Bull. London Math. Soc. 53 (2021) 1563–1582 doi:10.1112/blms.12537

The magical Ewens sampling formula

Simon Tavaré

This article is dedicated to Warren Ewens, friend, mentor and collaborator, in honour of the 50th anniversary of his celebrated sampling formula.

Abstract

Starting with n cooked spaghetti strands, tie randomly chosen ends together to produce a collection of spaghetti hoops. What is the expected number of hoops? What can be said about the distribution of the number of hoops of length 1, 2, ...? What is the behaviour of the longest hoops when n is large? What is the probability that all the hoops have different lengths? Questions like these appear in many guises in many areas of mathematics, one connection being their relation to the Ewens sampling formula (ESF). I will describe a number of related examples, including prime factorisation, random mappings and random permutations, illustrating the central role played by the ESF. I will also discuss methods for simulating decomposable combinatorial structures by exploiting another wonder of the ESF world, namely the Feller coupling (FC). Analysis of the spaghetti game in which ends of *different* strands must be tied shows that apparently small departures from the FC can open up a number of unsolved problems. Several past presidents of the London Mathematical Society (LMS) have contributed to the theory around the ESF, as I will illustrate.

1. *Introduction*

This article, based on my presidential address delivered 10 November 2017, gives a personal view of developments around the theme of the Ewens sampling formula (ESF) — some new, some old. The ESF has been studied extensively, and it arises in many different settings in probability and statistics. Arratia, Barbour and Tavaré [4] describe numerous applications in combinatorics, and James and Kerber [27, Chapter 41] and Crane [16] provide many other examples.

I have presented the topics in the order of my presidential address, and have used much the same motivation, in terms of simple games of chance. Probability theory developed from this perspective, and although the examples might seem somewhat light-hearted, they provide excellent motivation for deeper issues. I have made no attempt to be comprehensive — the field is too extensive, and the space here limited — and important topics such as the Pitman sampling formula [44, 46] and the connections with Bayesian statistics are absent; Crane [16] provides useful pointers to this literature.

The paper begins with a brief overview of mathematical population genetics, to highlight the scientific focus from which the ESF developed. Several former presidents of the London Mathematical Society (LMS) have contributed to the theory discussed here. Brief synopses of these contributions appear in the shadowed boxes in the text.

Received 12 May 2021; published online 27 October 2021.

2020 *Mathematics Subject Classification* 60C05, 60J10, 65C05, 60J10 (primary).

Presidential Address to the London Mathematical Society delivered at its Annual General Meeting in London on 10 November 2017.

2. *Mathematical population genetics*

Mathematical population genetics has a long history dating back to the seminal work of Fisher, Haldane and Wright in the 1920s and 1930s. They studied the effects of forces such as mutation, selection and recombination on the structure of genetic variation in natural populations, allowing for demographic effects such as migration, admixture, subdivision and fluctuations in population size. Much of the theory might be characterised as *prospective*: determine how the population will evolve in the future from its present state. The relative roles of selection and neutrality on the evolution of gene frequencies have been a major topic of research.

With the advent of molecular data in the late 1960s, the paradigm changed from prospective to *retrospective*: given information about molecular variability measured in a sample of individuals, identify the forces that acted in the past to produce the present state. Charlesworth and Charlesworth[**15**] provide a brief, accessible overview of the subject.

For our purposes, it is sufficient to highlight two early data sets, coming from the fruit fly species *Drosophila tropicalis* and *Drosophila simulans*. The data are in Table 1.

TABLE 1. *Allele frequencies observed at the Esterase-2 locus in* D. tropicalis *and the Esterase-2 locus in* D. simulans. *Each sample resulted in seven different types; the frequencies of the types are shown in the third column.*

	Sample size	Allele frequencies
D. tropicalis	298	234, 52, 4, 4, 2, 1, 1
D. simulans	308	91, 76, 70, 57, 12, 1, 1

The frequencies in the two samples seem to have different shapes, the *D. simulans* sample being ‘flatter’ than the *D. tropicalis* one. Wright [**55**, p. 303] argued that

> *. . . the observations do not agree at all with the equal frequencies expected for neutral alleles in enormously large populations.*

This raised the question of what the distribution of allele frequencies should look like under neutrality. The answer was provided by the ESF [**22**], to which we now turn.

2.1. *The Ewens sampling formula*

Ewens [**22**] derived the joint probability distribution of the number of selectively neutral alleles $C_j(n)$ represented j times $(j = 1, 2, \ldots, n)$ in a sample of n genes taken from a large population. For non-negative integers $c_1, c_2, \ldots, c_n$, he showed that

$$\mathbb{P}(C_1(n) = c_1, \ldots, C_n(n) = c_n) = \mathbb{1}\left(\sum_{j=1}^{n} jc_j = n\right) \frac{n!}{\theta_{(n)}} \prod_{j=1}^{n} \left(\frac{\theta}{j}\right)^{c_j} \frac{1}{c_j!}\,, \tag{1}$$

for $\theta \in (0, \infty)$, and $\theta_{(n)} := \theta(\theta + 1) \cdots (\theta + n - 1) = \Gamma(n + \theta)/\Gamma(\theta), \theta_{(0)} = 1$. This distribution is known as the ESF, $\theta > 0$ is being a parameter related to the rate at which novel alleles appear; we will see in Section 2.4 that the probability that gene $n + 1$ is a new type is $\theta/(\theta + n)$. In what follows we denote the law in (1) by ESF(θ).

2.2. *The number of types, K_n*

The number of types observed in the sample is $K_n = C_1(n) + \cdots + C_n(n)$; its probability distribution is given by

$$\mathbb{P}(K_n = k) = \frac{\theta^k \left[{n \atop k}\right]}{\theta_{(n)}}, \quad k = 1, 2, \ldots, n, \tag{2}$$

where $\left[{n \atop k}\right]$ is the unsigned Stirling number of the first kind, and its probability generating function is given by

$$\mathbb{E}s^{K_n} = \sum_{k=1}^{n} \mathbb{P}(K_n = k)s^k = \frac{(\theta s)_{(n)}}{\theta_{(n)}} = \prod_{j=1}^{n} \left(\frac{j-1}{\theta + j - 1} + \frac{\theta s}{\theta + j - 1} \right). \tag{3}$$

Equation (3) shows that K_n may be represented as the sum of n independent, but not identically distributed, Bernoulli random variables $\xi_1, \xi_2, \ldots, \xi_n$ satisfying

$$\mathbb{P}(\xi_j = 1) = \frac{\theta}{\theta + j - 1}, \quad \mathbb{P}(\xi_j = 0) = \frac{j-1}{\theta + j - 1}, j = 1, 2, \ldots, n; \tag{4}$$

as a consequence,

$$\mathbb{E}K_n = \sum_{j=1}^{n} \frac{\theta}{\theta + j - 1}.$$

As will be seen, the random variables $\xi_1, \ldots, \xi_n$ play an important role in the sequel.

2.3. *Consequences*

The arrival of the ESF had a dramatic effect on the development of statistical inference in population genetics, and in this section I outline why. First of all, dividing (1) by (2) shows that

$$\mathbb{P}(C_1(n) = c_1, \ldots, C_n(n) = c_n \mid K_n = k) = \mathbb{1}\left(\sum_{j=1}^{n} jc_j = n, \sum_{j=1}^{n} c_j = k \right) \frac{n!}{\left[{n \atop k}\right]} \prod_{j=1}^{n} \left(\frac{1}{j}\right)^{c_j} \frac{1}{c_j!}, \tag{5}$$

which is independent of θ. In statistical parlance, the number K_n of different alleles observed in the sample is sufficient for the parameter θ. It follows from the Rao–Blackwell theorem that estimation of θ should be based on K_n; earlier, estimation of θ had been based on the observed allele frequencies. Furthermore, the conditional distribution in (5) can be used for testing the goodness of fit of the model to data, a point to which we return later.

