The Annals of Applied Probability
2018, Vol. 28, No. 3, 1449–1490
https://doi.org/10.1214/17-AAP1334

STOCHASTIC COAGULATION-FRAGMENTATION PROCESSES WITH A FINITE NUMBER OF PARTICLES AND APPLICATIONS

BY NATHANAEL HOZE* AND DAVID HOLCMAN[†,‡,1]

ETH Zurich, University of Cambridge[†] and Ecole Normale Supérieure[‡]*

Coagulation-fragmentation processes describe the stochastic association and dissociation of particles in clusters. Cluster dynamics with cluster-cluster interactions for a finite number of particles has recently attracted attention especially in stochastic analysis and statistical physics of cellular biology, as novel experimental data are now available, but their interpretation remains challenging. We derive here probability distribution functions for clusters that can either aggregate upon binding to form clusters of arbitrary sizes or a single cluster can dissociate into two sub-clusters. Using combinatorics properties and Markov chain representation, we compute steady-state distributions and moments for the number of particles per cluster in the case where the coagulation and fragmentation rates follow a detailed balance condition. We obtain explicit and asymptotic formulas for the cluster size and the number of clusters in terms of hypergeometric functions. To further characterize clustering, we introduce and discuss two mean times: one is the mean time two particles spend together before they separate and the other is the mean time they spend separated before they meet again for the first time. Finally, we discuss applications of the present stochastic coagulation-fragmentation framework in cell biology.

Introduction. Clustering appears in various areas of science such as astrophysics, where masses can form aggregate under gravitation, biochemistry where molecules have to meet to react, colloids that aggregate in solution or ecology where prey–predator have to meet to stabilize populations. A century ago, Von Smoluchowski [33] described an irreversible aggregation of many particles in clusters. Later on, a set of coagulation-fragmentation equations was proposed by Becker–Döring when clusters can lose or gain only one particle at a time [6, 8, 21, 34]. Nowadays, continuous limit analysis [3, 26], determinist, stochastic, asymptotical and numerical methods are developed to study clustering and estimate the number of clusters [2, 9, 11, 27, 32] and their sizes.

The coagulation models mentioned above have been extended for cluster growth with a finite number of particles [23, 24], known as the Marcus–Lushnikov process, and more recently for discrete coagulation-fragmentation models with an

Received November 2016; revised May 2017.

[1] Supported by a Marie Curie Award and a Simons fellowship.

MSC2010 subject classifications. Primary 60J20; secondary 05A17.

Key words and phrases. Markov chain, coagulation-fragmentation processes, stochastic processes.

infinite number of particles [5]. Many statistical and probabilistic studies have been made on the Marcus–Lushnikov process [18]; however, much less is known about the statistical properties for the coagulation-fragmentation of a finite number of particles [13]. Of particular interest in stochastic biology are models of coagulation-fragmentation where the cluster size cannot exceed a given threshold [14, 35]. These models are used in genetics to describe the organization in clusters of the chromosome ends [14] or to model viral capsid assembly in cells [15, 16, 36]. This article presents several exact and asymptotic results and it aims to attract attention toward developments of applied probability with direct applications to interpret data.

We recall that the Smoluchowski equations for coagulation-fragmentation consist of an infinite system of differential equations for the number $n_j(t)$ of clusters of size j at time t in a population of infinite size [33]. The index j can take values between 1 and ∞ and

$$\begin{aligned}\frac{dn_j(t)}{dt} &= \frac{1}{2}\sum_{k=0}^{j-1} A(k, j-k)n_k(t)n_{j-k}(t) - n_j(t)\sum_{k=1}^{\infty} A(j,k)n_k(t)\\ &\quad - \frac{1}{2}n_j(t)\sum_{k=1}^{j-1} B(k, j-k) + \sum_{k=1}^{\infty} B(j,k)n_{k+j}(t),\end{aligned} \tag{1}$$

where the first line in the right-hand side corresponds to the coagulation and the second accounts for the fragmentation. The coagulation kernel $A(i, j)$ is the rate at which two clusters of size i and j coalesce to form a cluster of size $i + j$, while the fragmentation kernel $B(i, j)$ is the rate at which a cluster of size $i + j$ dissociates into a cluster of size i and a cluster of size j. When cluster sizes cannot exceed a threshold M, the system of equations (1) is truncated and the coagulation kernel is $A(i, j) = 0$ if $i + j \geq M$ [7, 10]. This system of equations is a mean-field deterministic model of the coagulation-fragmentation process that describes a discrete number of particles aggregating and dissociating, but it does not allow a complete analysis of the cluster distribution in the small size limit [13].

The goal of this paper is to study a stochastic system of coagulation-fragmentation with a limited number of particles. We present exact solutions and expressions for the distribution, statistical moments of clusters and number of particles per cluster, using the following generic rules of coagulation-fragmentation: a cluster of size $i + j$ can give rise to two clusters of size i and j at a rate $F(i, j)$, and two clusters of size i and j form a new cluster of size $i + j$ at a rate $C(i, j)$.

We focus our analysis on coagulation-fragmentation processes (CFP) that verify the detailed balance condition [12], for which there exists a function $a(i) = a_i$ such that $\forall i, j \in \mathbb{N}$:

$$\frac{C(i, j)}{F(i, j)} = \frac{a(i + j)}{a(i)a(j)}. \tag{2}$$

The detailed balance condition ensures that, at equilibrium, each elementary coagulation or fragmentation process is equilibrated by its reciprocal process. Using exact formulas for probabilities for these configurations when the total number of clusters is fixed, we compute the probability distribution function of the number of clusters. We also compute the probability distribution that the number of cluster of size i is m_i so that the distribution of sizes of the ensemble of clusters is $(m_1, \dots, m_n)$. When there are exactly N particles and the total number of clusters is fixed to K, we have the following identity for number conservation:

$$\sum_{i=1}^{N} m_i = K. \tag{3}$$

We shall show that when the total number of clusters is K, the conditional probability distribution function is given by

$$p'(m_1, \dots, m_N | K) = \frac{1}{C_{N,K}} \frac{a(1)^{m_1} \cdots a(N)^{m_N}}{m_1! \cdots m_N!}, \tag{4}$$

where the normalization constant $C_{N,K}$ can be computed explicitly [see formula (126)]. We will use this formula to compute the statistical moments for the cluster distributions. Using a combinatorial approach, we present several explicit formula for the steady-state distributions of clusters. Some of the results presented here were previously announced without proofs in the short letter [14].

The paper is organized as follows. In Section 1, we present the stochastic model of coagulation-fragmentation for N independent particles that we analyze by a Markov chain. We obtain explicit formula for the cluster configuration combinatorics using the partition of the integer N, for the distribution of particles in clusters. We derive the time evolution equations for the number of clusters. These equations represent a novel Markov chain, which allows us to determine the number of clusters at steady-state. By combining the expression for number of clusters with the distribution of cluster in configuration conditioned on the number of clusters, we obtain the distribution of the particles in clusters. In Section 3, we introduce two characteristic times for studying the time distribution two particles are in a cluster. These two times characterize the dynamics of exchange of particles between clusters: the first one is the mean time that two particles spend together before they separate, and the second is the mean time that they spend separated before they meet again for the first time. The colocalization probability of two particles is defined as the fraction of time, the particles spend together. Sections 4–6 provide direct applications of these results to specific CFPs: we focus specifically on the case of constant coagulation and fragmentation kernels in Section 4, for which we obtain analytical expressions for the clusters distributions. We also obtain several formula when the cluster size is limited.

In this article, to determine statistics of the cluster distribution, we alternatively study two different systems. First, we study the distribution of clusters using the integer partition of the total number of particles. We obtain the probability distribution of cluster configurations. In this process, we use the term *coagulation* when two clusters of a given size coalesce and form a new cluster. We use the term *fragmentation* to describe the separation of a cluster into two smaller ones. The coagulation and fragmentation kernels that we have chosen here allow us to perform another analysis: when the number of clusters is fixed, the overall rates of coagulation and fragmentation are independent of the configurations of the clusters. We will study aggregation-fragmentation, where the number of clusters is known. In that case, we shall use the following terminology: *formation* describes the change when a distribution of K becomes $K-1$ clusters. We use *separation* to describe the process by which a distribution of K clusters is transformed into a distribution of $K+1$ clusters.

1. Coagulation-fragmentation with a finite number of independent particles.

1.1. *Stochastic coagulation-fragmentation equations for a finite number of particles.* To describe the steady-state distribution for a CFP stochastic model with a finite number of N particles, we shall use a continuous-time Markov chain in the space of cluster configurations. The N particles distributed in clusters of size $(n_1, \dots, n_N)$ can undergo coagulation or fragmentation events under the constraint that

$$\sum_{k=1}^{N} n_k = N. \tag{5}$$

To study the distribution of particles in clusters, we use the decomposition of the integer N as the sum of positive integers (integer partition) [4]. The partitions of the integer N are described in N dimensions by the ensemble

$$P_N = \left\{(n_1, \dots, n_N) \in \mathbb{N}^N; \sum_{i=1}^{N} n_i = N \text{ and } n_1 \geq \cdots \geq n_N \geq 0\right\}. \tag{6}$$

The probability $P(n_1, \dots, n_N, t)$ of the configuration $(n_1, \dots, n_N)$ at time t satisfies an ensemble of closed equations. Indeed, by considering all the possible coagulation or fragmentation events, the master equation is obtained by considering the events occurring between time t and $t + \Delta t$:

- Two clusters of size n_i and n_j coagulate with a probability $C(n_i, n_j)\Delta t$ to form a cluster of size $n_i + n_j$.
- A cluster of size n_i dissociates into two clusters of size k and $n_i - k$ with a probability $F(k, n_i - k)\Delta t$.

- Nothing happens with the probability $1 - \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} C(n_i, n_j)\Delta t - \sum_{i=1}^{N} \sum_{k=1}^{n_i - 1} F(k, n_i - k)\Delta t$.

Thus, the master equations are

$$\begin{aligned}
&\frac{d}{dt} P(n_1, \dots, n_N, t) \\
&\quad = -\Biggl(\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} C(n_i, n_j) + \sum_{i=1}^{N} \sum_{k=1}^{n_i-1} F(k, n_i - k)\Biggr) P(n_1, \dots, n_N, t) \\
&\qquad + \sum_{k=1}^{N} \sum_{\substack{n'_i>0, n'_j>0 \\ n'_i + n'_j = n_k}} C(n'_i, n'_j) P(n_1, \dots, n'_i, \dots, n'_j, \dots, n_N, t) \\
&\qquad + \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} F(n_i, n_j) P(n_1, \dots, n_i + n_j, \dots, n_N, t).
\end{aligned} \tag{7}$$

Moreover, $C(n_i, n_j) = 0$ if either n_i or n_j is equal to 0. We now introduce an ensemble in $\mathbb{N}^N$ which consists of integer decompositions of clusters with a given size:

$$P'_N = \Biggl\{(m_1, \dots, m_N) \in \mathbb{N}^N; \sum_{i=1}^{N} i m_i = N \text{ and } m_1, \dots, m_N \geq 0\Biggr\}. \tag{8}$$

In the ensemble P'_N, m_i is the number of occurrence of the integer i in the partition of the integer N.The two ensembles P_N and P'_N correspond to different representations of the clusters distributions. For example, when there are $N = 9$ particles, distributed in two clusters of one particle, two clusters of two and one cluster of three. Then the distributions are $(3, 2, 2, 1, 1, 0, 0, 0, 0) \in P_9$, and $(2, 2, 1, 0, 0, 0, 0, 0, 0) \in P'_9$.

