

A Bose–Einstein approach to the random partitioning of an integer

This content has been downloaded from IOPscience. Please scroll down to see the full text.

J. Stat. Mech. (2011) P08021

(<http://iopscience.iop.org/1742-5468/2011/08/P08021>)

View [the table of contents for this issue](#), or go to the [journal homepage](#) for more

Download details:

This content was downloaded by: count0

IP Address: 134.102.186.160

This content was downloaded on 09/06/2016 at 12:03

Please note that [terms and conditions apply](#).

A Bose–Einstein approach to the random partitioning of an integer

Thierry E Huillet

Laboratoire de Physique Théorique et Modélisation, CNRS-UMR 8089 et
Université de Cergy-Pontoise, 2 Avenue Adolphe Chauvin, F-95302,
Cergy-Pontoise, France
E-mail: Thierry.Huillet@u-cergy.fr

Received 13 June 2011

Accepted 26 July 2011

Published 30 August 2011

Online at stacks.iop.org/JSTAT/2011/P08021

[doi:10.1088/1742-5468/2011/08/P08021](https://doi.org/10.1088/1742-5468/2011/08/P08021)

Abstract. Consider N equally spaced points on a circle of circumference N . Pick at random n points out of N on this circle and consider the discrete random spacings between consecutive sampled points, turning clockwise. This defines in the first place a random partitioning of N into n positive summands. Append then clockwise an arc of integral length k to each such sampled point, ending up with a discrete random set on the circle. Questions such as the evaluation of the probability of random covering or parking configurations, number and length of the gaps are addressed. For each value of k , asymptotic results are presented when n, N both go to ∞ according to two different regimes. In the first thermodynamical regime $n/N \rightarrow \rho$, the occurrence of, say, covering and parking configurations is exponentially rare in the whole admissible range of density ρ . We compute the rates from the equations of state. In the second one, they are macroscopically frequent. These questions require some understanding of both the smallest and largest extreme summands in the partition of N .

We consider next an urn model where N indistinguishable balls are assigned at random into N distinguishable boxes. This urn model consists of a random partitioning model of integer N into N non-negative summands. Given there are n non-empty boxes this gives back the original partitioning model of N into n positive parts. Following this circle of ideas, a grand canonical balls in boxes approach is supplied, giving some insight into the multiplicities of the box occupancies.

The random set model defines a k -nearest neighbor random graph with N vertices and kn edges. We shall also briefly consider the covering problem in the context of a random graph model with N vertices and n (out-degree 1) edges whose endpoints are no longer bound to be neighbors. In the latter setup,

connectivity is increased in that there exists a critical density ρ_c above which covering occurs with probability 1.

Keywords: rigorous results in statistical mechanics, stochastic processes (theory), random graphs, networks

ArXiv ePrint: [1106.2075](https://arxiv.org/abs/1106.2075)

Contents

1. Introduction	2
2. Random partition of an integer and discrete spacings	3
3. N-circle covering problems	6
4. Large deviation rate functions in the thermodynamical limit: hard rods, covering and parking configurations	8
4.1. Hard rods	9
4.2. Covering configurations	10
4.3. Parking configurations	12
5. The grand canonical partition of N	12
6. Random graph connectivity	16
References	18

1. Introduction

Many authors have considered the problems related to the coverage of the unit circle by arcs of equal sizes randomly placed on the circle, among which [22, 20, 6, 7, 19, 8, 9, 11]. In this paper, motivated by a remark in the paper ([2], p 18) on random graphs, we shall be concerned with a discrete version of the above problem, following [10] and [14].

Consider N equally spaced points (vertices) on the circle of circumference N , so with arc length 1 between consecutive points. Sample at random n out of these N points and consider the discrete random spacings between consecutive sampled points, turning clockwise on the circle. This defines a random partitioning of N into n positive summands.

Let then k be an integer and append clockwise an arc of length k to each sampled point, forming a random set of arcs on the circle. What is the probability that the circle is covered? If the circle is not covered, how many gaps do we have in the random set of arcs? What is the probability that no arcs overlap (the discrete hard rods model), what is the probability that no arcs overlap and that the gap lengths are smaller than k itself (the discrete version of Rényi's parking model). All these questions require some understanding of both the smallest and largest spacings in the sample (or equivalently the smallest and largest summands in the random partitioning point of view). They are the discrete versions of similar problems raised in the continuum. We will focus on the limiting thermodynamical regime, $n, N \rightarrow \infty$ while $n/N \rightarrow \rho$, and also, sometimes, on a regime where $n, N \rightarrow \infty$ while $n(1 - n/N)^k \rightarrow \alpha$, $0 < \alpha < \infty$. In the first regime, the occurrence

of say covering and parking configurations is exponentially rare in the whole admissible density range of ρ , whereas in the second one they are macroscopically frequent. At the heart of this model is the Bose–Einstein distribution for discrete spacings.

Finally, a Bosonic grand canonical approach to the above model will be considered where N indistinguishable balls are assigned at random into N distinguishable boxes. As it is defined, this urn model consists of a random partitioning model of integer N into N non-negative summands. Conditioning this balls in boxes model on having exactly n positive summands (or non-empty boxes) yields back the original random partitioning model of N into n positive summands. For this urn model, we will study the number of empty boxes and the number of boxes with i balls, giving some insight into the box occupancy multiplicities, in both the canonical and the grand canonical ensembles.

The model just developed may as well be viewed as a k -nearest neighbors random graph with N vertices and kn edges. In relation to this, in section 6 we shall consider a random graph model with N vertices and n (out-degree 1) edges whose endpoints are no longer necessarily neighbors, being now chosen at random on the whole set of vertices. In this model of a different kind, each of the n sampled points is allowed to create a link far away with any of the N vertices, not necessarily with neighbors. We estimate the covering probability for this random graph model in the spirit of Erdős–Rényi (see [1]). We show that, in sharp contrast to the k -nearest neighbor graph, there exists a critical density $\rho_c = 1 - e^{-1}$ above which covering occurs with probability one. The take-home message is to what extent when connections are not restricted to neighbors the chance of connectedness is increased.

To summarize: sampling at random n points on the circle out of N is related to the following.

- A random partitioning model of N into n positive summands.
- A discrete geometrical random set problem on the circle¹.
- A balls in boxes urn model and a random partitioning model of N into N non-negative summands.
- A k -nearest neighbors random graph model.

2. Random partition of an integer and discrete spacings

Consider a circle of circumference N , with N integer. Consider N equally spaced points on the circle so with arc length 1 between consecutive points. We shall call this discrete set of points the N -circle. Draw at random $n \in \{2, \dots, N-1\}$ points without replacement at the integer sites of this circle (thus, with M_1, \dots, M_n independent and identically distributed, say iid, and uniform on $\{1, \dots, N\}$). Pick at random one of the points M_1, \dots, M_n and call it $M_{1:n}$. Next, consider the ordered set of integer points $(M_{m:n}, m = 1, \dots, n)$, turning clockwise on the circle, starting from $M_{1:n}$. Let $N_{m,n} = M_{m+1:n} - M_{m:n}$, $m = 1, \dots, n-1$, be the consecutive discrete spacings, with $N_{n,n} = M_{1:n} - M_{n:n}$, modulo N , closing the

¹ In relation to this, it seems to be an open problem as to whether or not one can obtain some precise statistical information on the joint law of the sizes of the connected components. This program could be partly achieved (in [11]) for the continuous random partitioning of the unit circle problem but it is not clear whether similar tools can be used in the discrete setup.

loop. Under our hypothesis $N_{m,n} \stackrel{d}{=} N_n$, $m = 1, \dots, n$, independent of m , the distribution of which is $\bar{F}_{N_n}(k) := \mathbf{P}(N_n > k) = 1 - F_{N_n}(k) = \binom{N-k-1}{n-1} / \binom{N-1}{n-1}$, with $\mathbf{E}N_n = N/n$.

