logarithmic, such as integer partitions and set partitions, the total variation distance comparisons with independent processes are useful [27, 28].

For the large components of logarithmic combinatorial structures, where the limit is the Poisson-Dirichlet process, the total variation distance method is useless, since as always when compar-

ing a discrete distribution with a continuous distribution the total variation distance is identically one. One way around this is to use the Ewens Sampling Formula with parameter θ (example 5) as the comparison object in place of the Poisson-Dirichlet distribution with parameter θ. Here the ESF may be considered as the "discrete analog" of the Poisson-Dirichlet, in that both are one parameter families, with no place for "small scale" structure. For example, both random mappings and ESF($\theta = 1/2$) look almost the same at the large end, and this may be quantified by bounds on the total variation distance between their respective processes $(C_{b+1}(n), C_{b+2}(n), \ldots, C_n(n))$, observing only large components [5].

Two other metrics are especially useful for the processes in this paper. One is the Wasserstein distance on $\mathbb{Z}_+^n$ or $\mathbb{Z}_+^{\pi(n)}$, which in our context measures the expected number of changes needed to convert the dependent component counting process to its independent limit. Another is the l_1 Wasserstein distance on $\mathbb{R}^\infty$. In the context of prime factorizations and the Poisson-Dirichlet limit in (24), this is the infimum, over all conceivable couplings, of

$$\mathbb{E} \sum_{i \geq 1} |\log P_i(n) - (\log n) V_i|. \tag{28}$$

We conjecture that this is $O(1)$; see [1, 4].

References

[1] R. ARRIATA, *Independence of prime factors: Total variation and Wasserstein metrics, insertions and deletions, and the Poisson-Dirichlet process,* preprint, 1996.

[2] R. ARRATIA, A. D. BARBOUR, and S. TAVARÉ, *Poisson process approximations for the Ewens Sampling Formula*, Ann. Appl. Probab. **2**:519–535, 1992.

[3] ____, *On random polynomials over a finite field*, Math. Proc. Cambridge Philos. Soc. **114**:347–368, 1993.

[4] ____, *Expected l_1 distance in Poisson-Dirichlet approximations for random permutations: A tale of three couplings*, 1997 (in preparation).

[5] ____, *Logarithmic combinatorial structures*, 1997 (in preparation).

[6] R. ARRATIA, D. STARK, and S. TAVARÉ, *Total variation asymptotics for Poisson process approximations of logarithmic combinatorial assemblies*, Ann. Probab. **23**:1347–1388, 1995.

[7] R. ARRATIA and S. TAVARÉ, *The cycle structure of random permutations*, Ann. Probab. **20**:1567–1591, 1992.

[8] ____, *Limit theorems for combinatorial structures via discrete process approximations*, Random Structures Algorithms **3**:321–345, 1992.

[9] ____, *Independent process approximations for random combinatorial structures*, Adv. Math. **104**:90–154, 1994.

[10] R. A. ARRATIA and D. STARK, *A total variation distance invariance principle for primes, permutations, and Poisson-Dirichlet*, preprint, 1996.

[11] A. D. BARBOUR and S. TAVARÉ, *A rate for the Erdos-Turán law*, Combinatorics, Prob. Comput. **3**:167–176, 1994.

[12] P. BILLINGSLEY, *On the distribution of large prime divisors*, Period. Math. Hungar. **2**:283–289, 1972.

[13] P. DIACONIS and J. W. PITMAN, *Permutations, record values and random measures*, unpublished lecture notes, Statistics Department, University of California, Berkeley, 1986.

[14] W. J. EWENS, *The sampling theory of selectively neutral alleles*, Theoret. Population Biol. **3**:87–112, 1972.

[15] P. FLAJOLET and A. M. ODLYZKO, *Random mapping statistics*, Proc. Eurocrypt '89 (J.-J. Quisquater, ed.), Lecture Notes in Comput. Sci., Springer-Verlag, 1990, pp. 329–354.

[16] P. FLAJOLET and M. SORIA, *Gaussian limiting distributions for the number of components in combinatorial structures*, J. Combin. Theory A **53**:165–182, 1990.

[17] B. FRISTEDT, *The structure of random partitions of large integers*, Trans. Amer. Math. Soc. **337**:703–735, 1993.

[18] J. C. HANSEN, *Order statistics for decomposable combinatorial structures*, Random Structures Algorithms **5**:517–533, 1994.

[19] J. C. HANSEN and E. SCHMUTZ, *How random is the characteristic polynomial of a random matrix*? Math. Proc. Cambridge Philos. Soc. **114**:507–515, 1993.

[20] L. HOLST, *A unified approach to limit theorems for urn models*, J. Appl. Probab. **16**:154–162, 1979.

[21] N. S. JOHNSON, S. KOTZ, and N. BALAKRISHNAN, *Discrete multivariate distributions*, Wiley, New York, 1997.

[22] J.F.C. KINGMAN, *Random discrete distributions*, J. Roy. Statist. Soc. **37**:1–22, 1975.

[23] ____, *Poisson processes*, Oxford University Press, Oxford, 1993.

[24] D. E. KNUTH and L. TRABB PARDO, *Analysis of a simple factorization algorithm*, Theoret. Comput. Sci. **3**:321–348, 1976.

[25] V. F. KOLCHIN, *Random mappings*, Optimization Software, Inc., New York, 1986.

[26] J. KUBILIUS, *Probabilistic methods in the theory of numbers*, Transl. Math. Mono., vol. 11, Amer. Math. Soc., Providence, RI,1962.

[27] B. PITTEL, *On the likely shape of the random Ferrers diagram*, Adv. Appl. Math., 1997 (in press).

[28] ____, *Random set partitions: Asymptotics of subset counts*, J. Combin. Theory A 1997 (in press).

[29] L. A. SHEPP and S. P. LLOYD, *Ordered cycle lengths in a random permutation*, Trans. Amer. Math. Soc. **121**:340–357, 1966.

[30] D. STARK, *Total variation distance for independent process approximations of random combinatorial objects*, Ph.D. thesis, University of Southern California, 1994.

[31] G. TENENBAUM, *Introduction to analytic and probabilistic number theory*, Cambridge Stud. Adv. Math., vol. 46. Cambridge University Press, 1995.

[32] W. VERVAAT, *Success epochs in Bernoulli trials with applications in number theory*, Mathematical Center Tracts, vol. 42, Mathematisch Centrum, Amsterdam, 1972.

[33] G. A. WATTERSON, *The stationary distribution of the infinitely-many-alleles diffusion model*, J. Appl. Probab. **13**:639–651, 1976.