A sufficient condition to obtain an invariant measure of the steady-state probability is the reversibility of the CFP [20, 22] where the coagulation-fragmentation kernel satisfies the detailed balance condition: there exists a function $a(i) = a_i$ such that [12]

$$C(i, j) a_i a_j = F(i, j) a_{i+j}. \tag{9}$$

The functions a_i characterize the ratio of the coagulation and fragmentation rates, and any function of the form $a'_i = \alpha^i a_i$ with $\alpha \neq 0$ satisfies the above condition. This condition insures the reversibility of the Markov chain and guarantees the existence of an invariant measure [12], where the steady-state probability of a given configuration $(m_1, \dots, m_N) \in P'_N$ is

$$P'(m_1, \dots, m_N) = \frac{1}{C_N} \frac{a_1^{m_1} \cdots a_N^{m_N}}{m_1! \cdots m_N!}, \tag{10}$$

where C_N is a normalization constant. An explicit computation of the normalization constant is difficult [31]. Here, we propose to estimate the probability of occurrence of a certain cluster configuration $(m_1, \dots, m_N)$ by limiting the study to the configurations of a given number of particles.

At this stage, we shall explain the rational for computing expression (10). This expression is the distribution at equilibrium of particles in clusters, where the dissociation (resp., association) rate is proportional to the number of elements (minus one) (resp., the number of pairs of particles).

The equilibrium probability distributions associated to the Markov chain configuration $(m_1, \dots, m_N)$ is computed from analyzing the transition between the two neighboring states $(m_1, \dots, m_i - 1, \dots, m_j - 1, \dots, m_{i+j} + 1, \dots, m_N)$ and $(m_1, \dots, m_N)$. It is obtained first from the coagulation rate $\psi(i, j)$ of a cluster of size i with one of size j, given the distribution $(m_1, \dots, m_N)$. The rate $\psi(i, j)$ is given by

$$\psi(i, j) = \frac{1}{2} C(i, j) m_i m_j \qquad \text{if } i \neq j \tag{11}$$

$$= C(i, i) m_i (m_i - 1) \qquad \text{otherwise.} \tag{12}$$

The factor $\frac{1}{2}$ accounts for the symmetric cases $\psi(i, j)$ and $\psi(j, i)$. The fragmentation rate $\phi(i, j)$ from $i + j$ to (i, j), that accounts for the transition from the configuration $(m_1, \dots, m_i - 1, \dots, m_j - 1, \dots, m_{i+j} + 1, \dots, m_N)$ to $(m_1, \dots, m_N)$ (Figure 1) is

$$\phi(i, j) = F(i, j)(m_{i+j} + 1). \tag{13}$$

$$\begin{tikzcd}
(m_1, \dots, m_i - 1, \dots, m_j - 1, \dots, m_{i+j} + 1, \dots, m_N) \\
(m_1, \dots, m_N) \arrow[u, "\text{Coagulation } C(i, j) m_i m_j"'] \arrow[d, "\text{Fragmentation } F(k, i - k) m_i"'] \\
(m_1, \dots, m_{i-k} + 1, \dots, m_k + 1, \dots, m_i - 1, \dots, m_N)
\end{tikzcd}$$

FIG. 1. *Markov chain representation of the transitions. Starting with a configuration $(m_1, \dots, m_N)$, m_i is the number of clusters of size i and N is the total number of particles. The fragmentation rate of a cluster of size i into one cluster of size k and one of size $i - k$ is $F(k, i - k) m_i$, while the rate of formation of a cluster of size $i + j$ from two clusters of size i and j is equal to $C(i, j) m_i m_j$.*

Thus, the stationary probability π' satisfies the relation

$$\begin{aligned}
&\frac{\pi'(m_1,\dots,m_i-1,m_j-1,m_{i+j}+1,\dots,m_N)}{\pi'(m_1,\dots,m_N)}\\
&\qquad=\frac{\psi(i,j)}{\phi(i,j)}\\
&\qquad=\frac{1}{2}\frac{C(i,j)}{F(i,j)}\frac{m_i m_j}{m_{i+j}+1}=\frac{1}{2}\frac{a_i a_j}{a_{i+j}}\frac{m_i m_j}{m_{i+j}+1}.
\end{aligned}\tag{14}$$

A direct computation shows that the probability $P'(m_1,\dots,m_N)$ defined in equation (10) satisfies equation (14) (see Figure 1).

1.2. *Cluster partitions with a finite number of particles.* To determine the cluster distribution at equilibrium, we compute here the probability of a configuration when the number of clusters K is fixed. We also find the probability of having K clusters. The number of distributions of N particles into K clusters is the cardinal of the ensemble

$$P_{N,K}=\left\{(n_1,\dots,n_K)\in\mathbb{N}^K;\sum_{i=1}^K n_i=N \text{ and } n_1\ge\cdots\ge n_K\ge 1\right\},\tag{15}$$

which is also the ensemble of the partitions of the integer N as a sum of K integers. This ensemble is in bijection with

$$P'_{N,K}=\left\{(m_1,\dots,m_N)\in\mathbb{N}^N;\sum_{i=1}^N i m_i=N \text{ and } \sum_{i=1}^N m_i=K\right\},\tag{16}$$

where the transformation $P_{N,K}\to P'_{N,K}$ defined by

$$(n_1,\dots,n_K)\mapsto(m_1,\dots,m_N)=\left(\sum_{i=1}^K 1_{\{n_i=1\}},\dots,\sum_{i=1}^K 1_{\{n_i=N\}}\right)\tag{17}$$

maps the partition $(n_1,\dots,n_K)$, where N is written as a sum of K positive integers, to the number of occurrence of each integer into the image partition. The partitions of N are written as

$$P_N=\bigcup_K P_{N,K} \quad\text{and}\quad P'_N=\bigcup_K P'_{N,K}.\tag{18}$$

In Sections 4, 5 and 6, we derive explicitly expressions for the probabilities of configurations in $P'_{N,K}$.

1.3. *Statistical moments for the cluster configurations when the number of clusters is fixed.* We show now that the probability of configuration $(m_1, \dots, m_N)$, when the total number of clusters is equal to K, is given by

$$p'(m_1, \dots, m_N|K) = \frac{\frac{a_1^{m_1} \cdots a_N^{m_N}}{m_1! \cdots m_N!}}{C_{N,K}}, \tag{19}$$

where

$$C_{N,K} = \sum_{(m_i) \in P'_{N,K}} \frac{a_1^{m_1} \cdots a_N^{m_N}}{m_1! \cdots m_N!}. \tag{20}$$

The normalization factor of equation (19) is computed using the following result.

REMARK 1. We consider the functions

$$S(x) = \sum_{i=1}^{\infty} a_i x^i \tag{21}$$

and the partial sums

$$S_N(x) = \sum_{i=1}^{N} a_i x^i. \tag{22}$$

The Kth-power of these functions, that is, S^K and S_N^K have the same Nth-order coefficient and this coefficient determines $C_{N,K}$. Moreover, the function

$$g(x, y) = \exp\bigl(S(x)y\bigr) \tag{23}$$

is a generating function of the $C_{N,K}$.

PROOF. The number of configurations $(m_1, \dots, m_N)$ that satisfy the conditions $\sum_i m_i = K$ and $\sum_i i m_i = N$ is the Nth-order coefficient of the multinomial expansion

$$\begin{aligned}(a_1 x_1 + \cdots + a_N x_N)^K \\ = \sum_{(m_1,\dots,m_N), \sum m_i = K} \frac{K!}{m_1! \cdots m_N!} (a_1 x_1)^{m_1} \cdots (a_N x_N)^{m_N}.\end{aligned} \tag{24}$$

Using the N-tuple $(x_1, x_2, \dots, x_N) = (X, X^2, \dots, X^N)$, we obtain

$$\frac{(a_1 X + \cdots + a_N X^N)^K}{K!} = \sum_{(m_1,\dots,m_N), \sum m_i = K} \prod_i \frac{a_i^{m_i}}{m_i!} X^{\sum_i i m_i}. \tag{25}$$

We can group the terms by the exponents of X, which are equal to $\sum_i im_i$. In particular, for a partition $(m_1,\dots,m_N)\in P'_{N,K}$, the exponent is $\sum_i im_i = N$, and the Nth-order coefficient in expression (25) is equal to

$$\sum_{(m_1,\dots,m_N)\in P'_{N,K}} \frac{\prod a_i^{m_i}}{m_1!\cdots m_N!} = C_{N,K}. \tag{26}$$

More generally,

$$\frac{S^K(X)}{K!} = \sum_{n=K}^{\infty} \sum_{(m_1,\dots,m_N)\in P'_{n,K}} \frac{\prod a_i^{m_i}}{m_1!\cdots m_N!} X^n \tag{27}$$

$$= \sum_{n=K}^{\infty} \sum_{(m_1,\dots,m_N)\in P'_{n,K}} C_{n,K} X^n. \tag{28}$$

It follows that the function defined by

$$g(X,Y) = \exp\bigl(S(X)Y\bigr) \tag{29}$$

$$= \sum_{K=0}^{\infty} \frac{S^K(X)}{K!} Y^K \tag{30}$$

$$= \sum_{K=0}^{\infty} \sum_{n=K}^{\infty} C_{n,K} X^n Y^K \tag{31}$$

is a generating function of the $C_{N,K}$. □

REMARK 2. The coefficients $C_{N,K}$ defined by relation (20) satisfy the induction formula:

$$(N+1)C_{N+1,K} = \sum_{k=0}^{N-K+1} (k+1)a_{k+1}C_{N-k,K-1} \tag{32}$$

with

$$\begin{cases} C_{N,N} = \dfrac{a_1^N}{N!}, \\ C_{N,1} = a_N. \end{cases} \tag{33}$$

PROOF. We start with the formulas (33). The coefficient $C_{N,N}$ is obtained from the unique partition in $P'_{N,N}$, which is given by $m_1 = N$ and $m_i = 0$ if $i > 1$. Therefore, $C_{N,N} = \frac{a_1^N}{N!}$. The coefficient $C_{N,1}$ is obtained from the partition $m_N = 1$ and $m_i = 0$ if $i < N$, and thus $C_{N,1} = a_N$.

We now prove relation (32). Differentiating the function $\sigma(x) = \frac{S^K(x)}{K!}$, we get

$$\sigma'(x) = S'(x)\frac{S^{K-1}(x)}{(K-1)!}. \tag{34}$$

We evaluate the left-hand side of (34) using equation (28), and obtain

$$\begin{aligned}\sigma'(x) &= \left(\sum_{n=K}^{\infty} C_{n,K}x^n\right)' \\ &= \sum_{n=K-1}^{\infty} (n+1)C_{n+1,K}x^n \\ &= x^{K-1}\sum_{n=0}^{\infty}(n+K)C_{n+K,K}x^n.\end{aligned} \tag{35}$$

We evaluate the right-hand side of (34) using the definition of S (21) and we obtain

$$\begin{aligned}\sigma'(x) &= \left(\sum_{i=0}^{\infty}(i+1)a_{i+1}x^i\right)\left(\sum_{i=K-1}^{\infty} C_{i,K-1}x^i\right) \\ &= x^{K-1}\sum_{n=0}^{\infty}\left(\sum_{k=0}^{n}(k+1)a_{k+1}C_{n-k+K-1,K-1}\right)x^n.\end{aligned} \tag{36}$$

Thus, by equalizing the Nth-order coefficient of $\sigma'(x)$, we have

$$(N+1)C_{N+1,K} = \sum_{k=0}^{N-K+1} (k+1)a_{k+1}C_{N-k,K-1}. \tag{37}$$

□

Next, we estimate various moments when the number of clusters is fixed. We summarize the main result in the following.