Not all was good news, however. Estimation of θ is based on the distribution in (2), and Ewens showed that the maximum likelihood estimator (MLE) of θ is the solution $\hat{\theta}_n$ of the equation $k = \mathbb{E}K_n$, where k is the observed number of types in the sample. Furthermore, the MLE $\hat{\theta}_n$ is asymptotically Normal with mean θ and variance $\theta/\log n$. The slow rate is due to dependence among the genes in the sample, and symbolises why statistical inference in population genetics is hard: effective sample sizes are roughly the logarithm of the observed sample size.

Finally, what can be said about our two data sets? We noted that the conditional distribution in (5) may be used to assess goodness of fit. The statistic one chooses to design a test depends on the alternative hypothesis of interest. Watterson [53] suggested use of $H = \sum_{i=1}^{k} x_i^2$, where x_i is the relative frequency of the ith of the k types in the sample. H is larger for unbalanced allele frequencies; for the *tropicalis* data, $H = 0.647$, while for the *simulans* data, $H = 0.236$. Empirical p-values, calculated as the fraction of 10^5 simulated values of H that are smaller than the observed values, are 0.04 for the *simulans* sample and 0.87 for the *tropicalis* sample.

Thus the *uneven* allele frequencies seem to be consistent with the ESF, as opposed to the more even frequencies predicted by Wright. Presumably there are other genetic forces at work in the *simulans* sample. In Section 6 a method for simulating the null distribution of H is discussed in a little more detail.

2.4. *Laws of succession*

One step in Ewens' argument leading to (1) established that the probability that the $(n+1)$st gene sampled is a type that has not been found in the first n sampled genes is

$$\theta/(\theta+n), \tag{6}$$

while the probability it is a copy of a particular existing allele present in m copies is

$$m/(\theta+n). \tag{7}$$

The connection with (4) appears in Section 4.1.

It is convenient to imagine sequential sampling from an infinite collection of distinguishable species, whose random relative frequencies $P_1, P_2, \ldots$ satisfy

$$0 < P_j < 1, j = 1, 2, \ldots; \quad \sum_{j=1}^{\infty} P_j = 1.$$

Suppose that the history of the first n samples is given. Donnelly [20] showed that if the conditional probability that the next sampled animal is of a previously unsampled species depends only on n, then this probability must be of the form (6). Furthermore, if in addition the conditional probability that the next animal chosen is of a particular species seen $m > 0$ times in the sample depends only on m and n, then Zabell [57] showed that the counts of the number of species sampled once, twice, ... must be ESF(θ). The law of succession represented by (6) and (7) is related to work of De Morgan described briefly in Figure 1.

Consider a sequence of independent Bernoulli trials with unknown probability p of success. Laplace's rule of succession states that the probability that the $(n+1)$st trial results in a success, given r successes in the first n trials and all values of p equally likely, is $(r+1)/(n+2)$; see [56] for the background.

De Morgan [19] considered a related problem in which the set of possible outcomes is not known in advance. He defined the probability that a new type is found on the $(n+1)$st trial as $1/(n+k+1)$, and that a type with m prior occurrences is chosen as $(m+1)/(n+k+1)$, where k is the number of types observed in the first n trials. The resulting sampling formula is not the ESF.

FIGURE 1 (colour online). *A. De Morgan, first president* (1865–1866).

3. *Population processes*

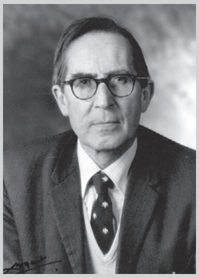
Stochastic models of growth in biological populations follow two major traditions, one originating from population genetics, the other from population dynamics. Population genetics models, including those of Wright, Fisher and Moran, initially assumed constancy of the population size through time. Results such as the ESF and Kingman's coalescent (see Figure 5) were derived in this setting. The population dynamics tradition, typified by branching process models (see Figure 2), emphasises growth and stochastic, fluctuations stemming from birth and death events of a finite collection of independent individuals. The two fields connect in many places, for example, [28, 34], and more recent work on modelling ancestral relationships among extant individuals through coalescent point processes [41].

One relevant connection centres around Fisher's logarithmic series distribution [23], which initiated the statistical modelling of the distribution of species counts obtained from specimens sampled from a population composed of different species. Letting C_j denote the number of species having j representatives in the sample, Watterson [52] interprets Fisher's model as one for means, rather than probabilities:

$$\mathbb{E}C_j \approx \theta\frac{x^j}{j}, \tag{8}$$

for some $x > 0$. Other connections between the ESF and species sampling are described in [17].

Karlin and McGregor [28] describe a birth-and-death process with immigration at rate θ and birth and death rates equal to 1 for which the counts $C_j = C_j(t)$ of family sizes at time t are independent Poisson-distributed random variables with means given by (8) with $x = t/(1+t)$. The connection with the ESF is made more transparent in Section 4.3, but we note for now that [48] uses the birth process with immigration as a convenient way to study the asymptotic behaviour of the ESF, and in particular the Poisson–Dirichlet distribution in Section 4.4.2.



© Godfrey Argent Studio

Kendall had a long-standing interest in the theory and applications of branching processes, particularly those of birth-and-death type [30, 31, 34]. His lecture on the occasion of the centenary of the London Mathematical Society [32] and his presidential address [33] described the history of branching processes.

[29] analysed the linear birth-and-death process with immigration, and discussed limiting regimes that result in the size N of an extant colony (individuals descended from a single immigrant) having the logarithmic series distribution of the form

$$\mathrm{P}(N=n) \propto x^n/n, \ \ n = 1, 2, \dots$$

FIGURE 2. *D. G. Kendall, 56th president* (1972–1974).

4. *The ESF and biased permutations*

Of particular interest here is its appearance as the distribution of the cycle counts of a θ-biased permutation. Let π be a permutation of $[n] := \{1, 2, \dots, n\}$ decomposed as a product of cycles, as illustrated in Figure 3. If π is chosen uniformly with probability $1/n!$, then Cauchy's formula establishes that the cycle counts $(C_1(n), \dots, C_n(n))$ have the ESF(1) law [25], and if a permutation π having k cycles is chosen with probability proportional to θ^k, then the cycle counts have the ESF(θ) law. In this case,

$$\mathbb{P}(\pi) = \frac{\theta^k}{\theta_{(n)}},$$

if the permutation π has k cycles.

The cycle counts may be studied in several ways, of which two probabilistic approaches use the independent Bernoulli random variables $\xi_1, \xi_2, \dots$ with distribution given by (4). The first method, the *Chinese restaurant process* (CRP), simulates a biased permutation of $[n]$ using the ξ_i in the order $\xi_1, \xi_2, \ldots, \xi_n$, whereas the second, the *Feller coupling* (FC), achieves the same end using the reverse order $\xi_n, \xi_{n-1}, \ldots, \xi_1$.

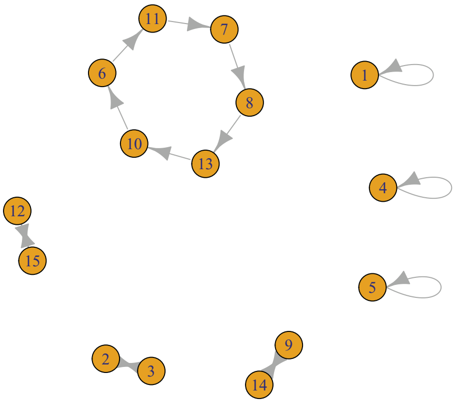


FIGURE 3 (colour online). *Random permutation of* $[15]$, *decomposed into cycles.* $K_{15} = 7$, $C_1(15) = 3$, $C_2(15) = 3$, $C_6(15) = 1$.

4.1. *The Chinese restaurant process*

The CRP, devised by Dubins and Pitman (see [1]), proceeds as follows. Integer 1 starts a cycle. Integer 2 is placed to the right of 1, in the same cycle, if $\xi_2 = 0$, or begins a new cycle if $\xi_2 = 1$.

Suppose, then, that the first $n-1$ integers have been assigned to cycles. Integer n starts a new cycle if $\xi_n = 1$, an event of probability $\theta/(\theta + n - 1)$, or is placed to the right of a uniformly chosen integer j, in the same cycle, if $\xi_n = 0$, $j = 1, 2, \ldots, n-1$. For any permutation π of $[n]$ having k cycles, $\mathbb{P}_\theta(\pi) = \theta^k/\theta_{(n)}$, so the CRP generates permutations with the ESF(θ) distribution.