It is indeed a result of considerable age (see e.g. [10]) that identically distributed (id) discrete spacings $\mathbf{N}_n := (N_{m,n}; m = 1, \dots, n)$ with $|\mathbf{N}_n| := \sum_m N_{m,n} = N$ can be generated as the conditioning

$$\mathbf{N}_n = \mathbf{G}_n \mid \{|\mathbf{G}_n| = N\}, \quad (1)$$

where $|\mathbf{G}_n| := \sum_{m=1}^n G_m$ is the sum of n iid geometric (α) random variables ≥ 1 (with $\mathbf{P}(G_1 \geq k) = \alpha^{k-1}$, $k \geq 1$, $\alpha \in (0, 1)$), and so N_n has the claimed Pólya–Eggenberger $PE(1, n-1)$ distribution: $\mathbf{P}(N_n = k) = \binom{N-k-1}{n-2} / \binom{N-1}{n-1}$, $k = 1, \dots, N-n+1$.

Note that as $n, N \rightarrow \infty$, while $n/N = \rho < 1$ is fixed, using the Stirling formula we get the convergence in distribution

$$N_n \xrightarrow{d} G, \quad (2)$$

where $G \geq 1$ is a discrete random variable (rv) with geometric $(1-\rho)$ distribution: $\mathbf{P}(G \geq m) = (1-\rho)^{m-1}$, $m \geq 1$. The limiting expected value of N_n is $1/\rho$.

With $\mathbf{k} := (k_m; m = 1, \dots, n)$, the joint law of \mathbf{N}_n is

$$\mathbf{P}(\mathbf{N}_n = \mathbf{k}) = \frac{1}{\binom{N-1}{n-1}} 1(|\mathbf{k}| = N), \quad (3)$$

which is the exchangeable uniform distribution on the restricted discrete N -simplex $|\mathbf{k}| := \sum_{m=1}^n k_m = N$, $k_m \geq 1$, also known as the Bose–Einstein distribution. This distribution occurs in the following Pólya–Eggenberger urn model context (see [15]). An urn contains n balls all of different colors. A ball is drawn at random and replaced together with adding another ball of the same color. Repeating this $N-n$ times, \mathbf{N}_n is the number of balls of different colors in the urn. See [10].

From the random model just defined we get

$$N = \sum_{m=1}^n N_{m,n}, \quad (4)$$

which corresponds to a random partition of N into n id parts or components ≥ 1 .

It also models the following random allocation problem (see [16]). N items are to be shared at random between n recipients. $N_{m,n}$ is the share of the N items allocated to recipient m . Although all shares are id, there is a great variability in the recipients' parts as will become clear from the detailed study of the smallest and largest shares in the sample.

This model is connected to the continuous spacings between n randomly placed points on the unit circle in the following way. As $N \rightarrow \infty$, $\mathbf{N}_n/N \xrightarrow{d} \mathbf{S}_n$, where $\mathbf{S}_n := (S_{1,n}, \dots, S_{n,n})$ has Dirichlet uniform density function on the continuous unit n -simplex [17]

$$f_{S_{1,n}, \dots, S_{n,n}}(s_1, \dots, s_n) = (n-1)! \cdot \delta(\sum_{m=1}^n s_m - 1). \quad (5)$$

Let $P_n(1) := \sum_{m=1}^n 1(N_{m,n} > 1)$ be the number of sampled points whose distance to their clockwise neighbors is more than one unit. There are $n - P_n(1)$ sampled points which are neighbors, therefore

$$\begin{aligned} N &= 1 \cdot (n - P_n(1)) + \sum_{m=1}^n N_{m,n} 1(N_{m,n} > 1) \\ &= n + \sum_{m=1}^n (N_{m,n} - 1)_+, \end{aligned}$$

where $i_+ = \max(i, 0)$. Appending an arc of length 1 clockwise to the n sampled points and considering the induced covered set from $\{1, \dots, N\}$, $\bar{\mathcal{L}}_n(1) := \sum_{m=1}^n (N_{m,n} - 1)_+$ represents the length of the gaps (the size of the uncovered set). So, from the model $\bar{\mathcal{L}}_n(1) = N - n$ is a constant and

$$N - n = \sum_{m=1}^n (N_{m,n} - 1)_+$$

corresponds to a random partition of $N - n$ into n id parts or components ≥ 0 . Stated differently, the length of the covered set $\mathcal{L}_n(1) = N - \bar{\mathcal{L}}_n(1)$ is constant, equal to n , which is obvious.

Of considerable interest is the sequence $(N_{m:n}; m = 1, \dots, n)$ obtained while ordering the components' sizes $(N_{m,n}; m = 1, \dots, n)$, with $N_{1:n} \leq \dots \leq N_{n:n}$.

By the exclusion–inclusion principle, the cumulative distribution function $F_{N_{m:n}}(k) = \mathbf{P}(N_{m:n} \leq k)$ is easily seen to be

$$F_{N_{m:n}}(k) = \frac{1}{\binom{N-1}{n-1}} \sum_{q=m}^n \binom{n}{q} \sum_{p=n-q}^n (-1)^{p+q-n} \binom{q}{n-p} \binom{N-pk-1}{n-1}, \quad (6)$$

which has been known for a while in the context of spacings in the continuum (see [22]).

In particular,

$$F_{N_{n:n}}(k) := \mathbf{P}(N_{n:n} \leq k) = \frac{1}{\binom{N-1}{n-1}} \sum_{p=0}^n (-1)^p \binom{n}{p} \binom{N-pk-1}{n-1} \quad (7)$$

and

$$\bar{F}_{N_{1:n}}(k) := \mathbf{P}(N_{1:n} > k) = \binom{N-nk-1}{n-1} / \binom{N-1}{n-1} \quad (8)$$

are the largest and smallest component size distributions in this case.

In the formula giving $F_{N_{n:n}}(k)$, with $[x]$ standing for the integral part of x , the sum should as well stop at $n \wedge [(N - n)/k]$, observing $\binom{i}{j} = 0$ if $i < j$.

Clearly, if $k = 1$, $\mathbf{P}(N_{n:n} = 1) = 0$ ($=1$) whatever $n < N$ (if $n = N$). If $k = 2$ and $N > 2n$, $\mathbf{P}(N_{n:n} \leq 2) = \mathbf{P}(N_{n:n} = 2) = 0$. If $N = 2n$, $\mathbf{P}(N_{n:n} = 2) = 1/\binom{2n-1}{n-1}$ is the probability of a regular configuration with all sampled points equally spaced by two arc length units. If $n < N < 2n$, $\mathbf{P}(N_{n:n} = 2)$ is the probability of a configuration with $2n - N$ neighbor points distant by one arc length unit and $N - n$ points distant by two units.

As $N, k \rightarrow \infty$ while $k/N \rightarrow s$

$$\binom{N - pk - 1}{n - 1} / \binom{N - 1}{n - 1} \rightarrow (1 - ps)_+^{n-1}.$$

With $0 < a < b \leq N$, the joint law of $(N_{1:n}, N_{n:n})$ is given by

$$\mathbf{P}(N_{1:n} > a, N_{n:n} \leq b) = \sum_{m=0}^n \frac{(-1)^m}{\binom{N-1}{n-1}} \binom{n}{m} \binom{N - (na + m(b-a)) - 1}{n-1}. \quad (9)$$

In the random partitioning of N 's image, it gives the probability that the shares of all n recipients all range between a and b . Putting $(a = k, b = N)$ and $(a = 0, b = k)$ gives $F_{N_{n:n}}(k)$ and $\bar{F}_{N_{1:n}}(k)$. This formula was first obtained by [3] in the continuum. Putting next $a = k, b = 2k$, we get

$$\mathbf{P}(N_{1:n} > k, N_{n:n} \leq 2k) = \frac{1}{\binom{N-1}{n-1}} \sum_{m=0}^n (-1)^m \binom{n}{m} \binom{N - (n+m)k - 1}{n-1}. \quad (10)$$

When $k = 1$, we have $\mathbf{P}(N_{1:n} > 1, N_{n:n} \leq 2) = \binom{N-1}{n-1}^{-1} 1_{N=2n}$. If $N = 2n$, $\binom{2n-1}{n-1}^{-1}$ is the probability of the configuration where the n sampled points are exactly equally spaced, each by two arc length units.