THEOREM 1.1. *When the number of clusters is equal to K for a total of N particles, the mean number of clusters of size i is*

$$\langle M_i\rangle_{N,K} = a_i\frac{C_{N-i,K-1}}{C_{N,K}}, \tag{38}$$

where a_i and $C_{N,K}$ are defined in (9) *and* (20), *respectively.*

PROOF. The mean number of clusters of size i when the total number of clusters is K is given by the following sum:

$$\langle M_i\rangle_{N,K} = \sum_{P'_{N,K}} m_i p'(m_1,\dots,m_N) \tag{39}$$

$$= \frac{1}{C_{N,K}} \sum_{P'_{N,K}} m_i \frac{a_1^{m_1} \cdots a_N^{m_N}}{m_1! \cdots m_N!}$$

$$= \frac{1}{C_{N,K}} \sum_{P'_{N,K}, m_i > 0} a_i \frac{a_1^{m_1} \cdots a_i^{m_i - 1} \cdots a_N^{m_N}}{m_1! \cdots (m_i - 1)! \cdots m_N!}, \tag{40}$$

where the subscript $P'_{N,K}, m_i > 0$ in the sum characterizes the partitions in $P'_{N,K}$ containing at least one occurrence of the integer i. By considering the partitions of $P'_{N,K}$ where i appears at least once and removing one i, we obtain exactly the partitions of $P_{N-i,K-1}$, except for the number of occurrence of i, where the corresponding partitions in both sets have the same number of repetitions, i.e.

$$\forall j \neq i, m_j \qquad \text{in } (m_1, \dots, m_{N-i}) \in P'_{N-i,K-1} \tag{41}$$

is equal to

$$m_j \qquad \text{in } (m_1, \dots, m_N) \in P'_{N,K} \tag{42}$$

and m_i in $(m_1, \dots, m_{N-i}) \in P'_{N-i,K-1}$ is equal to $m_i + 1$ in $(m_1, \dots, m_N) \in P'_{N,K}$. There are no clusters larger than $N - i$ in the partitions [$m_j = 0$ for $j > N - i$ for $(m_1, \dots, m_N) \in P'_{N,K}$, $m_i > 0$]. Thus

$$\langle M_i \rangle_{N,K} = \frac{1}{C_{N,K}} \sum_{P'_{N,K}, m_i > 0} a_i \frac{a_1^{m_1} \cdots a_i^{m_i - 1} \cdots a_N^{m_N}}{m_1! \cdots (m_i - 1)! \cdots m_N!}$$

$$= \frac{a_i}{C_{N,K}} \sum_{P_{N-i,K-1}} \frac{a_1^{m_1} \cdots a_i^{m_i} \cdots a_N^{m_N}}{m_1! \cdots m_i! \cdots m_N!} \tag{43}$$

$$= a_i \frac{C_{N-i,K-1}}{C_{N,K}}. \qquad \square$$

We further have the following.

REMARK 3. $\langle M_i \rangle_{N,K} = 0$ if $i > N - K + 1$.

PROOF. When the N particles are distributed in K clusters, the largest cluster contains at most $N - K + 1$ particles. The corresponding partition in $P'_{N,K}$ is given by $m_1 = K - 1$, $m_{N-K+1} = 1$, and $m_j = 0$ otherwise. $\square$

In the following, we determine the second moment of the number of clusters of size i:

$$\langle M_i^2 \rangle_{N,K} = \frac{1}{C_{N,K}} \sum_{P'_{N,K}} m_i^2 \frac{a_1^{m_1} \cdots a_N^{m_N}}{m_1! \cdots m_N!} \tag{44}$$

and the covariance of the number of clusters of size i and j

$$\langle M_{i,j}^2\rangle_{N,K} = \frac{1}{C_{N,K}} \sum_{P'_{N,K}} m_i m_j \frac{a_1^{m_1}\cdots a_N^{m_N}}{m_1!\cdots m_N!}. \tag{45}$$

THEOREM 1.2. *The second moment of the number of clusters of size i is*

$$\langle M_i^2\rangle_{N,K} = a_i^2 \frac{C_{N-2i,K-2}}{C_{N,K}} + a_i \frac{C_{N-i,K-1}}{C_{N,K}}, \tag{46}$$

and the covariance is

$$\langle M_{i,j}^2\rangle_{N,K} = a_i a_j \frac{C_{N-i-j,K-2}}{C_{N,K}}. \tag{47}$$

PROOF. The variance of the number of clusters of size i is given by

$$\begin{aligned}\langle M_i^2\rangle_{N,K} &= \frac{1}{C_{N,K}} \sum_{m_i\in P'_{N,K}} m_i^2 \frac{a_1^{m_1}\cdots a_N^{m_N}}{m_1!\cdots m_N!}\\ &= \frac{1}{C_{N,K}} \sum_{m_i\in P'_{N,K}} [m_i(m_i-1)+m_i] \frac{a_1^{m_1}\cdots a_N^{m_N}}{m_1!\cdots m_N!}\\ &= \frac{1}{C_{N,K}} \sum_{m_i\in P'_{N,K}, m_i>1} a_i^2 \frac{a_1^{m_1}\cdots a_i^{m_i-2}\cdots a_N^{m_N}}{m_1!\cdots(m_i-2)!\cdots m_N!} + \langle M_i\rangle_{N,K}.\end{aligned} \tag{48}$$

Using an argument similar to the proof of Theorem 1.1, we obtain

$$\begin{aligned}\langle M_i^2\rangle_{N,K} &= \frac{1}{C_{N,K}} \sum_{m_i\in P_{N-2i,K-2}} a_i^2 \frac{a_1^{m_1}\cdots a_i^{m_i}\cdots a_N^{m_N}}{m_1!\cdots m_i!\cdots m_N!} + \langle M_i\rangle_{N,K}\\ &= a_i^2 \frac{C_{N-2i,K-2}}{C_{N,K}} + a_i \frac{C_{N-i,K-1}}{C_{N,K}}.\end{aligned} \tag{49}$$

The covariance of the number of clusters of size i and j is obtained from the term

$$\langle M_{i,j}^2\rangle_{N,K} = \frac{1}{C_{N,K}} \sum_{P'_{N,K}} m_i m_j \frac{a_1^{m_1}\cdots a_N^{m_N}}{m_1!\cdots m_N!}, \tag{50}$$

which, by the same reasoning, leads to

$$\langle M_{i,j}^2\rangle_{N,K} = a_i a_j \frac{C_{N-i-j,K-2}}{C_{N,K}}. \tag{51}$$

□

2. Distribution of the number of clusters. In the previous section, we computed the probability distribution of a cluster configuration and determined the statistical moments for a fixed number of clusters. In this section, we study the statistics of the entire cluster configurations. We focus here on the probability distribution of the *number of clusters* and we shall compute the time dependent probability density function

$$P_K(t) = P\{K \text{ clusters at time } t\}, \tag{52}$$

which is a birth-and-death process that we investigate using a Markov chain.

The probability of having K clusters at time $t + \Delta t$ is the sum of the probability of starting at time t with $K - 1$ clusters and one of them dissociates into two smaller ones plus the probability of starting with $K + 1$ clusters and two of them associate plus the probability of starting with K and nothing happens (Figure 2). The first probability is the product of P_{K-1} by the transition rate $s_{K-1}\Delta t$ to go from state with $K - 1$ clusters to K, while the second is the transition from $K + 1$ to K, which is the product of P_{K+1} by the transition rate $f_{K+1}\Delta t$ of going from $K + 1$ clusters to K. Thus the master equations are given by

$$\begin{cases} \dot{P}_1(t) = -s_1 P_1(t) + f_2 P_2(t), \\ \dot{P}_K(t) = -(f_K + s_K)P_K(t) + f_{K+1}P_{K+1}(t) + s_{K-1}P_{K-1}(t), \\ \dot{P}_N(t) = -f_N P_N(t) + s_{N-1}P_{N-1}(t). \end{cases} \tag{53}$$

In Sections 4, 5 and 6, we will solve this system of equations explicitly at steady-state for particular formation and separation kernels. We now derive a general formula for the steady-state probability

$$\Pi_K = \lim_{t\to\infty} P_K(t) \tag{54}$$

of having K clusters at steady-state and express it in terms of the a_i. The steady-state probabilities of the number of clusters are solution of the system

$$\begin{cases} 0 = -s_1\Pi_1 + f_2\Pi_2, \\ 0 = -(f_K + s_K)\Pi_K + f_{K+1}\Pi_{K+1} + s_{K-1}\Pi_{K-1}, \\ 0 = -f_N\Pi_N + s_{N-1}\Pi_{N-1}, \end{cases} \tag{55}$$

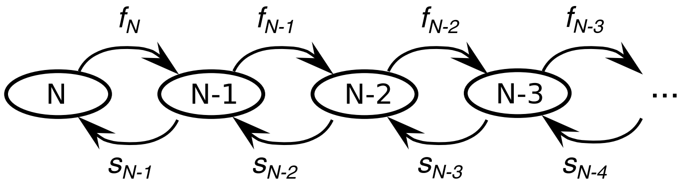

FIG. 2. *Markov chain representation for the number of clusters.* s_K (*resp.*, f_K) *is the separation* (*resp., formation*) *rate of a cluster when there are* K *clusters.*

with the normalization condition

$$\sum_{K=1}^{N} \Pi_K = 1. \tag{56}$$

The probabilities Π_K are given by the ratio

$$\frac{\Pi_K}{\Pi_{K-1}} = \frac{s_{K-1}}{f_K} \qquad \text{for } K \geq 2 \tag{57}$$

and the coefficients s_K and f_K are the mean-field separation and formation rates, respectively. Whereas the cluster configurations when the number of clusters is fixed depend only on the kernel a_i, the statistics of the number of clusters depend on the cluster fragmentation and coagulation rates F and C. In the following, we will focus on the coagulation condition $C(i, j) = 1$ and the fragmentation $F(i, j) = \frac{a_i a_j}{a_{i+j}}$ to state the following.

THEOREM 2.1. *When $C(i, j) = 1$ and $F(i, j) = \frac{a_i a_j}{a_{i+j}}$, the separation rate when there are K clusters is given by*

$$s_K = \frac{\sum_{i=1}^{N} \sum_{j=1}^{i-1} a_j a_{i-j} C_{N-i,K-1}}{C_{N,K}} \tag{58}$$

and the formation rate when there are K clusters is

$$f_K = \frac{K(K-1)}{2}. \tag{59}$$

PROOF. The total dissociation rate $d(n)$ of a cluster of size n is obtained by summing over all the possible sizes resulting from the dissociation and is given by

$$\begin{aligned} d(n) &= \sum_{i=1}^{n-1} F(i, n-i) \\ &= \sum_{i=1}^{n-1} \frac{a_i a_{n-i}}{a_n}. \end{aligned} \tag{60}$$

The rate at which a given configuration $(m_1, \dots, m_N) \in P'_{N,K}$ dissociates is $\sum_{i=1}^{N} d(i) m_i$. The separation rate for K clusters is thus

$$\begin{aligned} s_K &= \sum_{P'_{N,K}} \sum_{i=1}^{N} d(i) m_i p'(m_1, \dots, m_N) \\ &= \sum_{i=1}^{N} d(i) \langle M_i \rangle_{N,K}. \end{aligned} \tag{61}$$

Using relation (43), the separation rate becomes

$$s_K = \frac{\sum_{i=1}^{N} d(i) a_i C_{N-i,K-1}}{C_{N,K}}. \tag{62}$$

The separation rate of a distribution of K clusters equation (62) can thus be expressed as a function of the $C_{N,K}$:

$$s_K = \frac{\sum_{i=1}^{N} \sum_{j=1}^{i-1} a_j a_{i-j} C_{N-i,K-1}}{C_{N,K}}. \tag{63}$$

For $K = 1$, the only cluster is of size N and the separation rate is

$$s_1 = d(N). \tag{64}$$

The formation rates are given by

$$\begin{aligned} f_K = {} & \sum_{P'_{N,K}} \frac{1}{2}\left(\sum_{i=1}^{N} m_i(m_i - 1)C(i,i) + \sum_{i \neq j} m_i m_j C(i,j)\right) \\ & \times p'(m_1, \dots, m_N | K). \end{aligned} \tag{65}$$

For a distribution of K clusters, this is equal to the number of cluster pairs

$$f_K = \frac{K(K-1)}{2}. \tag{66}$$

□

We are now in position to study the statistics of the entire cluster configurations. Indeed, using Bayes' rule, the probability of a configuration $(m_1, \dots, m_N)$, that contains K clusters is the product of the conditional probability $p'(m_1, \dots, m_N|K)$ by the probability of having K clusters

$$p'(m_1, \dots, m_N, K) = p'(m_1, \dots, m_N|K)\Pi_K. \tag{67}$$

The mean number of clusters of size i is thus

$$\langle M_i \rangle_N = \sum_{K=1}^{N} \Pi_K \langle M_i \rangle_{N,K}. \tag{68}$$

3. Invariant of clusters dynamics. We introduce and compute here several measures of the cluster configurations that appear when the system is in a global steady-state. First, we compute the probability to find two particles in the same cluster, and second we measure two time scales associated to particle dynamics in an ensemble of clusters: (1) the mean time that two particles spend together before they separate and (2) the mean time that they spend separated before they meet again for the first time (Figure 3). The probability that two given particles are together is of interest in several cell biology examples: for instance, some genes can be silenced if the telomeres that carry them are forming a cluster [28].