The cycles generated in this way are ordered, in that the first contains the integer 1, the second cycle the smallest integer not in the first cycle and so on, and as the process evolves, cycle lengths may be changed.

4.2. *The Feller coupling*

The FC was introduced in [5] as a way to generate the cycles in a growing permutation one at a time, and it has proved very useful in the study of the asymptotics of properties of the ESF. The cycle counts are determined by the spacings between the 1s in realisations of $\xi_i, i \geqslant 1$. If we define $C_j(n)$ to be the number of spacings of length j between the 1s in $1\,\xi_2\,\xi_3 \cdots \xi_n\,1$, then the distribution of $(C_1(n), \ldots, C_n(n))$ is, of course, ESF(θ). The magic occurs because

$$Z_j = C_j(\infty) = \text{number of spacings of length } j \text{ in } 1\,\xi_2\,\xi_3\,\ldots \tag{9}$$

are independent Poisson-distributed random variables with $\mathbb{E}Z_j = \theta/j$. Further details may be found in [5] and [4, Chapter 5]. We note that, as in the description of the CRP, the permutations themselves, as opposed to just their cycle lengths, may be generated by an auxiliary randomisation; a new cycle is begun with the smallest unused integer, and a cycle is grown by adding a randomly chosen unused integer at its end.

Aside from its use as an analytically tractable coupling of $(C_1(n), \ldots, C_n(n))$ and $(Z_1, Z_2, \ldots)$, the FC is very useful for simulating permutations decomposed into cycles. In [4] it is shown that the CRP takes of order $O(n)$ calls to the random number generator to produce a sample of size n, whereas the FC takes of order $O(\log n)$.

4.3. *The conditioning relation*

Watterson [51] showed that $\mathrm{ESF}(\theta)$ is the distribution of independent Poisson random variables, conditioned on a weighted sum. Defining $Z_1(x), Z_2(x), \dots$ to be independent Poisson random variables with

$$\mathbb{E}Z_j(x) = \frac{\theta x^j}{j}, \quad j = 1, 2, \dots, \tag{10}$$

for $x > 0$, we have

$$\mathbb{P}(C_1(n) = c_1, \dots, C_n(n) = c_n) = \mathbb{P}(Z_1(x) = c_1, \dots, Z_n(x) = c_n \mid T_n(x) = n), \tag{11}$$

where

$$T_n(x) = Z_1(x) + 2Z_2(x) + \cdots + nZ_n(x).$$

We will see that the random variables $Z_j := Z_j(1)$ are the $C_j(\infty)$ in (9).

4.4. *Limit laws*

4.4.1. *Limit laws for the cycle counts.* We revisit the conditioning relation (11) with $x = 1$, and set $Z_j := Z_j(1)$, with means $\mathbb{E}Z_j = \theta/j$. One of the classical results about the ESF is that the probability distribution $\mathcal{L}(C_1(n), \dots, C_b(n))$ converges to that of $(Z_1, \dots, Z_b)$: for any $b \geqslant 1$,

$$(C_1(n), C_2(n), \dots, C_b(n)) \Rightarrow (Z_1, Z_2, \dots, Z_b) \quad \text{as } n \to \infty. \tag{12}$$

This is due to [25] for the case $\theta = 1$; see [4, Chapter 5.1] for history and further details. As it stands, (12) is only helpful in approximating results determined by fixed values of b as $n \to \infty$. However, there is now an extensive literature on metrising this convergence, particularly in the total variation metric. For example, [5] shows that for $b = b_n \leqslant n$,

$$d_{TV}(\mathcal{L}(C_1(n), \dots, C_b(n)), \mathcal{L}(Z_1, \dots, Z_b)) = O(b/n), \quad n \to \infty. \tag{13}$$

Error estimates of this type may be exploited to derive many limiting results for θ-biased permutations using simple methods; see [4, 9], for example.

© The Royal Society

Towards the end of his life Burnside focused on foundational issues in probability theory [12], giving principles with which probabilities could be computed. He discusses, *inter alia*, inclusion-exclusion, Bayes Theorem, geometrical probability, and (in our notation) showed that for $\theta = 1$,

$$\mathbb{P}(C_1(n) = r) \to \frac{e^{-1}}{r!}, \; r = 0, 1, 2, \dots$$

in the context of a problem involving the number of fixed points of a random permutation; see Section 7.2.2.

The book was published after Burnside's death with the help of AR Forsyth, twenty-first president (1904–1906), who also wrote Burnside's memoir for the Royal Society.

FIGURE 4. *W. Burnside, 22nd president* (1906–1908).

4.4.2. *Limit laws for the longest cycle lengths.* The limiting structure of the cycle lengths in decreasing order is given by the Poisson–Dirichlet distribution, denoted by $\mathrm{PD}(\theta)$; see Figure 5 for further details. Denoting by $L_1(n), L_2(n), \dots$ the length of the longest cycle, the next longest and so on, Kingman [35] showed that

$$n^{-1}(L_1(n), L_2(n), \dots) \Rightarrow (L_1, L_2, \dots),$$

where $(L_1, L_2, \ldots)$ has the PD(θ) distribution. Metric bounds on the distance to PD(θ) appear in [7]. The law of the size-biased PD(θ) distribution is known as the GEM(θ) distribution.

© University of Bristol

Kingman [**35**] defined the Poisson-Dirichlet law PD(θ) as the joint distribution of the points $\sigma_1 > \sigma_2 > \cdots$ of a Poisson process on $(0, \infty)$ with intensity $\theta e^{-x}/x$, normalized by their sum, σ; $(\sigma_1/\sigma, \sigma_2/\sigma, \ldots)$ and σ are independent.

There are many characterisations of PD(θ), summarised for example in [**4**, Chapter 5.5 and 5.7].

Kingman's coalescent [**38**, **36**, **37**], which models the ancestral relationships among a sample of individuals, revolutionised population genetics, and provided an alternative approach to properties of the ESF.

FIGURE 5. *J. F. C. Kingman, 65th president* (1990–1992).

5. *The combinatorics connection*

My address was motivated at least in part by a problem in Winkler's *Mathematical Mind-Benders* book [**54**]:

> *Spaghetti loops. The* 100 *ends of* 50 *strands of cooked spaghetti are paired at random and tied together. How many pasta hoops*[†]*, should you expect to result from this process, on average?*

This problem has arisen in many guises, for example, as the 'blades of grass' game described in [**24**] and an oft-asked Cambridge homework question. This begs more sophisticated questions, such as 'What is the chance that all the hoops have different lengths?' This can be addressed for a given value of n, say 50, or as $n \to \infty$. It is helpful to have a simple framework within which to analyse such questions.

Starting with $n = 50$ cooked strands, we have 100 ends. Artificially labelling these ends 1 to 100, the random choices begin with end 1 making a 99-way choice to determine which end to join; finishing a hoop at this first step corresponds to the event $\xi_{50} = 1$, having probability 1/99 and starting the new hoop with the smallest available unused strand. So if $n - r$ strands are already assigned to hoops, we have $\xi_r = 1$ if the last strand in the nascent hoop is joined to the other end of the strand that started that hoop (probability $1/[2(r-1)+1]$), and otherwise $\xi_r = 0$. In this way, the lengths of the hoops formed, in order, are the spacings between the ones in sequence $1\,\xi_n\,\xi_{n-1}\,\ldots\xi_1$.

This is an example of the FC, and it remains to compute θ. Since

$$\mathbb{P}(\xi_i = 1) = \frac{1}{2i-1} = \frac{1/2}{1/2 + i - 1},$$

we have identified $\theta = 1/2$, and this allows us to read off the structure of the hoop lengths from what we know about ESF($\theta = 1/2$), for fixed n and in the limit. In answer to the original problem, the number of hoops has distribution given by (2) with $\theta = 1/2$, so that

$$\mathbb{E}K_{50} = \sum_{j=1}^{50} \frac{1}{2j-1} \approx 2.938,$$

perhaps smaller than expected.

[†] Winkler's problem is phrased in terms of spaghetti *loops*, whereas I use the UK term, *hoops*.