As $n, N \rightarrow \infty$ while $n/N \rightarrow \rho < 1$, with E an rv with rate 1 exponential distribution,

$$-\log(1 - \rho) n (N_{1:n} - 1) \xrightarrow{d} E; \frac{\rho}{\log n} N_{n:n} \xrightarrow{a.s.} 1, \quad (11)$$

suggesting that the smaller (larger) integer component in the partition of N is of order n^{-1} (respectively $\log n$) in the considered asymptotic regime. More precisely, using the joint law of $(N_{1:n}, N_{n:n})$,

$$\left(-\log(1 - \rho) n (N_{1:n} - 1), N_{n:n} - \frac{\log n}{\rho} \right) \xrightarrow{d} (E, G), \quad (12)$$

where (E, G) are independent rvs on $\mathbb{R}_+ \times \mathbb{R}$, with distributions $\mathbf{P}(E > t) = e^{-t}$ and $\mathbf{P}(G \leq t) = e^{-e^{-t}}$ with $\mathbf{E}(G) = \gamma$ the Euler constant (exponential and Gumbel).

Although in the random partitioning of N , all parts attributed to each recipient are id, there is a great variability in the shares as the smallest one is of order 1 and the largest one of order $\log n$.

3. N -circle covering problems

Let $\mathcal{S}_n := \{M_1, \dots, M_n\}$ be the discrete set of points drawn at random on the N -circle with circumference N . Fix $k \in \{1, \dots, N\}$. Consider the coarse-grained discrete random set of intervals

$$\mathcal{S}_n(k) := \{M_1 + l, \dots, M_n + l, 1 \leq l \leq k\} \quad (13)$$

appending clockwise an arc of integral length $k \geq 1$ to each starting-point atom of \mathcal{S}_n .

The number of gaps and the length of the covered set. Let $P_n(k)$ be the number of gaps of $\mathcal{S}_n(k)$ (which is also the number of connected components), so with $P_n(k) = 0$ as soon as the N -circle is covered by $\mathcal{S}_n(k)$.

Let also $\mathcal{L}_n(k)$ be the total integral length of $\mathcal{S}_n(k)$. As there are $n - P_n(k)$ spacings covered by k and $P_n(k)$ gaps each contributing k to the covered length, it can be expressed as a contribution of two terms ($i \wedge j = \min(i, j)$),

$$\mathcal{L}_n(k) = \sum_{m=1}^{n-P_n(k)} N_{m:n} + kP_n(k) = \sum_{m=1}^n (N_{m,n} \wedge k). \quad (14)$$

Note also that the vacancy, which is the length of the N -circle not covered by any arc, is

$$\bar{\mathcal{L}}_n(k) := N - \mathcal{L}_n(k) = \sum_{p=1}^{P_n(k)} (N_{n-p+1:n} - k) = \sum_{m=1}^n (N_{m,n} - k)_+, \quad (15)$$

summing the gaps' lengths over the gaps (with $N_{n:n} - k$ the largest gap size and $N_{n-P_n(k)+1:n} - k$ the smallest gap size). We recover the result (i) originally due to [20] and its asymptotic consequences. The following statements are mainly due to Holst, see [10]. It holds that

(i) The distribution of $P_n(k)$ is

$$\mathbf{P}(P_n(k) = p) = \frac{\binom{n}{p}}{\binom{N-1}{n-1}} \sum_{m=p}^n (-1)^{m-p} \binom{n-p}{m-p} \binom{N-mk-1}{n-1}. \quad (16)$$

(ii) As $n, N \rightarrow \infty$, while $n(1 - (n/N))^k \rightarrow \alpha$, $0 < \alpha < \infty$,

$$P_n(k) \rightarrow \text{Poi}(\alpha), \quad (17)$$

where $\text{Poi}(\alpha)$ is a random variable with Poisson distribution of parameter α .

(iii) (a) Number of gaps. As $n, N \rightarrow \infty$ while $n/N \rightarrow \rho$, with $0 < \rho < 1$,

$$\frac{1}{\sqrt{n}} \left(P_n(k) - n(n/N)^k \right) \xrightarrow[N \rightarrow \infty]{d} \mathcal{N}(0, \sigma^2 = \rho^k(1 - \rho^k)), \quad (18)$$

where $\mathcal{N}(m, \sigma^2)$ stands for the normal law with mean m and variance σ^2 .

(b) Gap length:

$$\frac{1}{\sqrt{n}} \left(\bar{\mathcal{L}}_n(k) - N(1 - n/N)^k \right) \xrightarrow[N \rightarrow \infty]{d} \mathcal{N}(0, \sigma^2), \quad (19)$$

where $\sigma^2 = (1 + \bar{\rho} - \bar{\rho}^k)\bar{\rho}^k - (\bar{\rho} + k\rho)^2\bar{\rho}^{2k-1}$, $\bar{\rho} = 1 - \rho$.

The proofs of (ii) and (iiib) are in [10]. The proof of (iiia) follows from similar central limit theorem arguments developed there. In the first case (ii), $n \sim N(1 - (\alpha/N)^{1/k})$ and so n is very close to N ; because of that, there are finitely many gaps in the limit and the covering probability is $e^{-\alpha}$, so macroscopic. However, in the second case (iii), $n \sim \rho N$ is quite small: the number of gaps is of order $n\rho^k$ and the covering probability is expected to be exponentially small. Note from (iiib) that the variance of the limiting normal law is 0 when $k = 1$, in accordance with the fact that $\bar{\mathcal{L}}_n(1) = N - n$ remains constant.

The number of arcs needed to cover the N -circle. In (16), $\mathbf{P}(P_n(k) = 0)$ is the cover probability and $\mathbf{P}(P_n(k) = n)$ the probability that no overlap of arcs or rods takes place (the hard rods model). We have $\mathbf{P}(P_n(k) = 0) = \mathbf{P}(N_{n:n} \leq k)$.

The cover probability $\mathbf{P}(P_n(k) = 0)$ is also the probability that the number of arcs of length k (the sample size), say $N(k)$ required to cover the N -circle is less than or equal to n . We have $N(k) = \inf(n : N_{n:n} \leq k)$. In other words, $\mathbf{P}(N(k) > n) = \mathbf{P}(N_{n:n} > k)$ and so $\mathbf{E}N(k) = \sum_{n=1}^N \mathbf{P}(N_{n:n} > k)$, with

$$\mathbf{P}(N_{n:n} > k) = \frac{1}{\binom{N-1}{n-1}} \sum_{m=1}^n (-1)^{m-1} \binom{n}{m} \binom{N-mk-1}{n-1}.$$

We wish to estimate $\mathbf{E}N(k)$ as N grows large.

When $n(1 - (n/N))^k \rightarrow \alpha$, so when $n \sim N(1 - (\alpha/N)^{1/k})$, we have $\mathbf{P}(P_n(k) = 0) = \mathbf{P}(N(k) \leq n) \rightarrow e^{-\alpha}$. Therefore, as $N \rightarrow \infty$,

$$N^{1/k} \left(1 - \frac{N(k)}{N}\right) \xrightarrow{d} E_k, \quad (20)$$

where E_k has a Weibull(k) distribution with $\mathbf{P}(E_k > x) = e^{-x^k}$ and $\mathbf{E}(E_k) = \Gamma(1 + k^{-1})$. Thus

$$\mathbf{E}N(k) \sim_{N \rightarrow \infty} N \left(1 - \frac{\Gamma(1 + k^{-1})}{N^{1/k}} + o(N^{-1/k})\right) \quad (21)$$

is the estimated expected number of length- k arcs required to cover the N -circle.