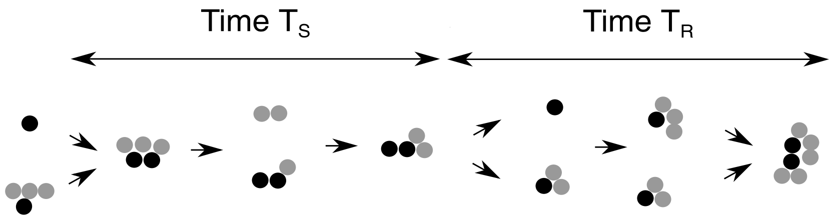

FIG. 3. *Cluster formation and dissociation. The time to separation T_S is the mean time for two specific particles (black) to spend in the same cluster. Before physical separation, the cluster can aggregate with other clusters (grey) or some particles except either of the two can be separated. The time to association T_R is the mean time the two particles meet again after separation for the first time.*

3.1. *The probability to find two particles in the same cluster.* When the mean number of clusters has reached its equilibrium, particles can still be exchanged between clusters. To characterize this exchange, we compute the probability to find two particles in the same cluster.

When the distribution of the clusters is $(n_1, \dots, n_N)$, the probability $P_2(n_1, \dots, n_N)$ to find two given particles in the same cluster is obtained by using the probability to choose the first particle in the cluster n_i, which is equal to the number of particles in the cluster divided by the total number of particles $\frac{n_i}{N}$. The probability to have the second particle in the same cluster is $\frac{n_i-1}{N-1}$. Summing over all possibilities, we get

$$P_2(n_1, \dots, n_N) = \sum_{i=1}^{K} \frac{n_i}{N} \frac{n_i - 1}{N - 1} = \frac{1}{N(N-1)} \left(\sum_{i=1}^{K} n_i^2 - N \right). \tag{69}$$

This probability is similar to Simpson's diversity index [29], a measure frequently used to quantify the diversity of ecosystems. We note that

$$\sum_{j=1}^{N} n_j^2 = \sum_{i=1}^{N} i^2 m_i. \tag{70}$$

Thus, when the distribution $(n_1, \dots, n_N)$ contains K clusters, we use $(n_1, \dots, n_K) \in P_{N,K}$ and obtain by summing over all configurations of K clusters

$$\sum_{(n_1,\dots,n_K) \in P_{N,K}} p(n_1, \dots, n_K | K) \sum_{j=1}^{K} n_j^2 = \sum_{(m_i) \in P'_{N,K}} p'(m_i) \sum_{j=1}^{N} j^2 m_j \tag{71}$$

$$= \sum_{j=1}^{N} j^2 \sum_{(m_i) \in P'_{N,K}} m_j p'(m_i) \tag{72}$$

$$= \sum_{j=1}^{N} j^2 \langle M_j \rangle_{N,K}, \tag{73}$$

where $\langle M_j \rangle_{N,K}$ is the mean number of clusters of size j, when there are N particles distributed in K clusters [equation (38)]. Taking into account all possible distributions of clusters, we obtain that the probability $\langle P_2 \rangle$ to find two particles in the same cluster is

$$\langle P_2 \rangle = \sum_{K=1}^{N} \sum_{(n_1,\dots,n_K)\in P_{N,K}} P_2(n_1,\dots,n_K)p(n_i)\Pi_K, \tag{74}$$

which can be written, using expressions (69) and (73) as

$$\langle P_2 \rangle = \frac{1}{N(N-1)} \sum_{K=1}^{N} \Pi_K \sum_{j=1}^{N} j^2 \langle M_j \rangle_{N,K} - \frac{1}{N-1}. \tag{75}$$

This approach can be generalized to the probability of having $n \geq 2$ particles together.

3.2. *Mean time for two particles to stay together in a cluster.* We define the mean time to separation (MTS) as the mean time between the arrival of two given particles in a cluster and their separation after dissociation of the cluster. We first compute the probability of a configuration $(m_1,\dots,m_N)$ conditioned on having two particles in the same cluster. We then derive the transition rates from this particular configuration to any of the states accessible by a single dissociation or association event (see Figure 1). The accessible states are divided into two classes: first, the configurations for which the particles are not initially in the same cluster (separated state) and second the ones where they are together. The current state is the configuration $(m_1,\dots,m_N)$, for which the particles are in the same cluster. Upon a coagulation event, the particles stay in the same cluster. A dissociation event can either occur for a cluster that did not contain the tracking particles or for the cluster that contained the particles. In the latter case, there are two possibilities: either the particles stay together after dissociation, or they are redistributed to two different clusters (separated state). The rate of dissociation from the ensemble $(m_1,\dots,m_N)$ to $(m_1,\dots,m_i+1,\dots,m_{k-i}+1,\dots,m_k-1,\dots,m_N)$ is $2F(i,k-i)m_k$ if $i \neq k/2$ and $F(\frac{k}{2},\frac{k}{2})m_k$ otherwise.

PROPOSITION 1. *The probability that two particles in a cluster of size k, given a configuration $(m_1,\dots,m_N) \in P'_{N,K}$ separate after a dissociation is*

$$p_S(k;m_1,\dots,m_N) = \frac{1}{k(k-1)} \frac{1}{\sum_{i=1}^{N} d(i)m_i} \sum_{i=1}^{k-1} i(2k-i)F(i,k-i). \tag{76}$$

PROOF. The probability that two particles in a cluster of size k separates after any dissociation event is equal to the product of the probability that the particles are in the cluster multiplied by the probability that this dissociation results in the effective separation of the particles. The first probability is obtained by considering the total dissociation rate $d(n)$ of a cluster of size n [see equation (60)]. The probability that the cluster containing the two particles effectively dissociates is thus proportional to $d(k)$ and is normalized by the total dissociation rate for the cluster configuration $(m_1, \dots, m_N)$:

$$\frac{d(k)}{\sum_{i=1}^{N} d(i) m_i}. \tag{77}$$

The probability of an effective separation P_{sep} is the complementary of the probability that the two particles stay together. When the two resulting clusters are of size i and $k - i$, this probability is equal to

$$P_{\text{sep}} = 1 - \frac{i(i-1) + (k-i)(k-i-1)}{k(k-1)}. \tag{78}$$

The probability that the two particles initially in a cluster of size k will be separated is thus equal to

$$\begin{aligned} p_S(k) &= \frac{d(k)}{\sum_{i=1}^{K} d(i) m_i} \sum_{i=1}^{k-1} \left(1 - \frac{i(i-1) + (k-i)(k-i-1)}{k(k-1)} \right) \\ &\quad \times \frac{F(i, k-i)}{d(k)} \\ &= \frac{1}{k(k-1)} \frac{1}{\sum_{i=1}^{N} d(i) m_i} \sum_{i=1}^{k-1} F(i, k-i) i (2k - i). \end{aligned} \tag{79}$$

□

The transition probability from the state $(m_1, \dots, m_N)$ to the separated one equals the sum of the probabilities $p_S(k)$ to be separated over all cluster sizes k,

$$P_S(m_1, \dots, m_N) = \sum_{k=1}^{K} p_S(k) m_k. \tag{80}$$

To derive the MTS, we first write the transition matrix of the Markov chain representing the transitions between the configurations of separated and nonseparated states, and second, we determine the mean transition times between the states.

Ordering the configurations $(m_1, \dots, m_N)$ by arbitrary indices I, we define the transition matrix T of size $(q(N) + 1) \times (q(N) + 1)$, where $q(N)$ is the total number of different partitions of the integer N. The elements of T of the first $q(N)$ rows and columns are the transition probabilities between states where the two particles are together, while index $q(N) + 1$ represents the separated state.

For $I, J \in (1, \dots, q(N))$, the matrix entries $[T]_{I,J}$ are either the coagulation or dissociation rates given above (see Figure 1), while the transitions $[T]_{I,q(N)+1}$ are equal to the rates to the separated state. Because the separated state is absorbing, we finally set $[T]_{q(N)+1,q(N)+1} = 1$. Additionally, the matrix is normalized into a stochastic matrix such that

$$\forall I, \qquad \sum_{J=1}^{q(N)+1} [T]_{I,J} = 1. \tag{81}$$

We now estimate the mean time the two particles stay together. The mean time $\tau(m_1, \dots, m_N)$ from a configuration $(m_1, \dots, m_N)$ to any configuration $(m'_1, \dots, m'_N)$ accessible by a single coagulation or fragmentation event is the reciprocal of the sum of all transition rates

$$\tau(m_1, \dots, m_N) = \left(\sum_{k=1}^{N} d(k) m_k + \frac{K(K-1)}{2}\right)^{-1}, \tag{82}$$

where $K = \sum_{i=1}^{N} m_i$ and the coagulation kernel is constant $C(i, j) = 1$. We represent the transition times in a vector $\boldsymbol{\tau}$

$$\boldsymbol{\tau} = \begin{bmatrix} \tau(N, 0, \dots, 0) \\ \tau(N-2, 1, \dots, 0) \\ \vdots \\ \tau(0, \dots, 0, 1) \end{bmatrix}. \tag{83}$$

We shall now compute the vector

$$\boldsymbol{t} = \begin{bmatrix} t(N, 0, \dots, 0) \\ t(N-2, 1, \dots, 0) \\ \vdots \\ t(0, \dots, 0, 1) \end{bmatrix} \tag{84}$$

which is the MTS for an ensemble of cluster configuration $(m_1, \dots, m_N)$: it is related to the vector $\boldsymbol{\tau}$ by [25]

$$\boldsymbol{t} = \sum_{n=0}^{\infty} T^n \boldsymbol{\tau}. \tag{85}$$

The vector $\boldsymbol{t}$ is computed using the matrices

$$A = I_{q(N)+1} - T \tag{86}$$

and

$$A^* = [A_{q(N)+1}]^{-1}, \tag{87}$$

where $A_{q(N)+1}$ is the matrix A from which the $q(N) + 1$ row and column were removed. Equation (85) is equivalent to

$$\boldsymbol{t} = A^* \boldsymbol{\tau}. \tag{88}$$

The MTS averaged over the equilibrium configuration distribution is

$$T_S = \boldsymbol{p}^{*T}\boldsymbol{t}, \tag{89}$$

where $\boldsymbol{p}^{*T} = [p^*(m_1, \dots, m_N)]_{(m_1,\dots,m_N)}$ and $p^*(m_1, \dots, m_N)$ is the probability distribution of the configuration $(m_1, \dots, m_N)$ when the two particles are in the same cluster. Using Bayes' theorem, the probabilities $p^*(m_1, \dots, m_N)$ are given by

$$p^*(m_1, \dots, m_N) = \frac{P_2(m_1, \dots, m_N)p(m_1, \dots, m_N)}{\langle P_2 \rangle}, \tag{90}$$

where we recall that $p(m_1, \dots, m_N)$ is the probability of the configuration $(m_1, \dots, m_N)$, $\langle P_2 \rangle$ is the probability that two specific particles are in the same cluster [see computation equation (75)], and $P_2(m_1, \dots, m_N)$ is the probability that the two particles are in the same cluster configuration $(m_1, \dots, m_N)$. The conditional probability $p^*(m_1, \dots, m_N)$ to select two particles in the same cluster, conditioned on the distribution $(m_1, \dots, m_N)$ is larger for clusters of larger sizes. We conclude this section with the following.

REMARK 4. The time that two particles spend separated T_R can be similarly determined, only the absorbing state in the transition matrix is the state at which the two particles are in the same cluster. Interestingly, the probability two particles are together is the fraction of time they spend together and is given by the ratio

$$\langle P_2 \rangle = \frac{T_S}{T_S + T_R}. \tag{91}$$

In the rest of the manuscript, we apply the previous analysis to three examples of coagulation-fragmentation with a finite number of particles.