5.1. *Random mappings*

A mapping from $[n]$ to $[n]$ may be constructed from independent and identically distributed random variables B_i, $i = 1, \dots, n$ satisfying

$$\mathbb{P}(B_i = j) = \frac{1}{n}, \quad j = 1, 2, \dots, n; \tag{14}$$

B_i is the image of i. The components of the mapping are formed by iteration: l and m are in the same component if some iterate of l equals some iterate of m. Components are therefore directed cycles of rooted labelled trees; an example with $n = 15$ appears in Figure 6.

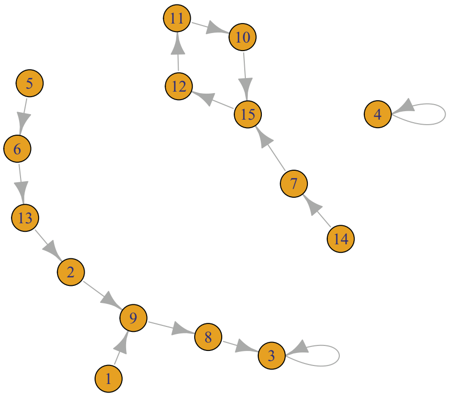


FIGURE 6 (colour online). *Random mapping of* [15], *decomposed into components.* $K_{15} = 3$, $C_1(15) = 1$, $C_6(15) = 1$, $C_8(15) = 1$. *There are six elements on cycles.*

Random mappings are a combinatorial assembly, the distribution of the component counts being given by (11), where the independent Poisson random variables $Z_j(x)$ having means given by

$$\mathbb{E}Z_j(x) = \frac{1}{j}\,\frac{m_j x^j}{(j-1)!} \tag{15}$$

where m_j, the number of possible structures of size j, is given by

$$m_j = e^j\,(j-1)!\,\mathbb{P}(\mathrm{Po}(j) < j)$$

rather than the values in (10); here $\mathrm{Po}(\lambda)$ denotes a random variable having a Poisson distribution with expectation λ.

Random mappings are one of the most studied combinatorial assemblies. Other values of the m_j lead to different assemblies, with component distribution determined as above, and the reader is referred to [4, Chapter 2.1] for further examples, such as polynomials over $\mathrm{GF}(q)$ and additive arithmetic semigroups. For the present purposes, we identify x to simplify the subsequent results, and this leads us to the class of logarithmic assemblies.

5.2. *The logarithmic assemblies*

We continue to assume that the $Z_j = Z_j(x)$ are independent Poisson-distributed random variables with means given in (15). If we can choose x so that $Z_j := Z_j(x)$ satisfies

$$j\,\mathbb{P}(Z_j = 1) \to \theta \quad \text{and} \quad j\mathbb{E}(Z_j) \to \theta, \quad \text{as } j \to \infty, \tag{16}$$

for some $\theta \in (0, \infty)$, we call the assembly *logarithmic*. For assemblies satisfying

$$\frac{m_j}{j!} \sim \frac{\theta y^j}{j} \quad \text{as } j \to \infty,$$

for some $y > 0$, $\theta > 0$, we can take $x = 1/y$ to express them in logarithmic form. For mappings,

$$\frac{m_j}{j!} = e^j\,\mathbb{P}(\mathrm{Po}(j) < j)/j,$$

so that $x = e^{-1}$ and

$$\mathbb{E}Z_j = \frac{1}{j}\mathbb{P}(\mathrm{Po}(j) < j);$$

in this case, we may take $\theta = 1/2$, since the central limit theorem shows that $\mathbb{P}(\mathrm{Po}(j) < j) \to 1/2$ as $j \to \infty$. For examples that are not logarithmic, see Figure 7.

The fact that the number m_j of trees on j labeled vertices is $m_j = j^{j-2}$ is known as Cayley's formula [**11**, **14**]; see also [**45**] for a beautiful double counting proof. It follows that the assembly of forests of labeled, unrooted trees is not logarithmic in the sense of (16), because

$$\frac{m_j}{j!} \sim \frac{1}{\sqrt{2\pi}}\frac{e^j}{j}\,j^{-3/2},\ j \to \infty.$$

Forests of labeled, rooted trees, for which $m_j = j^{j-1}$, are also not logarithmic.

FIGURE 7 (colour online). *A. Cayley, third president* (1868–1870).

6. *Simulating assemblies*

We have seen that samples from the ESF can be generated efficiently using the FC, but this leaves the issue of how to simulate from logarithmic assemblies. One possible solution is the rejection method described in [**3**]. Denoting the means $\mathbb{E}Z_j = \lambda_j/j$, we have

$$\begin{aligned}\mathbb{P}(C_1(n) = c_1, \ldots, C_n(n) = c_n) &\propto \prod_{j=1}^{n}\left(\frac{\lambda_j}{j}\right)^{c_j}\frac{1}{c_j!}\\ &\propto \left[\prod_{j=1}^{n}\left(\frac{\lambda_j}{\theta}\right)^{c_j}\right]\frac{n!}{\theta_{(n)}}\prod_{j=1}^{n}\left(\frac{\theta}{j}\right)^{c_j}\frac{1}{c_j!}. \qquad (17)\end{aligned}$$

The term in (17) is proportional to a function $h(c_1, \ldots, c_n)$ of $(c_1, \ldots, c_n)$ × the probability of $(c_1, c_2, \ldots c_n)$ under ESF(θ). It follows that if we can find θ such that $0 \leqslant \lambda_j \leqslant \theta$ for $j = 1, 2, \ldots, n$, we have a simple rejection algorithm:

- simulate $(c_1, \ldots, c_n)$ from ESF(θ);
- accept $(c_1, \ldots, c_n)$ as a realisation from $\mathcal{L}(C_1(n), \ldots, C_n(n))$ for the assembly with probability

$$h(c_1, \ldots, c_n) = \prod_{j=1}^{n}\left(\frac{\lambda_j}{\theta}\right)^{c_j} \leqslant 1;$$

- otherwise, reject $(c_1, \ldots, c_n)$ and start again.

The performance of this method is discussed in [**3**], where the asymptotic acceptance rate is given.

The conditioning relation may also be used to simulate samples, and the parameter x in the Poisson means may be chosen as a function of n to optimise the acceptance rate. The details of this method appear in [**3**], but for the present purposes this method is not competitive, and will not be discussed further here.

6.1. *Examples*

We begin with a comment on simulation of values of the test statistic H used in the genetic example in Section 2.3. The task is to simulate values from the distribution of the ESF conditional on a given number, k, of types. Recalling that this distribution is independent of θ, we are free to choose θ in a rejection method that uses the FC to simulate from ESF(θ), and accepts the cycle counts if the number of cycles equals k. The choice of θ is naturally the MLE of θ, as this, by definition, maximises the acceptance probability of the method. For both data sets, the MLE of θ is approximately 1.15, and this leads to an acceptance rate of about 17%, acceptable for these examples. For larger sample sizes, a method which uses every run seems necessary.

For the spaghetti hoops problem, simulation of 10^6 realisations using the FC with $n = 50$ estimates the probability that no hoop lengths are equal to be 0.838, in good agreement with the asymptotic value obtained in [**9**], namely $2e^{-\gamma/2}/\sqrt{\pi} \approx 0.8455$.

For a random mapping, we can take $\theta = 1/2$ and use the FC and rejection method. 10^6 simulations of a mapping with $n = 50$ estimated the chance it has no repeated component sizes as 0.888, close to the limiting value $\approx$0.896 found in [**3**].

7. *Variants on a theme*

7.1. *Spaghetti hoops, revisited*

This section introduces a twist to the spaghetti hoop problem mentioned in the introduction, by considering the case in which no strand can be tied directly to its other end. Of interest is the probability distribution of sizes of spaghetti hoops, their asymptotic behaviour, and the elucidation of their relationship to the ESF. We begin with a description of the process.

We build spaghetti hoops with a sequence of 0s and 1s, just as in the FC, but the joint distribution of the Bernoullis, denoted here by $\eta_{n+1}, \eta_n, \ldots, \eta_1$, is different: because no spaghetti strand can be tied directly to its other end, no singleton hoops can occur. $\eta_{n+1} = 1$ is playing the role of the 1 in position $n + 1$ in the FC.