4. Large deviation rate functions in the thermodynamical limit: hard rods, covering and parking configurations

k -hard rods configurations are those for which $N_{1:n} > k \geq 1$ (the smallest part in the decomposition of N exceeds the arc length k : appending an arc of length k to all sampled points does not result in overlapping of the added arcs). k -covering configurations with $k > 1$ are those for which $N_{n:n} \leq k$ (the largest part in the decomposition of N is smaller than the arc length k : appending an arc of length k to each sampled point results in the covering of all points of the N -circle, a connectedness property). k -parking configurations are those for which both $N_{1:n} > k$ and $N_{n:n} \leq 2k$ (the smallest part in the decomposition of N exceeds the arc length k and the largest part in the decomposition of N is smaller than twice the arc length k : appending an arc of length k to all sampled points results in a hard rods configuration where sampled points are separated by gaps of length at least k but with the extra excess gaps being smaller than k , so with no way to add a new rod (or car) of size k without provoking an overlap). All these configurations are exponentially rare in the thermodynamic limit $n, N \rightarrow \infty$ while $n/N \rightarrow \rho \in (0, 1)$. We make this statement precise by computing the large deviation rate functions in each case, extending to the discrete formulation similar results obtained in the continuum, see [18, 12, 21, 4].

4.1. Hard rods

k -hard rods configurations are those for which $N_{1:n} > k > 1$. (In the partitioning approach of the fortune N amongst n recipients, this event is realized if the share of the poorest is bounded below by k , a rare event.) When the number of sampled points n is a fraction of N (the case with a density $n = \rho N$), there are too few sampled points for a non-overlapping configuration to occur with a reasonably large probability. Rather, one expects that the probability of non-overlapping (hard rods) configurations tends to zero exponentially fast. To see this, we need to evaluate the large n expansion of $\mathbf{P}(N_{1:n} > k)$. Note that the event $N_{1:n} > k$ is an event with positive probability if and only if $N \geq n(k+1)$ so, in the following, we shall assume that $\rho < 1/(k+1)$, $k \geq 1$. We have

$$\mathbf{P}(N_{1:n} > k) = \frac{Z_{n,N}}{\sum_{k_1, \dots, k_n \geq 1} \prod_{m=1}^n 1_{\sum k_m = N}} = \frac{Z_{n,N}}{\binom{N-1}{n-1}} \sim C \left(\frac{n}{N}\right)^n \left(1 - \frac{n}{N}\right)^{N-n} Z_{n,N},$$

where $Z_{n,N} = \sum_{k_1, \dots, k_n \geq 1} \prod_{m=1}^n 1_{k_m > k} 1_{\sum k_m = N}$. In the limit $n, N \rightarrow \infty$ with $n/N \rightarrow \rho$

$$-\frac{1}{n} \log \mathbf{P}(N_{1:n} > k) \rightarrow -\frac{1}{\rho} (\rho \log \rho + (1 - \rho) \log (1 - \rho)) + \lim_{n \rightarrow \infty} -\frac{1}{n} \log Z_{n,N}. \quad (22)$$

In the limit $n, N \rightarrow \infty$ with fixed n/N limit, the quantity $\mathbf{P}(N_{1:n} > k)$ is easier to evaluate in an isobaric ensemble where the pressure p is held fixed instead of $\sum k_m$. Therefore, relaxing the constraint $\sum k_m = N$, we shall work instead with the modified random variables $\tilde{N}_{m,n}$, with exponentially tilted law

$$\mathbf{P}(\tilde{N}_{m,n} = k_m > k, m = 1, \dots, n) = \frac{\prod_{m=1}^n 1_{k_m > k} e^{-pk_m}}{Z_{n,p}}.$$

Here

$$Z_{n,p} = \sum_{k_1, \dots, k_n \geq 1} \prod_{m=1}^n 1_{k_m > k} e^{-pk_m} = \left(\sum_{l > k} e^{-pl} \right)^n = \left(\frac{e^{-p(k+1)}}{1 - e^{-p}} \right)^n$$

is the normalizing constant.

Defining $G_{n,p} := -\log Z_{n,p}$, we have $\partial_p G_{n,p} = \mathbb{E}_{N,p}(\sum \tilde{N}_{m,n} 1_{\tilde{N}_{m,n} > k})$ and one must choose p in such a way that $\partial_p G_{n,p} = N$, leading to $N = n(k+1 + e^{-p}/(1 - e^{-p}))$ or $1/\rho = e^{-p}/(1 - e^{-p}) + k + 1$, so $p = -\log((1 - \rho(k+1))/(1 - \rho k))$. The latter equation relating p, ρ and k is an equation of state. Due to the equivalence of ensembles principle, see [4] for similar arguments, we have $Z_{n,N} = e^{pN} Z_{n,p} O(N^{-1/2})$, leading to $-(1/n) \log Z_{n,N} \sim -(1/n) \log Z_{n,p} - p/\rho$. Proceeding in this way, we finally get

$$\begin{aligned} -\frac{1}{n} \log \mathbf{P}(N_{1:n} > k) &\rightarrow F_{\text{hr}}(p, \rho) \\ &= -\frac{1}{\rho} (\rho \log \rho + (1 - \rho) \log (1 - \rho)) - \frac{p}{\rho} - \log \left(\frac{e^{-p(k+1)}}{1 - e^{-p}} \right), \end{aligned} \quad (23)$$

with $\rho \in (0, 1/(k+1))$. Here, thermodynamical ‘pressure’ $p > 0$ and density ρ are related through the ‘state equation’ $\partial_p F_{\text{hr}}(p, \rho) = 0$ which can consistently be checked to be

$$\frac{1}{\rho} = k + 1 + \frac{e^{-p}}{1 - e^{-p}}, \quad (24)$$

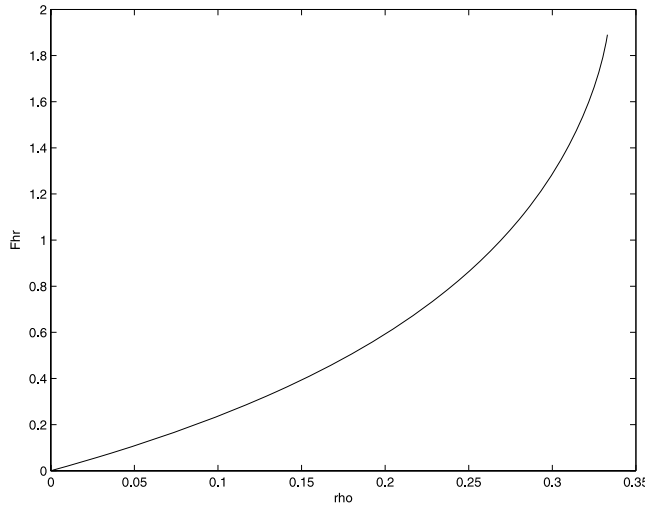


Figure 1. A plot of F_{hr} versus ρ when $k = 2$ or $\rho \in (0, 1/3)$.

leading to $p = -\log((1 - \rho(k + 1))/(1 - \rho k)) > 0$ (which is well defined and positive because $\rho < 1/(k + 1)$). Thus F_{hr} is an explicit entropy-like positive function of ρ and k , namely

$$F_{\text{hr}}(\rho) = -\frac{1}{\rho}((1 - \rho) \log(1 - \rho) - (1 - \rho(k + 1)) \times \log(1 - \rho(k + 1)) + (1 - \rho k) \log(1 - \rho k)). \quad (25)$$

In the thermodynamical limit, hard rods configurations are exceptional and the hard rods large deviation rate function F_{hr} is an explicit function of ρ and k . We conclude that with probability tending to 1, $N_{1:n} = 1$. In the partitioning approach of the fortune N amongst n recipients, the share of the poorer is the smallest possible.