4. Example 1: The case $a_i = a$.

We consider the case of a constant kernel $a_i = a$. To compute the separation and formation rates s_K and f_K, we use that $F(i, j) = a$ and $C(i, j) = 1$. This fragmentation kernel corresponds to the following model: a cluster of size n dissociates at a rate $\sum_{i=1}^{n-1} F(i, n-i) = (n-1)a$ and the sizes of the resulting clusters are uniformly distributed between 1 and $n-1$. When there are N particles and a total number of clusters K, the partition of clusters is denoted $(n_1, \dots, n_K) \in P_{N,K}$. The total transition rate from a configuration of K to $K+1$ clusters is the sum over all possible dissociation rates

$$s_K = \sum_{i=1}^{K}(n_i - 1)a = (N-K)a. \tag{92}$$

The formation rate is proportional to the number of pairs

$$f_K = \sum_{i=1}^{K-1}\sum_{j=i+1}^{K} C(n_i, n_j) = \frac{K(K-1)}{2}. \tag{93}$$

Following relation (53), the steady-state probability Π_K for the number of clusters of size K satisfies the time independent master equation

$$\begin{cases} s_1\Pi_1 = f_2\Pi_2, \\ (f_K + s_K)\Pi_K = f_{K+1}\Pi_{K+1} + s_{K-1}\Pi_{K-1}, \\ f_N\Pi_N = s_{N-1}\Pi_{N-1}, \end{cases} \tag{94}$$

which leads to the relation

$$\Pi_{K+1} = (2a)^K \frac{(N-1)!}{K!(K+1)!(N-K-1)!}\Pi_1. \tag{95}$$

Using the normalization condition $\sum_K \Pi_K = 1$, the probability Π_1 can be expressed as a hypergeometric series

$$\Pi_1 = \frac{1}{{}_1F_1(-N+1;2;-2a)}, \tag{96}$$

where

$${}_1F_1(a;b;z) = \sum_{n=0}^{\infty} \frac{(a)_n}{(b)_n}\frac{z^n}{n!}, \tag{97}$$

is Kummer's confluent hypergeometric function ([1], pages 503–535) and

$$(x)_n = x(x+1)\cdots(x+n-1) \tag{98}$$

is the Pochhammer symbol. The average number of clusters at steady-state

$$\begin{aligned} \mu_1(a) &= \sum_{K=1}^{N} K\Pi_K \\ &= \Pi_1 \frac{d}{dz}\big(z\,{}_1F_1(-N+1;2;z)\big)_{|z=-2a}. \end{aligned} \tag{99}$$

The derivative of the Kummer's function is

$$\frac{d}{dz}\,{}_1F_1(a;b;z) = \frac{a}{b}\,{}_1F_1(a+1;b+1;z). \tag{100}$$

Finally, the mean number of clusters is expressed as

$$\begin{aligned} \mu_1(a) &= 1 + a(N-1)\frac{{}_1F_1(-N+2;3;-2a)}{{}_1F_1(-N+1;2;-2a)} \\ &= 1 + a(N-1)G_1, \end{aligned} \tag{101}$$

where we note G_1 the function defined by

$$G_1 = \frac{{}_1F_1(-N+2;3;-2a)}{{}_1F_1(-N+1;2;-2a)}. \tag{102}$$

More generally, we introduce the functions G_i defined by

$$G_i = \frac{{}_1F_1(-N+1+i; 2+i; -2a)}{{}_1F_1(-N+1; 2; -2a)}. \tag{103}$$

Following the procedure presented above, all moments of the probability distribution Π_K can be computed and the nth-order moment μ_n is expressed using the operator H defined by

$$H(f)(z) = \frac{d}{dz} z f(z), \tag{104}$$

by

$$\mu_n = \sum_{n=1}^{N} K^n \Pi_K = \frac{H^{(n)}({}_1F_1(-N+1; 2; z))_{|z=-2a}}{{}_1F_1(-N+1; 2; -2a)}. \tag{105}$$

Using the differentiation formula for the hypergeometric function (100), the moments μ_n can be written as

$$\mu_n = \sum_{k=0}^{n} \alpha_k^n \frac{(N-1)!}{(k+1)!(N-1-k)!} 2^k a^k G_k \tag{106}$$

$$= \sum_{k=0}^{n} \alpha_k^n \frac{\Pi_{k+1}}{\Pi_1} G_k, \tag{107}$$

where the coefficients α_k^n are given by

$$\alpha_k^n = \begin{cases} k! \sum_{j=0}^{k/2} (-1)^j \frac{(k+1-j)^n + (j+1)^n}{(k-j)!} & \text{if } k \text{ is even,} \\ k! \sum_{j=0}^{(k-1)/2} (-1)^j \frac{(k+1-j)^n - (j+1)^n}{(k-j)!} & \text{if } k \text{ is odd,} \end{cases}$$

and $\alpha_0^n = \alpha_n^n = 1$. We can thus obtain the variance of the number of clusters, given by

$$\begin{aligned} \langle V_\infty(a) \rangle &= \mu_2 - \mu_1^2 \\ &= a(N-1)G_1(a,N) \\ &\quad + \frac{2}{3}a^2(N-1)(N-2)G_2(a,N) \\ &\quad - a^2(N-1)^2 G_1^2(a,N). \end{aligned} \tag{108}$$

4.1. *Asymptotic formulas for the mean and variance of the cluster number.* We provide here approximations for the functions G_n defined in equation (103). Kummer's function can be expressed in terms of generalized Laguerre polynomials

$$
{}_1F_1(-n;b;z) = \frac{n!}{(b)_n} L_n^{b-1}(z), \tag{109}
$$

where n is an integer. The asymptotic behavior of Laguerre polynomials for large n, fixed $x > 0$ and α, is given by [30]

$$
\begin{aligned}
L_n^{\alpha}(-x) = {} & \frac{n^{\frac{\alpha}{2}-\frac{1}{4}}}{2\sqrt{\pi}} \frac{e^{-\frac{x}{2}}}{x^{\frac{\alpha}{2}+\frac{1}{4}}} \exp\biggl(2\sqrt{x\biggl(n+\frac{\alpha+1}{2}\biggr)}\biggr) \\
& \times \Biggl(1 + \sum_{\nu=1}^{m/2} C_\nu(x) n^{-\nu/2} + O\bigl(n^{-m/2}\bigr)\Biggr),
\end{aligned}
$$

where m is a positive integer and $C_\nu(x)$ is a regular and bounded functions for $x > 0$, independent of N. Using the leading order term in the expansion for $m = 2$, we get

$$
\begin{aligned}
L_n^{\alpha}(-x) = {} & \frac{n^{\frac{\alpha}{2}-\frac{1}{4}}}{2\sqrt{\pi}} \frac{e^{-\frac{x}{2}}}{x^{\frac{\alpha}{2}+\frac{1}{4}}} \exp\biggl(2\sqrt{x\biggl(n+\frac{\alpha+1}{2}\biggr)}\biggr) \\
& \times \bigl(1 + C_1(x) n^{-1/2} + O\bigl(n^{-1}\bigr)\bigr).
\end{aligned} \tag{110}
$$

We can now evaluate G_1 by using the asymptotic expansion for ${}_1F_1(-N + 1; 2; -2a)$ and ${}_1F_1(-N+2;3;-2a)$. For large N, we have

$$
\begin{aligned}
{}_1F_1(-N+1;2;-2a) = {} & \frac{1}{N} \frac{(N-1)^{1/4}}{2\sqrt{\pi}} \frac{e^{-a}}{(2a)^{3/4}} \exp(2\sqrt{2aN}) \\
& \times \biggl(1 + O\biggl(\frac{1}{\sqrt{N}}\biggr)\biggr)
\end{aligned} \tag{111}
$$

and

$$
\begin{aligned}
{}_1F_1(-N+2;3;-2a) = {} & \frac{1}{N(N-1)} \frac{(N-2)^{3/4}}{2\sqrt{\pi}} \frac{e^{-a}}{(2a)^{5/4}} \\
& \times \exp\bigl(2\sqrt{2a(N-1/2)}\bigr)\biggl(1 + O\biggl(\frac{1}{\sqrt{N}}\biggr)\biggr).
\end{aligned} \tag{112}
$$

Finally,

$$
\begin{aligned}
G_1(a,N) = {} & \sqrt{\frac{2a}{N}} \exp\bigl(2\sqrt{2a(N-1/2)} - 2\sqrt{2aN}\bigr)\biggl(1 + O\biggl(\frac{1}{\sqrt{N}}\biggr)\biggr) \\
\approx {} & \sqrt{\frac{2}{aN}} \exp\biggl(-\sqrt{\frac{a}{2N}}\biggr) = \tilde{G}_1(a,N).
\end{aligned} \tag{113}
$$

Similarly, the present computation can be generalized to obtain the asymptotic approximation $\tilde{G}_n$ for the functions G_n, valid for large N and fixed n,

$$\tilde{G}_n(a,N) \approx \frac{(n+1)!}{(2aN)^{n/2}} \exp\left(-n\sqrt{\frac{a}{2N}}\right). \tag{114}$$

In Figure 4, we compare the exact value of G_1 [defined in (102)] with the approximation $\tilde{G}_1$ given by expression (113). The approximation is more accurate for intermediate values of a [see Figure 4(B)]. In addition, the error function $\frac{|G_1-\tilde{G}_1|}{G_1}$ has a discontinuity in the derivative for $a = 10$. Indeed, at $a = 10$, the function G_1 and $\tilde{G}_1$ cross each other, and thus $|G_1 - \tilde{G}_1|$ has a singular derivative, shown by a cusp type behavior in Figure 4(B). Moreover, the function G_1 is always decreasing

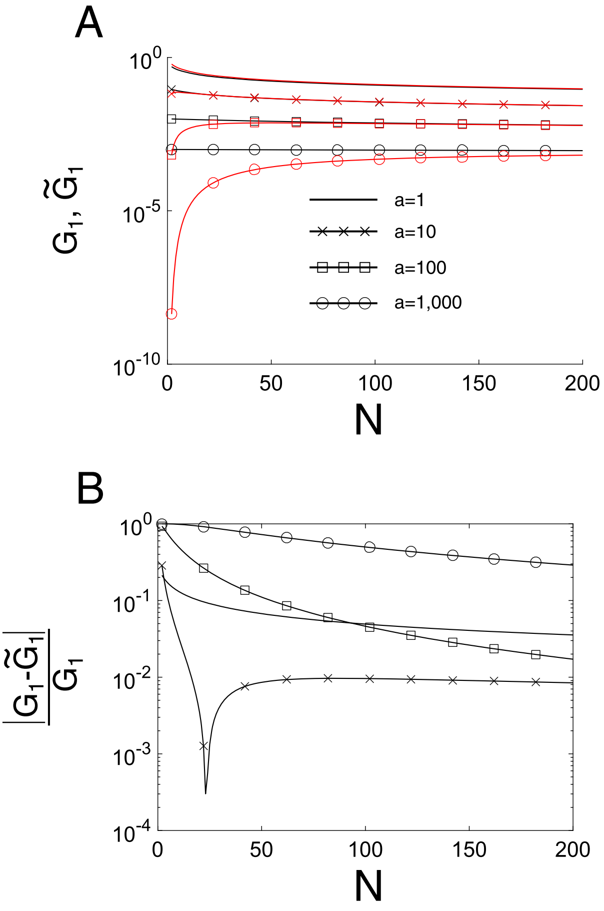

FIG. 4. *Approximation of* G_1. (A) *Plot of* G_1 (*black*) *and approximation* $\tilde{G}_1$ [*red, equation* (113)] *versus* N *for* $a = 1, 10, 100$ *and* 1000. (B) *Comparison of the values of* G_1 *and* $\tilde{G}_1$, *measured as* $\frac{|G_1-\tilde{G}_1|}{G_1}$.

with N, however, its approximation $\tilde{G}_1$ is nonmonotonic [Figure 4(A)]. Finally, by using the asymptotic expression (113), we find the approximation of the number of clusters for large N,

$$\mu_1(a) \approx 1 + \sqrt{2aN} \exp\biggl(-\sqrt{\frac{a}{2N}}\biggr). \tag{115}$$