It is convenient to label the spaghetti strands $\{1, 2, \ldots, n\}$. Strand 1 starts the process, and the nascent hoop is (1. Strand 1 cannot be tied to its other end, and so complete a hoop; the Bernoulli random variable η_n that indicates whether a hoop is completed satisfies $\mathbb{P}(\eta_n = 0 \mid \eta_{n+1} = 1) = 1$.

Strand 1 is tied at random to an end of one of the remaining $n - 1$ strands. Label the chosen strand as i_n, and the resulting, growing hoop becomes $(1\, i_n)$.

Strand i_n is tied at random to one of the remaining $2(n - 2) + 1$ free strand ends; call this strand j for now. If $j = 1$ a hoop is completed, and we set $\eta_{n-1} = 1$, close the current hoop and start the new hoop with the smallest unused strand, say i_{n-1}; this results in $(1\, i_n)(i_{n-1})$. The probability of this event is $\mathbb{P}(\eta_{n-1} = 1 \mid \eta_n = 0) = 1/(1 + 2(n - 2)) = 1/2/(n - 2 + 1/2)$. The next step involves a randomly chosen end from one of the $n - 3$ remaining strands, and we see that $\mathbb{P}(\eta_{n-2} = 0 \mid \eta_{n-1} = 1) = 1$.

In the event that strand $j \neq 1$, $\eta_{n-1} = 0$ and we set $i_{n-1} = j$ to get the current, growing hoop $(1\, i_n\, i_{n-1})$. We then have $\mathbb{P}(\eta_{n-1} = 0 \mid \eta_n = 0) = (n - 2)/(n - 2 + 1/2)$.

Continuing in this way we construct the spaghetti hoops in order. To calculate the probability of closing a hoop when there are $r - 1$ strands remaining to be chosen, there are two cases to consider. If $\eta_{r+1} = 0$ then $\eta_r = 1$ if the strand at the growing end of the hoop is tied to the free end of the strand that started that hoop. This event has probability $\mathbb{P}(\eta_r = 1 \mid \eta_{r+1} = 0) = 1/(2(r-1)+1) = 1/2/(r-1+1/2)$; otherwise, $\eta_r = 0$. On the other hand, if $\eta_{r+1} = 1$ then $\eta_r = 0$, as no hoops of length one are allowed; that is, $\mathbb{P}(\eta_r = 0 \mid \eta_{r+1} = 1) = 1$. An example is given in Figure 8.

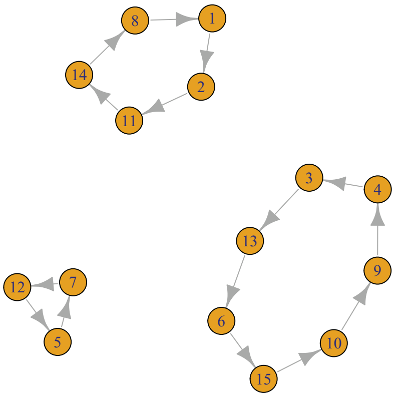


FIGURE 8 (colour online). *Random spaghetti hoops for* $n = 15$. $K_{15} = 3$, $C_3(15) = 1$, $C_5(15) = 1$, $C_7(15) = 1$.

We have now identified the structure of the process $\eta_{n+1}, \eta_n, \eta_{n-1}, \dots, \eta_1$. It is a (non-homogeneous) Markov chain that starts from $\eta_{n+1} = 1$ and has transition matrices P_r given by

$$P_r = \begin{pmatrix} \mathbb{P}(\eta_r = 0 \mid \eta_{r+1} = 0) & \mathbb{P}(\eta_r = 1 \mid \eta_{r+1} = 0) \\ \mathbb{P}(\eta_r = 0 \mid \eta_{r+1} = 1) & \mathbb{P}(\eta_r = 1 \mid \eta_{r+1} = 1) \end{pmatrix}$$
$$= \begin{pmatrix} \dfrac{r-1}{r-1+1/2} & \dfrac{1/2}{r-1+1/2} \\ 1 & 0 \end{pmatrix}, \tag{18}$$

for $r = n,\ n-1, \dots, 3$. There are two further loose ends to be tied up, namely when $r = 2$ and $r = 1$:

$$P_2 = \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}, \quad P_1 = \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix}.$$

As in the earlier spaghetti hoop model, the ordered hoop lengths are the spacings between the 1s in the sequence $\eta_1, \eta_2, \dots, \eta_n, 1$ read from right to left. The transition matrices in (18) are the special case $\theta = 1/2$ of the more general model that has

$$P_r = \begin{pmatrix} \dfrac{r-1}{r-1+\theta} & \dfrac{\theta}{r-1+\theta} \\ 1 & 0 \end{pmatrix}, \tag{19}$$

for some $\theta \in (0, \infty)$.

7.1.1. *Enumerating the state space.* The hoop lengths formed from n strands are found from the spacings between the 1s in the sequence $\eta_1 = 1, \eta_2 = 0, \dots, \eta_n = 0, \eta_{n+1} = 1$, where each $\eta_i \in \{0, 1\}$ and there are no consecutive 1s in the string. Writing h_n for the number of such strings, and recalling that the number b_n of binary strings of length n with no repeated 1s is $b_n = F_{n+2}$, where F_j is the jth Fibonacci number, we see that $h_n = b_{n-3} = F_{n-3+2} = F_{n-1}$. Hence

$$h_4 = 2, h_5 = 3, h_6 = 5, h_7 = 8, h_8 = 13, h_9 = 21, h_{10} = 34, h_{11} = 55, \dots, h_{50} = 7{,}778{,}742{,}049.$$

The Markov chain may be used to compute probabilities of patterns for small values of n. Table 2 gives an example for the case $n = 7, \theta = 0.5$; we return to this later. The Markov chain may also be used to simulate realisations of the hoop lengths, and so estimate probabilities of events of interest. For example, for $n = 50$ the chance that no hoops have the same length is ≈ 0.912.

TABLE 2. *Ordered cycle length probabilities for $n = 7$, $\theta = 0.5$. Patterns correspond to $\eta_1 = 1, \eta_2, \dots, \eta_7, \eta_8 = 1$.*

Pattern	Probability from (18)
10000001	0.55411
10100001	0.13853
10101001	0.02020
10100101	0.01558
10010001	0.11544
10010101	0.01299
10001001	0.08081
10000101	0.06234

As might be expected, it is the structure of the small hoop lengths that distinguishes the present example from the setting in Section 5. The asymptotic structure of the first, second, ... hoop lengths generated by the Markov chain is, however, the same for both models: direct calculation shows that when scaled by $1/n$ the relative sizes are asymptotically GEM(θ) with parameter $\theta = 1/2$, and the relative sizes of the largest, second largest, ... hoops are asymptotically Poisson–Dirichlet with $\theta = 1/2$; see [4, Chapter 5.4] for the ESF(θ) case. The asymptotic distribution of the number of hoops of size 2, 3, 4, ... is unknown, although empirically it seems to be close to Poisson.

7.2. *Other games*

7.2.1. *Children's playground game.* This example describes a playground game played by children at camp. The n children in the group stand close to each other, and some child, labelled #1, with one of their hands, grasps a hand of a different randomly chosen child (labelled #2). Child #2, with their free hand, grasps the randomly chosen hand of some child, (possibly the free hand of child #1). The process continues like this, with children only grasping 'free' hands. Thus one forms a collection of circles of children holding hands. Note though that no child is allowed to grasp their other hand.

A few moments thought, and allowance for finishing the game with all children involved, reveals that the sequence $\eta_{n+1} = 1, \eta_n, \dots, \eta_1 = 1$ that can be used to describe the ordered circle lengths behaves *precisely* as in (18). At its heart, this playground game is richer than the spaghetti hoop game in Section 7.1; for example, it is of interest to consider cases where all the children are looking in, or looking out. We hope to explore this in more detail elsewhere.