As $\rho \uparrow 1/(k + 1)$, pressure tends to ∞ and $F_{\text{hr}}(\rho) \rightarrow (k + 1) \log(k + 1) - k \log(k) > 0$. As $\rho \downarrow 0$, pressure tends to 0 and $F_{\text{hr}}(\rho) \rightarrow 0$. Figure 1 shows a graph of $F_{\text{hr}}(\rho)$ when $k = 2$.

4.2. Covering configurations

Covering configurations are those for which we have $N_{n:n} \leq k$. In the partitioning approach of the fortune N amongst n recipients, this event is realized when the share of the richest is bounded above by k (a rare event). Assume $n, N \rightarrow \infty$ with $n/N \rightarrow \rho \in (1/k, 1)$ where $k > 1$ is fixed. One also expects that the probability of covering configurations by arcs of length k tends to zero exponentially fast. Working now with

$$Z_{n,p} = \sum_{k_1, \dots, k_n \geq 1} \prod_{m=1}^n 1_{k_m \leq k} e^{-pk_m} = \left(e^{-p} \frac{1 - e^{-pk}}{1 - e^{-p}} \right)^n$$

and proceeding as for the hard rods case we easily get

$$-\frac{1}{n} \log \mathbf{P}(N_{n:n} \leq k) \rightarrow F_c(p, \rho),$$

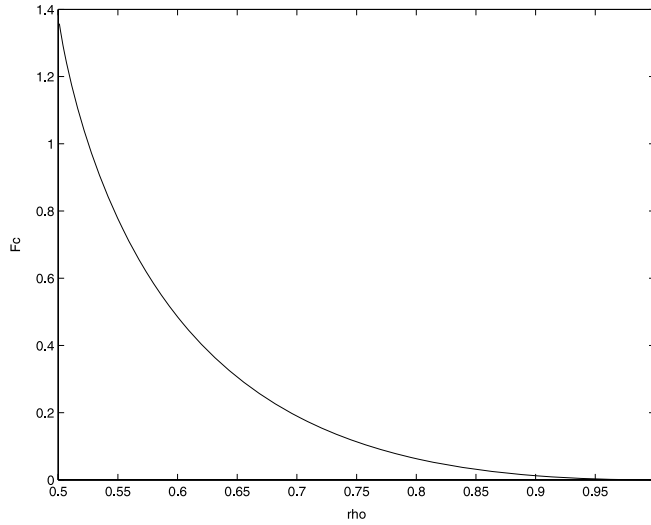


Figure 2. A plot of F_c versus ρ when $k = 2$ or $\rho \in (1/2, 1)$.

where the covering large deviation rate function is

$$F_c(p, \rho) = -\frac{1}{\rho} (\rho \log \rho + (1 - \rho) \log(1 - \rho)) - \frac{p}{\rho} - \log \left(\frac{e^{-p}(1 - e^{-pk})}{1 - e^{-p}} \right). \quad (26)$$

Here, thermodynamical pressure p and density $\rho \in (1/k, 1)$ are related through the covering state equation $\partial_p F_c(p, \rho) = 0$, namely

$$\frac{1}{\rho} = 1 + \frac{e^{-p}}{1 - e^{-p}} - \frac{ke^{-pk}}{1 - e^{-pk}}. \quad (27)$$

For all finite arc lengths k , k -covering configurations are also exceptional. The k -covering large deviation rate function F_c is in general an implicit function of ρ and k , $\rho \in (1/k, 1)$. When $\rho \downarrow 1/k$, pressure tends to $-\infty$ and $F_c(\rho) \rightarrow k \log k - (k - 1) \log(k - 1) > 0$. As $\rho \uparrow 1$, pressure tends to ∞ and $F_c(\rho) \rightarrow 0$. By continuity, there is a value of ρ_0 inside the definition domain of ρ where $p = 0$. We have $F_c(p, \rho_0) = -(1/\rho_0)(\rho_0 \log \rho_0 + (1 - \rho_0) \log(1 - \rho_0)) - \log k$. In the partitioning approach of the fortune N amongst n recipients, the share of the richest is bounded above with probability tending to 0 exponentially fast.

Remark. When $k = 2$, the covering equation of state can be solved explicitly because it boils down to a second degree equation in e^{-p} . One finds $p = -\log((1 - \rho)/(2\rho - 1))$. Plugging in this expression of p in $F_c(p, \rho)$ with $k = 2$ gives

$$F_c = -\frac{1}{\rho} (2\rho \log \rho - (2\rho - 1) \log(2\rho - 1)),$$

an explicit function of $\rho \in (1/2, 1)$. Note that $p \uparrow \infty$ as $\rho \uparrow 1$, $p \downarrow -\infty$ as $\rho \downarrow 1/2$ and $p = 0$ when $\rho = 2/3$. We have $F_c(p, 2/3) = \frac{3}{2} \log 3 - 2 \log 2$. Figure 2 shows a graph of $F_c(\rho)$ when $k = 2$.

4.3. Parking configurations

Parking configurations are those for which we have $N_{1:n} > k$, $N_{n:n} \leq 2k$. In the partitioning approach of the fortune N amongst n recipients, this event is realized if the share of the richest is bounded above by twice the share of the poorest. Assume $n, N \rightarrow \infty$ with $n/N \rightarrow \rho \in (1/(2k), 1/(k+1))$. One expects that the probability of k -parking configurations tends to zero exponentially fast. Working now with

$$Z_{n,p} = \sum_{k_1, \dots, k_n \geq 1} \prod_{m=1}^n 1_{k < k_m \leq 2k} e^{-pk_m} = \left(e^{-p(k+1)} \frac{1 - e^{-pk}}{1 - e^{-p}} \right)^n$$

and proceeding as for the hard rods case we easily get

$$-\frac{1}{n} \log \mathbf{P}(N_{1:n} > k, N_{n:n} \leq 2k) \rightarrow F_\pi(p, \rho),$$

where the parking large deviation rate function is

$$F_\pi(p, \rho) = -\frac{1}{\rho} (\rho \log \rho + (1 - \rho) \log(1 - \rho)) - \frac{p}{\rho} - \log \left(\frac{e^{-p(k+1)} (1 - e^{-pk})}{1 - e^{-p}} \right). \quad (28)$$

Here, thermodynamical pressure p and density $\rho \in (1/k, 1)$ are related through the parking equation of state $\partial_p F_\pi(p, \rho) = 0$, namely

$$\frac{1}{\rho} = k + 1 + \frac{e^{-p}}{1 - e^{-p}} - \frac{ke^{-pk}}{1 - e^{-pk}}. \quad (29)$$

The parking configuration large deviation rate function F_π is an implicit function of ρ and k with $\rho \in (1/(2k), 1/(k+1))$. The latter formula can be extended to the border case $k = 1$. Indeed, when $k = 1$, $\rho = 1/2$, pressure tends to ∞ and $F_\pi(p, \rho) = 2 \log 2$. From the Stirling formula, this is in agreement with the fact $\mathbf{P}(N_{1:n} > 1, N_{n:n} \leq 2) = \binom{N-1}{n-1}^{-1} \neq 0$ only if $N = 2n$, which is the probability of the regular configuration where the n sampled points are all exactly equally spaced by two arc length units.

Remark. When $k = 2$, the parking equation of state can be solved explicitly to give $p = -\log((1 - 3\rho)/(4\rho - 1))$. Plugging in this expression of p into $F_\pi(p, \rho)$ with $k = 2$ gives F_π as an explicit function of $\rho \in (1/4, 1/3)$. Note that $p \uparrow \infty$ as $\rho \uparrow 1/3$, $p \downarrow -\infty$ as $\rho \downarrow 1/4$ and $p = 0$ when $\rho = 2/7$. figure 3 shows a graph of $F_\pi(\rho)$ when $k = 2$.