To obtain a numerical approximation of G_1 for small a, we used the finite continued fraction decomposition for the ratio of hypergeometric functions ${}_1F_1$ [19], and obtain

$$\begin{aligned} G_1(a, N) &= \frac{{}_1F_1(-N+2; 3; -2a)}{{}_1F_1(-N+1; 2; -2a)} \\ &= \cfrac{1}{1 + \cfrac{(N+1)a/3}{1+\cfrac{(N-2)a/6}{1+\cfrac{(N+2)a/10}{1+\cfrac{(N-3)a/15}{\ddots + \frac{a/(N-1)(2N-3)}{1+a/(N-1)}}}}}}. \end{aligned} \tag{116}$$

We obtain the Taylor expansion of G_1 for small a from the continued fraction,

$$\begin{aligned} &G_1(a, N) \\ &\quad = 1 - \frac{N+1}{3}a + \biggl(\frac{(N+1)(N-2)}{18} + \frac{(N+1)^2}{9}\biggr)a^2 \\ &\qquad - \biggl(\frac{(N+1)(N-2)(N+2)}{180} + \frac{(N+1)(N-2)^2}{108} + \frac{(N+1)^3}{27}\biggr)a^3 \\ &\qquad + o(a^3). \end{aligned} \tag{117}$$

This expression is computed from the continued fraction G_1 written

$$\begin{aligned} G_1(a) &= \frac{1}{1 + F_1(a)a} \\ &= \frac{1}{1 + \frac{f_1 a}{1+F_2(a)}}, \end{aligned} \tag{118}$$

where the general term of the sequence F_n is

$$F_n(a) = \frac{f_n}{1 + F_{n+1}(a)a} \tag{119}$$

with $F_{n+1}(a) = O(1)$. G_1 can also be written as

$$G_1(a) = \cfrac{1}{1 + \cfrac{af_1}{1+\cfrac{af_2}{1+\frac{\cdots}{1+af_n}}}}. \tag{120}$$

To obtain a third-order Taylor expansion of G_1, we simply used F_1, F_2, F_3 and we have truncated the rest of continued fraction to obtain

$$G_1(a) = 1 - F_1(a)a + F_1^2(a)a^2 - F_1^3(a)a^3 + o(a^3). \tag{121}$$

The expansion of F_1 is

$$\begin{aligned} F_1(a) &= \frac{f_1}{1 + F_2(a)a} \\ &= f_1\bigl(1 - F_2(a)a + F_2^2(a)a^2\bigr) + o(a^2) \\ &= f_1\bigl(1 - f_2(1 - f_3a)a + f_2^2(a)a^2\bigr) + o(a^2) \\ &= f_1 - f_1f_2a + (f_1f_2f_3 + f_1f_2^2)a^2 + o(a^2). \end{aligned} \tag{122}$$

Using F_1 expansion into G_1, we finally get

$$G_1(a) = 1 - f_1a + (f_1^2 + f_1f_2)a^2 - (f_1f_2f_3 + f_1f_2^2 + f_1^3)a^3 + o(a^3). \tag{123}$$

4.2. *Statistics for the number of clusters of a given size.* We now compute several moments for the size of clusters when there are N particles. Using relation (38), we first obtain the expression for the mean number of clusters of size n when there are K clusters

$$\begin{aligned} \langle M_n \rangle_{N,K} &= \sum_{(m_i)\in P'_{N,K}} m_n p'(m_i|K) \\ &= a\frac{C_{N-n,K-1}}{C_{N,K}}. \end{aligned} \tag{124}$$

Determining this relation requires computing the normalizing constant $C_{N,K}$ given in equation (20). The normalizing constant $C_{N,K}$ is the Nth-order coefficient of S^K, where S is the generating function (21)

$$S(x) = \sum_{i=1}^{\infty} a_i x^i = a\frac{x}{1-x}. \tag{125}$$

The coefficient $C_{N,K}$ is thus equal to the $(N-K)$th-order coefficient of $\frac{1}{K!}\frac{a^K}{(1-x)^K}$ (Remark 1). We obtain this coefficient by differentiating $N-K$ times $\frac{1}{(1-x)^K}$ and estimating the derivative at $x=0$. We obtain that

$$\begin{aligned} C_{N,K} &= \frac{1}{K!}a^K\frac{1}{(N-K)!}K(K+1)\cdots(K+N-K-1) \\ &= \frac{a^K}{K!}\frac{(N-1)!}{(K-1)!(N-K)!}. \end{aligned} \tag{126}$$

Thus, by combining (124) and (126), we obtain that the number of cluster of size n, when the N particles are distributed in K clusters is

$$\langle M_n \rangle_{N,K} = \frac{(N-n-1)!K!(N-K)!}{(N-1)!(K-2)!(N-n-K+1)!}, \tag{127}$$

which we remark to be independent of a. The mean number of clusters of size n is obtained by summing over all possible configurations with K clusters,

$$\begin{aligned} \langle M_n \rangle &= \sum_{K=1}^{N} \langle M_n \rangle_{N,K} \Pi_K \\ &= \frac{(N-n-1)!}{(N-1)!} \sum_K \frac{K(K-1)(N-K)!}{(N-n-K+1)!} \Pi_K. \end{aligned} \tag{128}$$

Using expression (95) for Π_K, we obtain

$$\langle M_n \rangle = 2a \frac{{}_1F_1(-N+1+n; 2; -2a)}{{}_1F_1(-N+1; 2; -2a)} \qquad \text{if } n < N \tag{129}$$

and

$$\langle M_N \rangle = \frac{1}{{}_1F_1(-N+1; 2; -2a)}. \tag{130}$$

The mean number of clusters of size N is exactly equal to the probability $\Pi_1(N)$ of having one cluster when there is N particles [see equation (96)]. Indeed this is the only configuration where a cluster of size N can appear. The mean number of clusters of size n is $2a\frac{\Pi_1(N)}{\Pi_1(N-n)}$, which means that it is given by the ratio of the probability of having one cluster when there are N particles over the probability of having one cluster when there are $N-n$ particles.

The number of clusters can also be written using the function G_k defined in (103),

$$\langle M_n \rangle = \sum_{k=0}^{n} (-1)^k \frac{(2a)^{k+1}}{(k+1)!} \frac{n!}{(n-k)!k!} G_k \tag{131}$$

$$= 2a \sum_{k=0}^{n} (-1)^k \frac{\Pi_{k+1}(n)}{\Pi_1(n)} G_k, \tag{132}$$

where $\Pi_k(n)$ is the probability of having k clusters in a system of n particles. To summarize this analysis, we plotted in Figure 5 the mean number of clusters of size n for $N = 5$ particles.

4.3. *Probability to find two particles in the same cluster.* We now evaluate the probability to find two particles in the same cluster for a constant kernel. When there are N particles, we first prove the following.

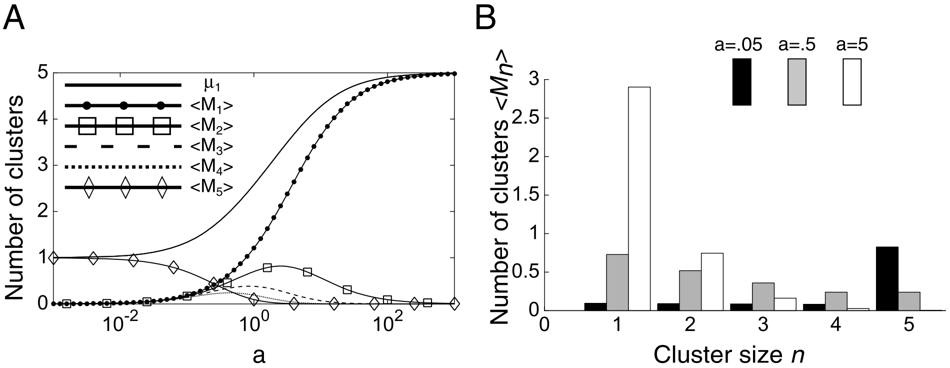

FIG. 5. (A) *Mean number of clusters of size n as a function of the parameter a for $N = 5$ particles* [*equation* (130)], *and total number of clusters* μ_1. (B) *Mean number of clusters* $\langle M_n \rangle$ *as a function of the cluster size n*, *for* $a = 0.05$, $a = 0.5$ *and* $a = 5$.

LEMMA 1.

$$\sum_{j=1}^{N} j^2 \langle M_j \rangle_{N,K} = N + 2N \frac{N-K}{K+1}, \tag{133}$$

where

$$\langle M_j \rangle_{N,K} = \frac{(N-j-1)!K!(N-K)!}{(N-1)!(K-2)!(N-j-K+1)!} \tag{134}$$

is given by relation (127).

PROOF. Using formula (127), we have

$$\begin{aligned} \sum_{j=1}^{N} j^2 \langle M_j \rangle_{N} &= \sum_{j=1}^{N} j^2 \frac{(N-j-1)!K!(N-K)!}{(N-1)!(K-2)!(N-j-K+1)!} \\ &= \frac{K(K-1)(N-K)!}{(N-1)!} \\ &\quad \times \sum_{j=1}^{N-K+1} j^2 \frac{(N-j-1)!}{(N-j-K+1)!}. \end{aligned} \tag{135}$$

To determine the sum in equation (135), we introduce the sum

$$S_{N,K} = \sum_{j=1}^{N-K+1} j^2 \frac{(N-j-1)!}{(N-j-K+1)!}, \tag{136}$$

and prove now that

$$S_{N,K} = \frac{N!}{(N-K)!} \frac{2N-K+1}{(K-1)K(K+1)}. \tag{137}$$

We first obtain a recurrence relation between $S_{N+1,K}$ and $S_{N,K}$

$$\begin{aligned} S_{N+1,K} &= \sum_{j=1}^{N-K+2} j^2(N-j)(N-1-j)\cdots(N-K+3+j) \\ &= \sum_{j=0}^{N-K+1} (j+1)^2(N-j-1)(N-j-2)\cdots(N-K+2+j) \\ &= \frac{(N-1)!}{(N-K+1)!} \\ &\quad + \sum_{j=1}^{N-K+1} \big(j^2+2j+1\big)(N-j-1)(N-j-2)\cdots(N-K+2+j) \\ &= \frac{(N-1)!}{(N-K+1)!} + S_{N,K} + A + 2B, \end{aligned} \tag{138}$$

where

$$A = \sum_{j=1}^{N-K+1} (N-j-1)(N-j-2)\cdots(N-K+2+j) \tag{139}$$

and

$$B = \sum_{j=1}^{N-K+1} j(N-j-1)(N-j-2)\cdots(N-K+2+j). \tag{140}$$

We first use a change of variable for A and find

$$\begin{aligned} A &= \sum_{j=K-2}^{N-2} j(j-1)\cdots(j-K+3) \\ &= \sum_{j=K-2}^{N-2} \frac{j!}{(j-K+2)!} = (K-2)! \sum_{j=K-2}^{N-2} \binom{j}{K-2}. \end{aligned} \tag{141}$$

Using the binomial relation

$$\sum_{j=k}^{n} \binom{j}{k} = \binom{n+1}{k+1}, \tag{142}$$

we obtain

$$A = (K-2)!\binom{N-1}{K-1} = \frac{(N-1)!}{(N-K)!(K-1)}. \tag{143}$$

Similarly, we obtain for B

$$\begin{aligned} B &= \sum_{j=K-2}^{N-2} (N-j-1)\frac{j!}{(j-K+2)!} \\ &= NA - \sum_{j=K-2}^{N-2} \frac{(j+1)!}{(j-K+2)!} \\ &= NA - (K-1)! \sum_{j=K-1}^{N-1} \binom{j}{K-1} \\ &= \frac{N!}{(N-K)!(K-1)} - \frac{N!}{(N-K)!K} \\ &= \frac{N!}{(N-K)!(K-1)K}. \end{aligned} \tag{144}$$

Finally, we obtain the induction relation for $S_{N,K}$

$$\begin{aligned} S_{N+1,K} &= S_{N,K} + 2\frac{N!}{(N-K)!(K-1)K} + \frac{(N-1)!}{(N-K)!(K-1)} \\ &\quad + \frac{(N-1)!}{(N-K+1)!} \\ &= S_{N,K} + \frac{N!(2N-K+2)}{(N-K+1)!(K-1)K}. \end{aligned} \tag{145}$$

Using that $S_{K,K} = (K-2)!$, we finally evaluate the sum

$$\begin{aligned} S_{N,K} &= \frac{1}{(K-1)K} \sum_{j=K}^{N} \frac{(j-1)!(2j-K)}{(j-K)!} \\ &= \frac{2}{(K-1)K} \sum_{j=K}^{N} \frac{j!}{(j-K)!} - \frac{1}{(K-1)!} \sum_{j=K}^{N} \frac{(j-1)!}{(j-K)!} \\ &= 2(K-2)! \sum_{j=K}^{N} \binom{j}{K} - (K-2)! \sum_{j=K-1}^{N-1} \binom{j}{K-1} \end{aligned} \tag{146}$$

$$
\begin{aligned}
&= 2(K-2)!\binom{N+1}{K+1} - (K-2)!\binom{N}{K} \\
&= \frac{(K-2)!}{(N-K)!}\left(2\frac{(N+1)!}{(K+1)!} - \frac{N!}{K!}\right) \\
&= \frac{N!}{(N-K)!}\frac{2N-K+1}{(K-1)K(K+1)},
\end{aligned}
$$

which is formula (133). □

THEOREM 4.1. *The probability to find two particles in the same cluster is*

$$\langle P_2\rangle = G_1, \tag{147}$$

where G_1 is defined in (102).