7.2.2. *Derangements.* Students of probability often meet random derangements in versions of the letters-and-envelopes problem, quoted here from [12, p. 21]:

> *There are n letters and n corresponding envelopes, and one letter is put into each envelope. This can be done in $n!$ ways. It is assumed that each two of*

these distributions are equally likely. What is the probability (i) *that just r letters go into their corresponding envelopes,* (ii) *that no letter goes into its corresponding envelope?*

If we define a permutation π by setting $\pi(i)$ to be the label of the envelope to which letter i is returned, the probability we seek in answer to (ii) is

$$\mathbb{P}(\pi \text{ is a derangement}) = \frac{D_n}{n!} = \sum_{l=0}^{n} \frac{(-1)^l}{l!},$$

where D_n is the nth derangement number. For the answer to (i), see Figure 4.

The distribution of the cycle length counts of a random derangement is precisely that of ESF(1) conditioned to have $C_1(n) = 0$, and the corresponding case for any θ is also of interest. Once more, these lengths can be generated by a Markov chain $\eta_{n+1} = 1, \eta_n, \dots, \eta_1 = 1$ with the same state space as that described in Section 7.1.1, but the transition matrix is *not* that in (19). By reverse engineering the probabilistic structure of the FC, we can construct a Markov chain with the property that it generates realisations of the original process $\xi_n, \dots, \xi_2, \xi_1$ in such a way that the joint law of $\eta_{n+1} = 1, \eta_n, \dots, \eta_1 = 1$ is the same as that of $1, \xi_n, \dots, \xi_2, \xi_1$ conditioned on having no consecutive 1s. Da Silva, Jamshidpey and Tavaré [**18**] establish that the transition matrices are given by

$$P_r = \begin{pmatrix} \dfrac{(\theta+r-1)\lambda_r(\theta)}{(\theta+r-1)\lambda_r(\theta)+\theta\lambda_{r-1}(\theta)} & \dfrac{\theta\lambda_{r-1}(\theta)}{(\theta+r-1)\lambda_r(\theta)+\theta\lambda_{r-1}(\theta)} \\ 1 & 0 \end{pmatrix}, \tag{20}$$

for $r = n-1, \dots, 3$,

$$P_2 = \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}, \quad P_1 = \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix}.$$

In (20) the quantity $\lambda_n(\theta)$ is the probability of the conditioning event obtained from ESF(θ):

$$\lambda_n(\theta) := \mathbb{P}(C_1(n) = 0) = \frac{n!}{\Gamma(n+\theta)} \sum_{j=0}^{n} (-1)^j \frac{\theta^j}{j!} \frac{\Gamma(n+\theta-j)}{(n-j)!},$$

with $\lambda_0(\theta) = 1, \lambda_1(\theta) = 0$.

Warren Ewens conjectured that the distribution of the hoop length counts in Section 7.1 is ESF$(\theta = 1/2)$, conditioned on having no singleton hoops. This is not the case, as the comparison in Table 3 shows.

TABLE 3. *Ordered cycle length probabilities for $n = 7$, $\theta = 0.5$. Patterns correspond to $\eta_1 = 1, \eta_2, \dots, \eta_7, \eta_8 = 1$.*

Pattern	Probability from (18)	Probability from (20)
10000001	0.55411	0.58323
10100001	0.13853	0.14581
10101001	0.02020	0.01823
10100101	0.01558	0.01458
10010001	0.11544	0.09721
10010101	0.01299	0.00972
10001001	0.08081	0.07290
10000101	0.06234	0.05832

7.2.3. *Screaming Toes game.* The following problem comes from Cameron [13, p. 154]:

> *n people stand in a circle. Each player looks down at someone else's feet (i.e., not at their own feet). At a given signal, everyone looks up from the feet to the eyes of the person they were looking at. If two people make eye contact, they scream. What is the probability q_n, say, of at least one pair screaming?*

To set this in our framework, we can define a mapping by exploiting independent, but this time not identically distributed, random variables satisfying

$$\mathbb{P}(B_i = j) = \frac{1}{n-1}, \quad j \neq i, i = 1, \dots, n.$$

Just as in the classical case defined by (14), the mapping is decomposed into components defined by iteration, the core of the mapping (those elements in the cycles) being a derangement. An example is given in Figure 9.

A detailed analysis of this type of random mapping is given in [49], and we record the essential details for comparison with the standard random mapping in Section 5.1. The number of structures of size j is now given by

$$\tilde{m}_j = (j-1)!\, e^j \sum_{l=0}^{j-2} \frac{e^{-j} j^l}{l!} = (j-1)!\, e^j\, \mathbb{P}(\mathrm{Po}(j) < j-1), \quad j \geqslant 2,$$

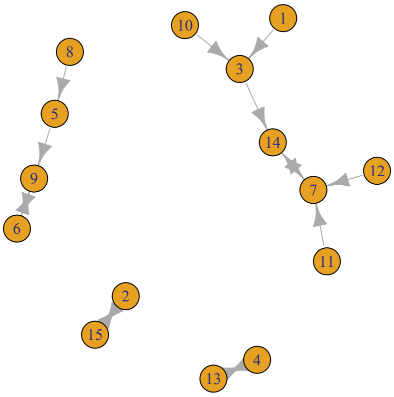


FIGURE 9 (colour online). *Random mapping for the Screaming Toes game with $n = 15$, decomposed into components. $K_{15} = 4$, $C_2(15) = 2$, $C_4(15) = 1$, $C_7(15) = 1$. There are eight elements on cycles, and four screaming pairs.*

and the joint law of $(\tilde{C}_2(n), \dots, \tilde{C}_n(n))$ is given by a version of the conditioning relation (11), but with the $Z_j(x)$ having means

$$\mathbb{E}Z_j(x) = \frac{\tilde{m}_j x^j}{j!}, \quad j = 2, 3, \dots,$$

for any $x > 0$. To make the structure logarithmic we set $x = e^{-1}, \theta = 1/2$ as before, resulting in

$$\mathbb{E}Z_j := \frac{\lambda_j}{j} = \frac{1}{j}\, \mathbb{P}(\mathrm{Po}(j) < j-1), \quad j = 2, 3, \dots \tag{21}$$

which should be compared to the standard case in (10).

Denoting the number of cycles of length two in the core by $\tilde{C}_2^*(n)$, the answer to Cameron's question is given by

$$q_n = \mathbb{P}(\tilde{C}_2^*(n) > 0),$$

which may be found from the distribution of $C_2^*(n)$ in [49, Lemma 2]:

$$\mathbb{P}(\tilde{C}_2^*(n)=0)=\sum_{l=0}^{\lfloor n/2\rfloor}(-1)^l\left(\frac{1}{2}\right)^l\frac{1}{l!}\,\frac{n_{[2l]}}{(n-1)^{2l}},$$

where $n_{[k]}=n(n-1)\cdots(n-k+1), n_{[0]}=1$. Asymptotically, $\tilde{C}_2^*(n)\Rightarrow \mathrm{Po}(1/2)$, so that the probability that no-one screams $\to 1-e^{-1/2}\approx 0.393$ as $n\to\infty$.

Simulation of 10^6 mappings of size $n=50$ provides estimates of the probability that all the component sizes are distinct of 0.976, that all the cycle lengths in the core are distinct of 0.839, and that all components lengths are distinct and all cycle lengths are distinct of 0.829. The asymptotic values of the first two probabilities are, from (21):

$$\prod_{j=2}^{\infty}\left(1+\frac{\lambda_j}{j}\right)e^{-\lambda_j/j}\approx 0.982 \quad\text{and}\quad \prod_{j=2}^{\infty}\left(1+\frac{1}{j}\right)e^{-1/j}=e^{1-\gamma}/2\approx 0.763, \tag{22}$$

respectively, using the Poisson approximation heuristic described below.

7.3. *Poisson approximation*

Asymptotic estimates such as those in (22) can be guessed by using the fact that for logarithmic assemblies, the analogues of (12) and (13) hold. Consider, for example, the difference D_n between the number of components and the number of distinct components lengths:

$$D_n=\sum_{j=1}^{n}(C_j(n)-1)_+,$$

where $(x)_+=\max(0,x)$. The probability of no repeated component sizes is then $\mathbb{P}(D_n=0)$.