Equations (25), (26) and (28), as given from their respective equations of state, constitute the discrete versions of the large deviation rate functions occurring in the continuum in [18, 12, 21], [12], [21, 4], respectively, for hard rods, covering and parking configurations.

5. The grand canonical partition of N

Suppose N indistinguishable balls are assigned at random into N distinguishable boxes. Let $N_{n,N} \geq 0$ be the number of balls in box number n . This leads to a random partition

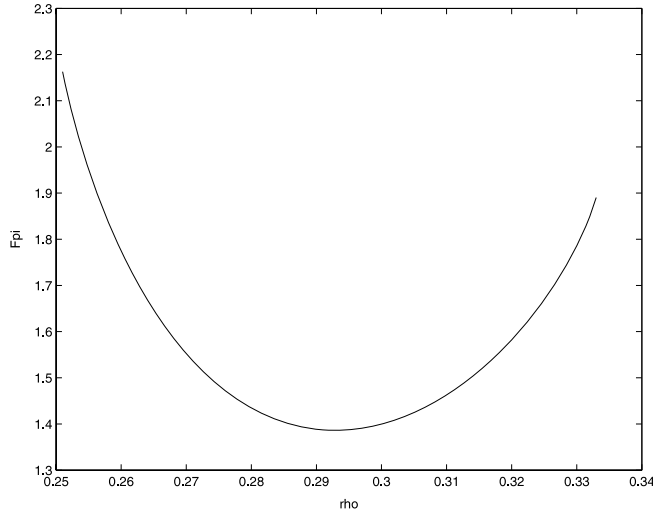


Figure 3. A plot of F_π versus ρ when $k = 2$ or $\rho \in (1/4, 1/3)$.

of N now into N id summands which are ≥ 0 :

$$N = \sum_{n=1}^N N_{n,N}. \quad (30)$$

We have

$$\mathbf{P}(N_{1,N} = k_1, \dots, N_{N,N} = k_N) = \frac{1}{\binom{2N-1}{N}}, \quad (31)$$

which is a Bose–Einstein distribution on the full N -simplex:

$$\left\{ k_n \geq 0 \text{ satisfying } \sum_{n=1}^N k_n = N \right\}.$$

Summing over all the k_n but one, the marginal distribution of $N_{1,N}$ is easily seen to be

$$\mathbf{P}(N_{1,N} = k) = \frac{\binom{2N-k-2}{N-k}}{\binom{2N-1}{N}}, \quad k = 0, \dots, N. \quad (32)$$

Let $P_N = \sum_{n=1}^N 1(N_{n,N} > 0)$ count the number of summands which are strictly positive (the number of non-empty boxes). With $k_m \geq 1$ satisfying $\sum_{m=1}^n k_m = N$, we obtain

$$\mathbf{P}(N_{1,N} = k_1, \dots, N_{n,N} = k_n; P_N = n) = \frac{\binom{N}{n}}{\binom{2N-1}{N}}, \quad (33)$$

which is independent of the filled box occupancies (k_1, \dots, k_n) (the probability being uniform).

As there are $\binom{N-1}{n-1}$ sequences $k_m \geq 1$, $m = 1, \dots, n$ satisfying $\sum_{m=1}^n k_m = N$, summing over the $k_m \geq 1$, we get the hypergeometric distribution for P_N :

$$\mathbf{P}(P_N = n) = \frac{\binom{N}{n} \binom{N-1}{n-1}}{\binom{2N-1}{N}}, n = 1, \dots, N. \quad (34)$$

This distribution occurs in the following urn model. Draw N balls without replacement from an urn containing $2N - 1$ balls in total, N of which are white, $N - 1$ are black. The law of P_N describes the probability that there are n white balls drawn from the urn. Its mean is $N^2/(2N - 1) \sim N/2$ and its variance is $(N^2(N - 1))/(2(2N - 1)^2) \sim N/8$.

As a result,

$$\mathbf{P}(N_{1,N} = k_1, \dots, N_{n,N} = k_n \mid P_N = n) = \frac{1}{\binom{N-1}{n-1}} 1(|\mathbf{k}| = N), \quad (35)$$

which is the spacings conditional Bose–Einstein model with $\mathbf{k} \geq \mathbf{1}$ described in (3). The balls in boxes model just defined is therefore an extension of the conditional Bose–Einstein model allowing the number of sampled points to be unknown and random.

Repetitions (grand canonical). It is likely that some boxes contain the same number of particles. To take these multiplicities into account, let $A_{i,N}$, $i \in \{0, \dots, N\}$ count the number of boxes with exactly i balls, that is

$$A_{i,N} = \# \{n \in \{1, \dots, N\} : N_{n,N} = i\} = \sum_{n=1}^N 1(N_{n,N} = i). \quad (36)$$

Then $\sum_{i=0}^N A_{i,N} = N$, where $\sum_{i=1}^N A_{i,N} = P_N$ is the number of filled boxes and $A_{0,N} = N - P_N$ the number of empty ones. The joint probability of the $A_{i,N}$ is given by the Ewens formula (see [5, 13])

$$\mathbf{P}(A_{0,N} = a_0, A_{1,N} = a_1, \dots, A_{N,N} = a_N) = \frac{1}{\binom{2N-1}{N}} \frac{N!}{\prod_{i=0}^N a_i!}, \quad (37)$$

on the set $\sum_{i=0}^N a_i = \sum_{i=1}^N i a_i = N$.

Let us now investigate the marginal law of the $A_{i,N}$. Firstly, the law of $A_{0,N} = N - P_N$ clearly is

$$\mathbf{P}(A_{0,N} = a_0) = \frac{\binom{N}{a_0} \binom{N-1}{a_0}}{\binom{2N-1}{N}}, a_0 = 0, \dots, N-1, \quad (38)$$

with $\mathbf{E}(A_{0,N}) \sim N/2$. Secondly, recalling $A_{i,N} = \sum_{n=1}^N 1(N_{n,N} = i)$, with $(N)_l := N(N-1) \cdots (N-l+1)$, using the exchangeability of $(N_{1,N}, \dots, N_{N,N})$, the probability generating function of $A_{i,N}$ ($i \neq 0$) reads

$$\mathbf{E}(z^{A_{i,N}}) = 1 + \sum_{l \geq 1} \frac{(z-1)^l}{l!} (N)_l \mathbf{P}(N_{1,N} = i, \dots, N_{l,N} = i).$$

Using $\mathbf{P}(N_{1,N} = k_1, \dots, N_{l,N} = k_l) = \binom{2N-l-\sum_{m=1}^l k_m - 1}{N-l-1} / \binom{2N-1}{N-1}$, we get the falling factorial moments of $A_{i,N}$ as

$$m_{l,i}(N) := \mathbf{E}[(A_{i,N})_l] = (N)_l \binom{2N-l-li-1}{N-l-1} / \binom{2N-1}{N-1}, \quad (39)$$

where $l \in \{0, \dots, l(i) = (N-1) \wedge [N/i]\}$. The marginal distribution of $A_{i,N}$ is thus

$$\mathbf{P}(A_{i,N} = a_i) = \sum_{l=a_i}^{l(i)} \frac{(-1)^{l-a_i}}{l!} \binom{l}{a_i} m_{l,i}(N), \quad a_i \in \{0, \dots, l(i)\}. \quad (40)$$

If $l = 1$, $\mathbf{E}(A_{i,N}) = N \binom{2N-i-2}{N-2} / \binom{2N-1}{N-1}$. The variance of $A_{i,N}$ is $\sigma^2(A_{i,N}) = m_{2,i}(N) + m_{1,i}(N) - m_{1,i}(N)^2$.