PROOF. Using equation (75) and Lemma 1, we can now compute the probability to find two particles in the same cluster

$$
\begin{aligned}
\langle P_2\rangle &= \frac{1}{N(N-1)}\sum_{K=1}^{N}\Pi_K\sum_{j=1}^{N} j^2\langle M_j\rangle_{N,K} - \frac{1}{N-1} \\
&= \frac{1}{N(N-1)}\sum_{K=1}^{N}\Pi_K\left(N + 2N\frac{N-K}{K+1}\right) - \frac{1}{N-1}.
\end{aligned}
\tag{148}
$$

Thus,

$$
\begin{aligned}
\langle P_2\rangle &= \frac{2}{N-1}\sum_{K=1}^{N}\Pi_K\frac{N-K}{K+1} \\
&= -\frac{2}{N-1} + 2\frac{N+1}{N-1}\sum_{K=1}^{N}\frac{1}{K+1}\Pi_K.
\end{aligned}
\tag{149}
$$

Following formula (99), the sum in equation (149) can be expressed by integrating the Kummer's function

$$\sum_{K=1}^{N}\frac{1}{K+1}\Pi_K = \frac{\Pi_1}{2}\frac{\int z\,{}_1F_1(-N+1;2;z)\,dz}{z^2}\bigg|_{z=2a}. \tag{150}$$

Integrating the hypergeometric series gives

$$\int z\,{}_1F_1(-N+1;2;z)\,dz = z^2{}_2F_2(-N+1,2;2,3;z), \tag{151}$$

which leads to

$$\sum_{K=1}^{N}\frac{1}{K+1}\Pi_K=\frac{1}{2}\frac{{}_2F_2(-N+1,2;2,3;-2a)}{{}_1F_1(-N+1;2;-2a)}\tag{152}$$
$$=\frac{1}{2}\frac{{}_1F_1(-N+1;3;-2a)}{{}_1F_1(-N+1;2;-2a)}.$$

By combining equations (149) and (152), we obtain

$$\langle P_2\rangle=-\frac{2}{N-1}+\frac{N+1}{N-1}\frac{{}_1F_1(-N+1;3;-2a)}{{}_1F_1(-N+1;2;-2a)}.\tag{153}$$

The three-term recurrence relation for Kummer's function ([1] equations 13.4.1–13.4.6) gives

$${}_1F_1(-N+1;3;-2a)=\frac{N-1}{N+1}{}_1F_1(-N+2;3;-2a)\tag{154}$$
$$+\frac{2}{N+1}{}_1F_1(-N+1;2;-2a).$$

Finally, using equation (102), we obtain

$$\langle P_2\rangle=G_1.\tag{155}$$ □

To finish this section, we note that the large N asymptotic of the probability that two particles are in the same cluster is

$$\langle P_2\rangle\approx\sqrt{\frac{2}{aN}}.\tag{156}$$

Many results presented in this section can be used to study the distribution of clusters in biological systems such as telomere organization in yeast. We provided here the explicit derivations of the exact and asymptotic formulas that can be used to analyze experimental and simulation results [14].

5. Example 2: The case $a_i = a$ for $i < M$ and $a_i = 0$ if $i \geq M$.

In this section, we consider N particles that can associate or dissociate at a constant rate, but in addition they cannot form clusters of more than M particles. The configuration space for distributions of N particles in K clusters of size less than M is

$$P'_{N,K,M}=\left\{(m_i)_{1\le i\le M};\sum_{i=1}^{M}im_i=N,\sum_{i=1}^{M}m_i=K\right\}.\tag{157}$$

First, the minimal number of clusters is necessarily bounded by $K \geq N/M$, since the opposite would imply a cluster of at least $M+1$ particles. The probability of a configuration $(m_1,\dots,m_M)\in P'_{N,K,M}$ is equal to

$$\Pr\{(m_1,\dots,m_M)\in P'_{N,K,M}\}=\frac{1}{C_{N,K,M}}\frac{1}{m_1!\cdots m_M!},\tag{158}$$

where the normalization constant $C_{N,K,M}$ is the Nth-order coefficient of

$$(aX+aX^2+\cdots+aX^M)^K = a^K X^K\left(\frac{X^M-1}{X-1}\right)^K$$

$$= a^K\left(\frac{X}{1-X}\right)^K \sum_{n=0}^{K}\binom{K}{n}(-1)^n X^{nM} \tag{159}$$

$$= a^K \frac{1}{(1-X)^K}\sum_{n=0}^{K}\binom{K}{n}(-1)^n X^{nM+K}.$$

Then the Nth-order coefficient of the polynomial is obtained by finding the $(N-nM-K)$th-order coefficient of $(1-X)^{-K}$

$$\begin{aligned} C_{N,K,M} = a^K \sum_{n=0}^{K}\binom{K}{n}(-1)^n \frac{1}{(N-(nM+K))!}D^{(N-(nM+K))} \\ \times\left(\frac{1}{(1-X)^K}\right)_{|X=0}, \end{aligned} \tag{160}$$

where we write $D^{(n)}(f)$ as the nth-order derivative of some function f. Thus, setting $K_0 = \lfloor\frac{N-K}{M}\rfloor$, where $\lfloor\cdot\rfloor$ is the floor function, we have

$$C_{N,K,M} = a^K K\sum_{n=0}^{K_0}\frac{(N-nM-1)!}{n!(K-n)!(N-(nM+K))!}(-1)^n. \tag{161}$$

For $M=N$, we find $K_0=0$ and the normalization constant

$$C_{N,K,N} = a^K\frac{(N-1)!}{(K-1)!(N-K)!}, \tag{162}$$

is equal to the normalization constant $C_{N,K}$ obtained for the constant kernel in Section 4.

The mean number of clusters of size $i\le M$ conditioned on the number of clusters K is

$$\begin{aligned} \langle M_i\rangle_K &= \sum_{m_i\in P'_{N,K,M}} m_i p(m_1,\dots,m_M) \\ &= a\frac{C_{N-i,K-1,M}}{C_{N,K,M}}. \end{aligned} \tag{163}$$

To find the probability to have K clusters, we now redefine the formation rate. In Section 4, the formation rate was proportional to the number of pairs of particles since all of them could form a new cluster. In the present case, two clusters of size

i and j can form a new cluster only if $i + j \leq M$. The formation rate when there are K clusters is thus

$$f_K = \sum_{(m_i) \in P'_{N,K,M}} p(m_1, \ldots, m_N) \times \left(\sum_{i=1}^{M/2} \frac{m_i(m_i - 1)}{2} + \sum_{\substack{i,j=1 \\ i+j \leq M; i \neq j}}^{M} m_i m_j \right). \tag{164}$$

The formation rate can be written as a function of the coefficients $C_{N,K,M}$ as

$$f_2 = C_{N,2,M}, \tag{165}$$

and for $K > 2$

$$f_K = \frac{K(K-1)}{2} \sum_{i=1}^{\min(\frac{M}{2}, \frac{N-K+2}{2})} C_{N-2i,K-2,M} + \frac{K(K-1)}{2} \sum_{\substack{i,j=1 \\ i+j \leq M}}^{\min(M-1,N-K+1)} C_{N-i-j,K-2,M}. \tag{166}$$

The separation rate remains unchanged $s_K = (N - K)a$, and the probabilities at steady-state are given by

$$\Pi_K = \frac{f_{K+1}}{s_K} \Pi_{K+1}. \tag{167}$$

A simple expression is certainly hopeless, but the limit $a \to 0$ is informative: contrary to the previous case with a constant kernel ($M = N$), where the particles form a single cluster of size N, the present system contains multiple steady-state distributions. The clusters grow independently and reach either their limit size M or are configured such that the sum of the sizes of each pair of cluster is larger than M. All possible configurations contain exactly $\lceil N/M \rceil$ clusters, where $\lceil . \rceil$ is the ceiling function.

We illustrate the limit case $a \to 0$ for $N = 9$, $M = 4$ (Figure 6). Because $a > 0$, all partitions are accessible, but as $a \to 0$, the steady-state configurations are dominated by the configurations with the largest possible cluster size (4, 4, 1), (4, 3, 2) and (3, 3, 3). Applying formulas (161) and (163), we obtain the limit cluster configuration probabilities

$$\begin{aligned} p(4,4,1) &= \frac{3}{10}, \\ p(4,3,2) &= \frac{6}{10}, \\ p(3,3,3) &= \frac{1}{10}. \end{aligned} \tag{168}$$

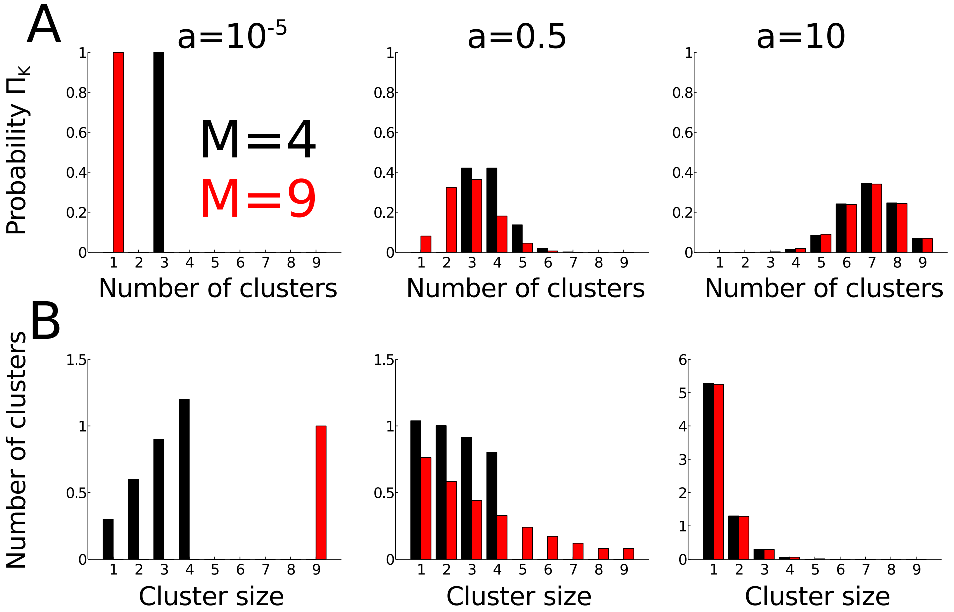

FIG. 6. (A) *Distribution of the number of clusters* Π_K *for* $N=9$, *when cluster sizes are limited* ($M=4$, *black*) *and not limited* ($M=9$, *red*). *There is a minimum of* $\lceil N/M \rceil$ *clusters. From left to right* : $a=10^{-5}$, $a=0.5$, $a=10$. (B) *Mean number of clusters of each size* $\langle M_n \rangle$. *For* $a \to 0$, *for* $N=9$ *and* $M=4$ *the clusters organize in three different cluster configurations, while for* $M=N$ *a single cluster containing N particles is formed.*

These steady-state probabilities do not depend on the initial particles configurations as long as $a \neq 0$. For $a=0$, there are three possible configurations $(4,4,1)$, $(4,3,2)$ and $(3,3,3)$: once equilibrium is attained, the clusters will remain unchanged. The probability to get to equilibrium depends on the configuration and the order of clustering events. When there is no limitation in the cluster formation ($M=N=9$), a single cluster containing all particles is formed (Figure 6, left panel). For large values of a, most clusters are very small, and the distributions are similar for $M=4$ and $M=9$ (Figure 6, right panel).