For a typical logarithmic assembly, the Poisson approximation heuristic suggests that

$$D_n\Rightarrow D:=\sum_{j\geqslant 1}(Z_j-1)_+\quad\text{as } n\to\infty,$$

where the Z_j are independent Poisson random variables with means identified in (16). The proof follows from a truncation argument [3], and use of the total variation bound in (13). The left-most quantity in (22) is then just $\mathbb{P}(D=0)=\prod_{j\geqslant 2}\mathbb{P}(Z_j\leqslant 1)$.

8. *Prime factorisation*

There are intriguing connections between the probabilistic structure of combinatorial assemblies and the factorisation of integers into products of primes, outlined, for example, in [4, 6]. By way of example, Rényi [47] studied the probability that an integer chosen at random from $[n]$ is square-free, the natural analogue of the probability that a logarithmic assembly has no repeated component sizes. The strategy illustrated in the previous section suggests the approach here.

Writing N_n for the random integer with distribution

$$\mathbb{P}(N_n=j)=\frac{1}{n},\quad j=1,2,\ldots,n,$$

the prime factorisation of N_n is

$$N_n=\prod_p p^{C_p(n)},$$

where $C_p(n)=0$ if $p>n$. Analogous to (12), we have

$$(C_2(n),C_3(n),\ldots)\Rightarrow(Z_2,Z_3,\ldots)\quad\text{as } n\to\infty,$$

where the Z_p are independent geometrically distributed random variables satisfying

$$\mathbb{P}(Z_p = k) = \left(1 - \frac{1}{p}\right)\left(\frac{1}{p}\right)^k, \quad k = 0, 1, \ldots; \tag{23}$$

here $\mathbb{E}Z_p = 1/(p-1)$. The bound (13) is replaced by Kubilius's fundamental lemma [40], which shows that

$$d_{TV}(\mathcal{L}(C_p(n), p \leqslant b), \mathcal{L}(Z_p, p \leqslant b)) = O(e^{-cu}) \text{ for some } c > 0,$$

where $u = \log n / \log b$; see also [50]. Thus $d_{TV} \to 0$ if $\log b / \log n \to 0$.

It now follows that the difference D_n between the number of prime factors, with and without multiplicity, satisfies

$$D_n = \sum_p (C_p(n) - 1)_+ \Rightarrow D = \sum_p (Z_p - 1)_+ \quad \text{as } n \to \infty,$$

so that the probability the N_n is square-free is, asymptotically,

$$\mathbb{P}(D = 0) = \prod_p \mathbb{P}(Z_p \leqslant 1) = \prod_p \left(1 - \frac{1}{p}\right)\left(1 + \frac{1}{p}\right) = \prod_p \left(1 - \frac{1}{p^2}\right) = \frac{1}{\zeta(2)} = \frac{6}{\pi^2} \approx 0.608,$$

giving another proof of Rényi's result. By way of comparison, the probability that a random permutation has no repeated cycle lengths [26] is asymptotically $e^{-\gamma} \approx 0.562$.

This example has concentrated on the small prime factors, but the large prime factors are also of interest. Writing

$$N_n = P_1(n) P_2(n) \cdots,$$

for $P_1(n) \geqslant P_2(n) \geqslant \cdots$ prime or 1, Billingsley's theorem [10] shows that

$$(\log n)^{-1}(\log P_1(n), \log P_2(n), \ldots) \Rightarrow (L_1, L_2, \ldots), \tag{24}$$

where $(L_1, L_2, \ldots)$ has the Poisson–Dirichlet law with parameter $\theta = 1$. [2] establishes that there are couplings for which

$$\mathbb{E} \sum |\log P_i(n) - (\log n) L_i| = O(\log \log n).$$

Billingsley's theorem [10] has a complicated proof that identifies the joint distribution function of the limit random vector $(L_1, L_2, \ldots, L_r)$ for $r \geqslant 1$, and subsequent authors have sought to make the proof more transparent. For example, Donnelly and Grimmett [21] identified Billingsley's densities as being those of the marginals of PD(1), and they used a size-biassing argument and the connection between GEM(1) and PD(1) to provide another proof. Further discussion, and a simple, direct proof appear in [8].

Kingman [39] and Lloyd [42] started from a slightly different probability model, one in which the integer N is chosen from the zeta distribution with parameter $s > 1$:

$$\mathbb{P}(N = n) = \zeta(s)^{-1} n^{-s}, \quad n = 1, 2, \ldots.$$

Writing $n = \prod_p p^{c_p(n)}$, it follows that

$$\mathbb{P}(N = n) = \prod_p \left(1 - \frac{1}{p^s}\right)\left(\frac{1}{p^s}\right)^{c_p(n)},$$

showing that N may be constructed from independent, geometrically distributed random variables $C_p(n)$ having the distribution (23) with p replaced by p^s. Kingman [39] and Lloyd [42] showed that $(\log n)^{-1}(\log P_1, \log P_2, \ldots) \Rightarrow \text{PD}(1)$ as $s \to 1$.

The paper concludes with some speculations about probabilistic thinking, described in Figure 10.

© Master and Fellows of Trinity College, Cambridge

On thinking probabilistically . . .

The number-theoretic result needed for the proofs of (the two versions of) Billingsley's Theorem is Mertens Theorem [**43**], which states that

$$\sum_{p\leq x}\frac{1}{p} = \log\log x + M + O(1/\log x),$$

where M is the Meissel-Mertens constant.

It is tempting to believe that Hardy and Littlewood, number theoretic masters, would have discovered Billingsley's result had they been more persuaded by probabilistic reasoning. It was to be some time before there was a real synergy between pure mathematics and probability theory in Cambridge, due to the arrival of DG Kendall in 1962.

FIGURE 10. *G. H. Hardy, 32nd president* (1926–1928), *39th president* (1939–1941)*; J. E. Littlewood, 40th president* (1941–1943).

Acknowledgements. I thank Professors Richard Arratia and Andrew Barbour for their stimulating and productive collaboration over the last 30 years on topics relating to this work. In particular, many of the then-new results presented in the presidential address are from [**3**], co-authored with Warren Ewens. I thank Professors Peter Cameron, Charles Goldie, Frank Kelly, John Kingman, Andy Lynch and Ian Strachan for helpful discussions on the topics in this presentation. Any obscurities and errors in the paper remain my fault. I thank Elizabeth Fisher of the LMS, who collected the presidential photos used in the paper. Figures 1 and 7 are reproduced courtesy of the LMS. Figures 2 and 4 are reproduced courtesy of the Royal Society. Figure 5 is reproduced courtesy of the University of Bristol. Figure 10 is reproduced courtesy of the Master and Fellows of Trinity College, Cambridge. Finally, my thanks go to Drs. Poly da Silva and Arash Jamshidpey for many helpful comments on earlier drafts of this article.

References

1. D. J. Aldous, 'Exchangeability and related topics', *Ecole d'été de probabilités de Saint-Flour XIII*, Lecture Notes in Mathematics 1117 (Springer, Berlin, 1985) 1–198.

2. R. Arratia, 'On the amount of dependence in the prime factorization of a uniform random integer', *Contemporary combinatorics*. Bolyai Society Mathematical Studies 10 (János Bolyai Mathematical Society, Budapest, 2002) 29–91.

3. R. Arratia, A. Barbour, W. Ewens and S. Tavaré, 'Simulating the component counts of combinatorial structures', *Theoret. Popul. Biol.* 122 (2018) 5–11.

4. R. Arratia, A. Barbour and S. Tavaré, *Logarithmic combinatorial structures: a probabilistic approach* (European Mathematical Society, Zurich, 2003).

5. R. Arratia, A. D. Barbour and S. Tavaré, 'Poisson process approximations for the Ewens sampling formula', *Ann. Appl. Probab.* 2 (1992) 519–535.

6. R. Arratia, A. D. Barbour and S. Tavaré, 'Random combinatorial structures and prime factorizations', *Notices Amer. Math. Soc.* 44 (1997) 903–910.

7. R. Arratia, A. D. Barbour and S. Tavaré, 'A tale of three couplings: Poisson-Dirichlet and GEM approximations for random permutations', *Combin. Probab. Comput.* 15 (2006) 31–62.