In particular, we find that $\mathbf{E}(A_{1,N}) = N(N+1)/(2(2N+1)) \sim N/4$ is the mean number of singleton boxes in the grand canonical model: when N is large, about one fourth out of the N boxes is filled by singletons (recall that one half of N is filled by no ball). The variance of $A_{1,N}$ is $\sigma^2(A_{1,N}) \sim N/4$ so we expect that $A_{1,N}$, properly normalized, converges to a normal distribution. Next, we can check that $\mathbf{E}(A_{2,N}) \sim N/8$ and further that $\mathbf{E}(A_{i,N}) \sim N/2^{i+1}$, showing a geometric decay in i of $\mathbf{E}(A_{i,N})$.

Finally, note that the probability that $A_{i,N}$ takes its maximal possible value $l(i)$ is

$$\mathbf{P}(A_{i,N} = l(i)) = m_{l(i),i}(N) / l(i)! = \binom{N}{l(i)} / \binom{2N-1}{N-1}.$$

For example $\mathbf{P}(A_{1,N} = N-1) = N / \binom{2N-1}{N-1}$ is the (exponentially small) probability that all N boxes are filled by singletons.

Multiplicities and conditioning. Let us now investigate the same problem while conditioning on $P_N = n$.

Firstly, note that $\sum_{i=1}^N i A_{i,N}(i) = N$ is the total number of balls. Using the multinomial formula, with $\sum_{i=1}^N i a_i = N$ and $\sum_{i=1}^N a_i = n$, we thus get

$$\mathbf{P}(A_{1,N} = a_1, \dots, A_{N,N} = a_N, P_N = n) = \frac{\binom{N}{n}}{\binom{2N-1}{N}} \frac{n!}{\prod_{i=1}^N a_i!} \quad (41)$$

and

$$\mathbf{P}(A_{1,N} = a_1, \dots, A_{N,N} = a_N \mid P_N = n) = \frac{n!}{\binom{N-1}{n-1}} \frac{1}{\prod_{i=1}^N a_i!}. \quad (42)$$

The latter formulas give the joint (Ewens-like) distributions of the repetition vector count.

Let us investigate the marginal distribution of the $A_{i,N}$ conditional given $P_N = n$. Firstly, the law of $A_{0,N} = N - P_N$ is $\mathbf{P}(A_{0,N} = a_0 \mid P_N = n) = \delta_{a_0-(N-n)}$.

Secondly, recalling $A_{i,N} = \sum_{n=1}^N 1(N_{n,N} = i)$, with $(n)_l = n(n-1) \cdots (n-l+1)$ (and $(n)_0 := 1$), using the exchangeability of $(N_{1,N}, \dots, N_{N,N})$, the conditional probability generating function of $A_{i,N}$ reads

$$\mathbf{E}(z^{A_{i,N}} \mid P_N = n) = 1 + \sum_{l \geq 1} \frac{(z-1)^l}{l!} (n)_l \mathbf{P}(N_{1,N} = i, \dots, N_{l,N} = i \mid P_N = n).$$

Using $\mathbf{P}(N_{1,N} = k_1, \dots, N_{l,N} = k_l \mid P_N = n) = \binom{N - \sum_{l=1}^l k_l - 1}{n - l - 1} / \binom{N - 1}{n - 1}$, we get the conditional falling factorial moments of $A_{i,N}$ as

$$m_{l,i}(n, N) := \mathbf{E}[(A_{i,N})_l \mid P_N = n] = (n)_l \binom{N - li - 1}{n - l - 1} / \binom{N - 1}{n - 1}, \quad (43)$$

where $l \in \{0, \dots, l(i) = (n - 1) \wedge [(N - 1)/i]\}$. The conditional marginal distribution of $A_{i,N}$ is thus

$$\mathbf{P}(A_{i,N} = a_i \mid P_N = n) = \sum_{l=a_i}^{l(i)} \frac{(-1)^{l-a_i}}{l!} \binom{l}{a_i} m_{l,i}(n, N). \quad (44)$$

If $l = 1$, $\mathbf{E}(A_{i,N} \mid P_N = n) = n \binom{N-i-1}{n-2} / \binom{N-1}{n-1}$. In particular, $\mathbf{E}(A_{1,N} \mid P_N = n) = n(n-1)/(N-1)$ is the mean number of singleton boxes. In the thermodynamical limit $n, N \rightarrow \infty$, $n/N \rightarrow \rho$, $\mathbf{E}(A_{1,N} \mid P_N = n) \sim \rho n$ and a fraction ρ of the n filled boxes is filled with singletons. For the variance, we have $\sigma^2(A_{1,N} \mid P_N = n) \sim \rho n$. We can also check that a fraction $\rho(1-\rho)$ of the n filled boxes is filled with doubletons, $\mathbf{E}(A_{2,N} \mid P_N = n) \sim \rho(1-\rho)n$, and more generally that $\mathbf{E}(A_{i,N} \mid P_N = n) \sim \rho(1-\rho)^i n$.

Finally, note that the probability that $A_{i,N}$ reaches its maximal possible value $l(i)$ is

$$\mathbf{P}(A_{i,N} = l(i) \mid P_N = n) = m_{l(i),i}(n, N) / l(i)! = \binom{n}{l(i)} / \binom{N-1}{n-1}.$$

For example $\mathbf{P}(A_{1,N} = n - 1 \mid P_N = n) = n / \binom{N-1}{n-1}$ is the probability that $n - 1$ boxes are filled by singletons and one box by $N - n + 1$ balls, which is obvious.

6. Random graph connectivity

The latter model may be viewed as a clockwise k -nearest neighbor graph with N vertices and kn edges. Consider as before N equally spaced points (vertices) on the N -circle so with arc length 1 between consecutive points. Draw at random $n \in \{2, \dots, N - 1\}$ points without replacement at the integer vertices of this circle. Assume $N \leq 2n$ and draw an edge at random from each of the n sampled points, removing each sampled point once it has been paired. At the end of this process, we get a random graph with N vertices and n (out-degree 1) edges whose endpoints are no longer neighbors, being now chosen at random on $\{1, \dots, N\}$. We wish to estimate the covering probability for this new model in the spirit of Erdős–Rényi random graphs.

Let B_m , $m = 1, \dots, n$ be a sequence of independent (but not id) Bernoulli rvs with success probabilities $p_m = ((N - n)/(N - (m - 1)))$, $m = 1, \dots, n$. With $[z^k]\phi(z)$ the z^k -coefficient of $\phi(z)$, the N -covering probability is

$$\mathbf{P}_{n,N}(\text{cover}) = \mathbf{P}\left(N - n \leq \sum_{m=1}^n B_m \leq n\right) = \sum_{k=N-n}^n [z^k] \mathbf{E}\left(z^{\sum_{m=1}^n B_m}\right), \quad (45)$$

which is just the probability of hitting all points of the un-sampled set $\{n + 1, \dots, N\}$ at least once in a uniform pairing without replacement of the n -sample. This covering probability is the probability of connectedness of the random graph with N vertices and

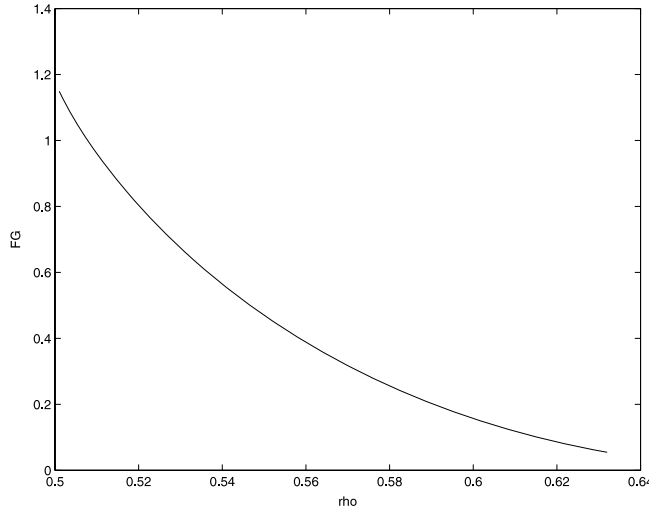


Figure 4. A plot of F_G versus ρ when $\rho \in (1/2, 1 - e^{-1})$.