8. R. Arratia and F. Kochman, 'A simple direct proof of Billingsley's theorem', Preprint, 2014, arXiv:1401.1553.

9. R. Arratia and S. Tavaré, 'Limit theorems for combinatorial structures via discrete process approximations', *Rand. Struct. Alg.* 3 (1992) 321–345.

10. P. Billingsley, 'On the distribution of large prime divisors', *Period. Math. Hungar.* 2 (1972) 283–289.

11. C. W. Borchardt, 'Über eine Interpolationsformel für eine Art Symmetrischer Functionen und über Deren Anwendung', *Math. Abh. der Akademie der Wissenschaften zu Berlin* (1860) 1–20.

12. W. Burnside, *Theory of probability* (Cambridge University Press, London, 1928).

13. P. J. Cameron, *Notes on counting: an introduction to enumerative combinatorics* (Cambridge University Press, Cambridge, 2017).

14. A. Cayley, 'A theorem on trees', *Quart. J. Pure Appl. Math.* 23 (1889) 376–378.

15. B. Charlesworth and D. Charlesworth, 'Population genetics from 1966 to 2016', *Heredity* 118 (2017) 2–9.

16. H. Crane, 'The ubiquitous Ewens sampling formula', *Statist. Sci.* 31 (2016) 1–19.

17. P. H. da Silva, A. Jamshidpey, P. McCullagh and S. Tavaré, 'Fisher's measure of variability in repeated samples', submitted.

18. P. H. da Silva, A. Jamshidpey and S. Tavaré, 'The Feller Coupling for random derangements', *Stochastic Process. Appl.* (2022), to appear.

19. A. De Morgan, *An essay on probabilities, and on their application to life contingencies and insurance offices* (Longman et al., London, 1838).

20. P. Donnelly, 'Partition structures, Pólya urns, the Ewens sampling formula, and the ages of alleles', *Theoret. Popul. Biol.* 30 (1986) 271–288.

21. P. Donnelly and G. Grimmett, 'On the asymptotic distribution of large prime factors', *J. Lond. Math. Soc.* 47 (1993) 395–404.

22. W. J. Ewens, 'The sampling theory of selectively neutral alleles', *Theoret. Popul. Biol.* 3 (1972) 87–112.

23. R. A. Fisher, A. S. Corbet and C. B. Williams, 'The relation between the number of species and the number of individuals in a random sample from an animal population', *J. Animal Ecology* 12 (1943) 42–58.

24. M. Gardner, *The sixth book of mathematical puzzles and diversions from scientific American* (W.H. Freeman, New York, 1971).

25. V. L. Goncharov, 'Some facts from combinatorics', *Izvestia Akad. Nauk. SSSR, Ser. Mat.* 8 (1944) 3–48; see also: On the field of combinatory analysis. *Trans. Amer. Math. Soc.* 19, 1–46.

26. D. H. Greene and D. E. Knuth, *Mathematics for the analysis of algorithms*, 2nd edn (Birkhaüser, Boston, MA, 1982).

27. G. James and A. Kerber, *The representation theory of the symmetric group*, Encyclopedia of Mathematics and Its Applications 16 (Addison-Wesley, Reading, MA, 1981).

28. S. Karlin and J. McGregor, 'The number of mutant forms maintained in a population', *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (ed. L. LeCam and J. Neyman; University of California Press, Berkeley, CA, 1967) 415–438.

29. D. G. Kendall, 'On the generalized "birth-and-death" process', *Ann. Math. Statist.* 19 (1948) 1–15.

30. D. G. Kendall, 'On some modes of population growth leading to R. A. Fisher's logarithmic series distribution', *Biometrika* 35 (1948) 6–15.

31. D. G. Kendall, 'Stochastic processes and population growth', *J. R. Stat. Soc. Ser. B* 11 (1949) 230–264.

32. D. G. Kendall, 'Branching processes since 1873', *J. Lond. Math. Soc.* 41 (1966) 385–406.

33. D. G. Kendall, 'The genealogy of genealogy. Branching processes before (and after) 1873', *Bull. Lond. Math. Soc.* 7 (1975) 225–253.

34. D. G. Kendall, 'Some problems in mathematical genealogy', *Perspectives in Probability and Statistics: Papers in Honour of M.S. Bartlett* (ed. J. GANI; Academic Press, London, 1975) 325–345.

35. J. F. C. Kingman, 'Random discrete distributions', *J. R. Stat. Soc.* 37 (1975) 1–22.

36. J. F. C. Kingman, 'Exchangeability and the evolution of large populations', *Exchangeability in probability and statistics* (ed. G. Koch and F. Spizzichino; North-Holland, Amsterdam, 1982) 97–112.

37. J. F. C. Kingman, 'On the genealogy of large populations', *J. Appl. Prob.* 19A (1982) 27–43.

38. J. F. C. Kingman, 'The coalescent', *Stoch. Process. Appl.* 13 (1982) 235–248.

39. J. F. C. Kingman, 'The Poisson-Dirichlet distribution and the frequency of large prime divisors', *Technical report* (Isaac Newton Institute for Mathematical Sciences, Cambridge, 2004).

40. J. Kubilius, *Probabilistic methods in the theory of numbers*, AMS Translations of Mathematical Monographs 11 (American Mathematical Society, Providence, RI, 1964; Russian edition 1962).

41. A. Lambert and T. Stadler, 'Birth–death models and coalescent point processes: the shape and probability of reconstructed phylogenies', *Theoret. Popul. Biol.* 90 (2013) 113–128.

42. S. P. Lloyd, 'Ordered prime divisors of a random integer', *Ann. Probab.* 12 (1984) 1205–1212.

43. F. Mertens, 'Ein Beitrag zur analytischen Zahlentheorie', *J. reine angew. Math.* 78 (1874) 46–62.

44. J. W. Pitman, 'Exchangeable and partially exchangeable random partitions', *Prob. Theory Related Fields* 102 (1995) 145–158.

45. J. W. Pitman, 'Coalescent random forests', *J. Combin. Theory, Ser. A* 85 (1999) 165–193.

46. J. W. Pitman, *Combinatorial stochastic processes*, Lecture Notes in Mathematics 1875 (Springer, Berlin, 2006).

47. A. Rényi, 'On the density of certain sequences of integers', *Publ. Inst. Math. Acad. Serbe Sci.* 8 (1955) 157–162.

48. S. Tavaré, 'The birth process with immigration, and the genealogical structure of large populations', *J. Math. Biol.* 25 (1987) 161–168.

49. S. Tavaré, 'A note on the Screaming Toes game', *J. Appl. Prob.* 59 (2022), to appear.

50. G. Tenenbaum, 'Crible d'Ératosthène et modèle de Kubilius', *Number theory in progress*, vol. 2 (Zakopane-Kościelisko, 1997; de Gruyter, Berlin, 1999) 1099–1129.

51. G. A. Watterson, 'The sampling theory of selectively neutral alleles', *Adv. Appl. Prob.* 6 (1974) 463–488.

52. G. A. Watterson, 'Models for the logarithmic species abundance distributions', *Theor. Popul. Biol.* 6 (1974) 217–250.

53. G. A. Watterson, ‘The homozygosity test of neutrality’, *Genetics* 88 (1978) 405–417.
54. P. Winkler, *Mathematical mind-benders* (A. K, Peters, Wellesley, MA, 2007).
55. S. Wright, *Evolution and the genetics of populations*, Variability Within and Among Natural Populations 4 (University of Chicago Press, Chicago, IL, 1978).
56. S. L. Zabell, ‘The rule of succession’, *Erkenntnis* 31 (1989) 283–321.
57. S. L. Zabell, ‘The continuum of inductive methods revisited’, *The cosmos of science: essays of exploration* (ed. J. Earman and J. D. Norton; University of Pittsburg Press, Pittsburg, PA, 1997) 351–385.

Simon Tavaré
Department of Applied Mathematics and Theoretical Physics
University of Cambridge, Centre for Mathematical Sciences
Wilberforce Road
Cambridge, CB3 0WA
United Kingdom

and

Herbert and Florence Irving Institute for Cancer Dynamics
and Department of Statistics
Columbia University
1255 Amsterdam Avenue
New York, NY 10027
USA

st3193@columbia.edu