n out-degree 1 edges. It is of course zero if $N > 2n$. Let $\bar{p}_n = (1/n) \sum_{m=1}^n p_m$ be the sample mean of the Bernoulli rvs. The covering probability can be bounded by

$$\mathbf{P}_{n,N}(\text{cover}) \leq \sum_{k=N-n}^n \binom{n}{k} (1 - \bar{p}_n)^k \bar{p}_n^{n-k}. \quad (46)$$

Assume $n, N \rightarrow \infty$ while $n/N \rightarrow \rho$, so with $\rho \in (1/2, 1)$. Then

$$\bar{p}_n \rightarrow -\frac{1-\rho}{\rho} \log(1-\rho) =: \mu(\rho).$$

Clearly $\sigma^2(B_m) < \infty$ and $\sum_{m=1}^n m^{-2} \sigma^2(B_m)$ has a finite limit. By the Kolmogorov strong law of large numbers $(1/n) \sum_{m=1}^n B_m \xrightarrow{a.s.} \mu(\rho)$ and so $\mathbf{P}_{n,N}(\text{cover}) \rightarrow 1$ if $\rho \geq \rho_c := 1 - e^{-1}$ because in this case the probability to estimate is

$$\mathbf{P}\left(\frac{1}{n} \sum_{m=1}^n B_m \geq \frac{1-\rho}{\rho}\right),$$

with $\mu(\rho) \in [(1-\rho)/\rho, 1]$.

Meanwhile, when $\rho < 1 - e^{-1}$, the bound for the covering probability can be estimated by

$$\begin{aligned} \mathbf{P}_{n,N}(\text{cover}) &\leq N \int_{1-\rho}^{\rho} \binom{N\rho}{Nx} (1 - \mu(\rho))^{Nx} \mu(\rho)^{N(\rho-x)} dx \\ &\sim CN \int_{1-\rho}^{\rho} e^{NH_{\rho}(x)} dx, \end{aligned}$$

where $H_{\rho}(x) = \rho \log \rho - x \log x - (\rho - x) \log(\rho - x) + x \log(1 - \mu(\rho)) + (\rho - x) \log \mu(\rho)$. The function $x \rightarrow H_{\rho}(x)$ is concave and attains its maximum at $x = \rho(1 - \mu(\rho)) < 1 - \rho$,

which is outside the integration interval $[1 - \rho, \rho]$. By the saddle point method, when $\rho \in (1/2, \rho_c)$

$$\liminf_{n, N \rightarrow \infty, n/N \rightarrow \rho} -\frac{1}{n} \log \mathbf{P}_{n, N}(\text{cover}) = F_G(\rho) := -\frac{1}{\rho} H_\rho(1 - \rho) > 0. \quad (47)$$

So only in the low-density range $\frac{1}{2} < \rho < \rho_c := 1 - e^{-1}$ is the graph's connectedness probability exponentially small. Note that the graph's large deviation rate function F_G is maximal (minimal) at $\rho = 1/2$ ($\rho_c = 1 - e^{-1}$), with $F_G(\rho) \xrightarrow{\rho \rightarrow \rho_c^-} -((3 - e)/(e - 1)) \log(e - 2) > 0$. Figure 4 shows a graph of $F_G(\rho)$ when $\rho \in [\rho_c, 1)$, $F_G(\rho) = 0$.

We conclude that in the random graph approach to the covering problem, in sharp contrast to the k -nearest neighbor graph (compare with (26) with $k = 2$ and $\rho \in (1/2, 1)$), there exists a critical density $\rho_c = 1 - e^{-1}$ above which covering occurs with probability one. These results illustrate to what extent, when connections are not restricted to neighbors, the chance of connectedness is increased. This question was also raised in ([2], p 18) in relation to small-world graphs.

References

- [1] Bollobás B, 2001 *Random Graphs* (Cambridge Studies in Advanced Mathematics vol 73) 2nd edn (Cambridge: Cambridge University Press)
- [2] Cannings C, *Modelling protein–protein interactions networks from yeast-2-hybrid screens with random graphs*, 2006 *Statistics in Genomics and Proteomics* ed A Urfer and M A Turkman (Coimbra: Centro Internacional de Matematica)
- [3] Darling D A, *On a class of problems related to the random division of an interval*, 1953 *Ann. Math. Stat.* **24** 239
- [4] Dunlop F and Huillet T, *Hard rods: statistics of parking configurations*, 2003 *Physica A* **324** 698
- [5] Ewens W J, *The sampling theory of selectively neutral alleles*, 1972 *Theoret. Popul. Biol.* **3** 87
Ewens W J, 1972 *Theoret. Popul. Biol.* **3** 240 (erratum)
Ewens W J, 1972 *Theoret. Popul. Biol.* **3** 376 (erratum)
- [6] Feller W, 1971 *An Introduction to Probability Theory and its Applications* 2nd edn, vol 2 (New York: Wiley)
- [7] Flatto L and Konheim A G, *The random division of an interval and the random covering of a circle*, 1962 *SIAM Rev.* **4** 211
- [8] Holst L, *A note on random arcs on the circle*, 1983 *Probability and Mathematical Statistics* (Essays in Honour of Carl–Gustav Esseen) ed A Gut and L Holst, Uppsala pp 40–5
- [9] Holst L and Hüsler J, *On the random coverage of the circle*, 1984 *J. Appl. Probab.* **21** 558
- [10] Holst L, *On discrete spacings and the Bose–Einstein distribution*, 1985 *Contributions to Probability and Statistics* (Essays in Honour of Gunnar Blom) ed J Lanke and G Lindgren, Lund pp 169–77
- [11] Huillet T, *Random covering of the circle: the size of the connected components*, 2003 *Adv. Appl. Probab.* **35** 563
- [12] Huillet T, *Random covering of the circle: the configuration-space of the free deposition process*, 2003 *J. Phys. A: Math. Gen.* **36** 12143
- [13] Huillet T, *Sampling formulae arising from random Dirichlet populations*, 2005 *Commun. Stat. Theory Methods* **34** 1019
- [14] Ivchenko G I, *On the random covering of a circle: a discrete model*, 1994 *Diskret. Math.* **6** 94
- [15] Johnson N L and Kotz S, *Urn models and their application*, 1977 *An Approach to Modern Discrete Probability Theory* (Wiley Series in Probability and Mathematical Statistics) (New York: Wiley) p xiii+402
- [16] Kolchin V F, Sevastyanov B A and Chistyakov V P, 1978 *Random Allocations* Translated from the Russian. Translation edited by A V Balakrishnan (*Scripta Series in Mathematics*) (Washington, DC: V H Winston & Sons) (distributed by Halsted Press (New York: Wiley))
- [17] Pyke R, *Spacings (with discussion)*, 1965 *J. R. Stat. Soc. Ser. B* **27** 395
- [18] Salzburg Z, Zwanzig R and Kirkwood J, *Molecular distribution functions in a one-dimensional fluid*, 1953 *J. Chem. Phys.* **21** 1098

- [19] Siegel A F, *Random arcs on the circle*, 1978 *J. Appl. Probab.* [15 774](#)
- [20] Stevens W L, *Solution to a geometrical problem in probability*, 1939 *Ann. Eugen.* **9** 315
- [21] Tarjus G and Viot P, *Statistical mechanical description of the parking-lot model for vibrated granular materials*, 2004 *Phys. Rev. E* [69 011307](#)
- [22] Withworth W A, 1897 *Exercises on Choice and Chance* (Cambridge: Deighton Bell and Co.) Republished by Hafner, New York, 1959