The probability for two particles to be in the same cluster provides a good estimation for the cluster distribution for various values of the parameter a (Figure 7). When a is large, most particles are contained in very small clusters and the probability $\langle P_2 \rangle$ is similar for the cases $M=4$ and $M=9$. When $a \to 0$, particles tend to form larger clusters. A single cluster containing all particles is formed and $\langle P_2 \rangle \to 1$ when $M=9$, but the maximal value of $\langle P_2 \rangle$ is less than 1 when the maximal cluster size is limited. We can explicitly compute $\langle P_2 \rangle$ in the limit case $a \to 0$. For example, for $M=4$, using equation (69), and summing over all possi-

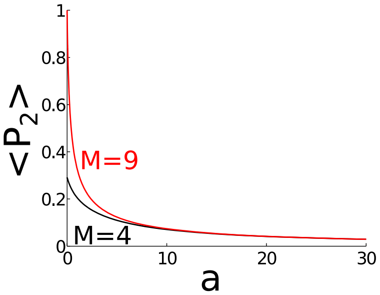

FIG. 7. *Probability $\langle P_2\rangle$ that two particles are in the same cluster. The parameters are $N = 9$ and $M = 4$ (black), $M = 9$ (red). For large values of $a \gg 1$, only small clusters are present and the steady-state distributions are similar for the cases $M = 4$ and $M = 9$. When $a \to 0$, the clusters organize in three different cluster configurations, while for $M = N$ a single cluster containing N particles is formed.*

ble configurations (168), we obtain

$$\begin{aligned}\langle P_2\rangle &= p(4,4,1)P_2(4,4,1) + p(4,3,2)P_2(4,3,2) + p(3,3,3)P_2(3,3,3)\\ &= \frac{3}{10}\frac{24}{72} + \frac{6}{10}\frac{20}{72} + \frac{1}{10}\frac{18}{72}\\ &= \frac{7}{24}.\end{aligned}$$

6. Example 3: Application to the case $a_i = ai$. We finally consider the case $a_i = ai$. It has been shown [12] that the number of clusters of size i is asymptotically

$$\langle M_i\rangle = aie^{-i\sqrt{2a/N}}. \tag{169}$$

The generating function S [equation (21)] is given by

$$S(x) = a\frac{x}{(1-x)^2}, \tag{170}$$

which gives that

$$\begin{aligned}C_{N,K} &= a^K\frac{1}{(N-K)!}D^{N-K}\frac{1}{(1-x)^{2K}}_{|x=0}\\ &= a^K\frac{1}{(N-K)!}\frac{(N+K-1)!}{(2K-1)!}\\ &= a^K\binom{N+K-1}{N-K}.\end{aligned} \tag{171}$$

We thus obtain using formula (43) that

$$\langle M_i \rangle_{N,K} = i \frac{\binom{N-i+K-2}{N-K-i+1}}{\binom{N+K-1}{N-K}}. \tag{172}$$

To obtain the number of clusters of size i, we determine the probability of a distribution of K clusters Π_K. We consider the coagulation kernel $C(i,j) = 1$ and the fragmentation kernel $F(i,j) = a\frac{ij}{i+j}$, and obtain that

$$d(n) = \sum_{i=1}^{n-1} a \frac{i(n-i)}{n} = \frac{a(n^2-1)}{6}. \tag{173}$$

The separation rates are

$$s_1 = \frac{a(N^2-1)}{6} \tag{174}$$

and for $K \geq 2$

$$\begin{aligned} s_K &= \frac{\sum_{i=1}^{N-K+1} d(i) a_i C_{N-i,K-1}}{C_{N,K}} \\ &= \frac{a}{6} \frac{\sum_{i=1}^{N-K+1} i(i^2-1)\binom{N-i+K-2}{N-i-K+1}}{\binom{N+K-1}{N-K}} \\ &= \frac{a}{6} \frac{1}{\binom{N+K-1}{N-K}} \frac{1}{(2K-3)!} \sum_{i=1}^{N-K+1} \frac{i(i^2-1)(N-i+K-2)!}{(N-i-K+1)!}. \end{aligned} \tag{175}$$

To go further in the determination of s_K, we evaluate the sum

$$\varphi_{N,K} = \frac{1}{(2K-3)!} \sum_{i=1}^{N-K+1} \frac{(i^3-i)(N-i+K-2)!}{(N-i-K+1)!}. \tag{176}$$

After the change of variables $j = N-i-K+1$ in the sum, we obtain

$$\begin{aligned} \varphi_{N,K} &= \frac{1}{(2K-3)!} \sum_{j=0}^{N-K} \bigl((N-K-j+1)^3 - (N-K-j+1)\bigr) \\ &\qquad \times \frac{(j+2K-3)!}{j!} \\ &= \sum_{j=0}^{N-K} \bigl((N-K-j+1)^3 - (N-K-j+1)\bigr) \binom{j+2K-3}{2K-3} \\ &= \sum_{j=2K-3}^{N+K-3} \bigl((N+K-2-j)^3 - (N+K-2-j)\bigr) \binom{j}{2K-3}. \end{aligned} \tag{177}$$

We expand the sum and write $\varphi_{N,K}$, with $G(N,K)=N+K-2$, as

$$\begin{aligned}\varphi_{N,K} &= \big(G(N,K)^3-G(N,K)\big)\sum_{j=2K-3}^{G(N,K)-1}\binom{j}{2K-3}\\ &\quad+\big(1-3G(N,K)\big)\sum_{j=2K-3}^{G(N,K)-1} j\binom{j}{2K-3}\\ &\quad+3G(N,K)\sum_{j=2K-3}^{G(N,K)-1} j^2\binom{j}{2K-3}\\ &\quad-\sum_{j=2K-3}^{G(N,K)-1} j^3\binom{j}{2K-3}.\end{aligned}\tag{178}$$

We evaluate the sums using the formulas

$$\sum_{j=k}^{n}\binom{j}{k}=\binom{n+1}{k+1},\tag{179}$$

$$\sum_{j=k}^{n} j\binom{j}{k}=(k+1)\binom{n+2}{k+2}-\binom{n+1}{k+1},\tag{180}$$

$$\begin{aligned}\sum_{j=k}^{n} j^2\binom{j}{k}&=(k+1)(k+2)\binom{n+3}{k+3}\\ &\quad-3(k+1)\binom{n+2}{k+2}+\binom{n+1}{k+1},\end{aligned}\tag{181}$$

$$\begin{aligned}\sum_{j=k}^{n} j^3\binom{j}{k}&=(k+1)(k+2)(k+3)\binom{n+4}{k+4}\\ &\quad-6(k+1)(k+2)\binom{n+3}{k+3}\\ &\quad+7(k+1)\binom{n+2}{k+2}-\binom{n+1}{k+1}.\end{aligned}\tag{182}$$

So the sum $\varphi_{N,K}$ is equal to

$$\begin{aligned}\varphi_{N,K} &= \big(G(N,K)^3-G(N,K)\big)\binom{G(N,K)}{2K-2}\\ &\quad+\big(1-3G(N,K)\big)(2K-2)\binom{G(N,K)+1}{2K-1}\end{aligned}$$

$$
\begin{aligned}
&-\big(1-3G(N,K)\big)\binom{G(N,K)}{2K-2}\\
&+3G(N,K)(2K-2)(2K-1)\binom{G(N,K)+2}{2K}\\
&-9(2K-2)G(N,K)\binom{G(N,K)+1}{2K-1}+3G(N,K)\binom{G(N,K)}{2K-2}\\
&-(2K-2)(2K-1)2K\binom{G(N,K)+3}{2K+1}\\
&+6(2K-2)(2K-1)\binom{G(N,K)+2}{2K}\\
&-7(2K-2)\binom{G(N,K)+1}{2K-1}+\binom{G(N,K)}{2K-2},
\end{aligned}\tag{183}
$$

which can be simplified into

$$
\begin{aligned}
\varphi_{N,K}={}&\binom{G(N,K)}{2K-2}\big(G(N,K)^3+5G(N,K)\big)\\
&-6\binom{G(N,K)+1}{2K-1}\big(1+2G(N,K)\big)(2K-2)\\
&+3\binom{G(N,K)+2}{2K}\big(G(N,K)+2\big)(2K-2)(2K-1)\\
&-\binom{G(N,K)+3}{2K+1}(2K-2)(2K-1)2K.
\end{aligned}\tag{184}
$$

We now use that $\binom{n+1}{k+1}=\frac{n+1}{k+1}\binom{n}{k}$ to write $\varphi_{N,K}$ as a function of $\binom{G(N,K)+1}{2K-1}$. Using equation (175), we obtain

$$
\begin{aligned}
s_K={}&\frac{a}{6}\frac{2K-1}{G(N,K)+1}\big(G(N,K)^3+5G(N,K)\big)\\
&-a\big(1+2G(N,K)\big)(2K-2)\\
&+\frac{a}{2}\big(G(N,K)+2\big)^2\frac{(2K-2)(2K-1)}{2K}\\
&-\frac{a}{6}\big(G(N,K)+2\big)\big(G(N,K)+3\big)\frac{(2K-1)(2K-2)}{2K+1}.
\end{aligned}\tag{185}
$$

The formation rates are given by

$$f_K = \frac{K(K-1)}{2} \tag{186}$$

and the probability of having K clusters is given by the relation

$$\Pi_K = \frac{f_{K+1}}{s_K}\Pi_{K+1}. \tag{187}$$

7. Conclusion.

In this paper, we investigated a certain class of discrete coagulation-fragmentation processes with a finite number of particles. We determined the steady-state probability distribution when the number of clusters is fixed. We studied the cluster distributions using the partitions of the total number of particles with a given number of clusters. We computed the distribution probability function in terms of multinomial coefficients.

This approach allows computing various statistical quantities and moments, including the mean number of clusters of a given size conditioned on the total number of clusters. However, computing other quantities, such as the size of the largest clusters cannot be derived from the present results and requires novel methods of calculation.

Finally, we defined two new times to characterize the cluster dynamics: the first one is the mean time that two particles spend together and the second is the mean time they spend separated. We computed here the fraction of these times, which is the probability that two particles are in the same cluster. We have applied these results to specific coagulation-fragmentation kernels. For the constant kernel, we obtained exact expressions of the number of clusters in terms of hypergeometric function. When the size of the cluster is limited, we obtained a model of nucleation in the limit $a \to 0$ and found multiple steady-state distributions, depending on the initial number of particles and the limit size.

Our study on coagulation-fragmentation of a finite number of particles was motivated by stochastic processes in chemical reaction theory and in molecular and cell biology of the cell nucleus organization. When the clusters and free particles evolve in a homogeneous region in dimension 2 or 3, the time two particles spend separated (recurrence time) is exactly the reciprocal of the forward rate of a chemical reaction. However, when the region is not homogenous so that clustering can occur preferentially in some subregions, this is not anymore the case, and the recurrence time can be shorter than the meeting time, as discussed in the context of telomere clustering in yeast [17]. In that case, clustering favors encounter. The time that two particles stay in the same cluster is an indicator of the possible exchange of genetic information between clustered telomeres, a process that remains to be studied both experimentally and theoretically.

Acknowledgments. David Holcman thanks the Institute of Mathematics, Oxford University, for its hospitality.

REFERENCES

[1] ABRAMOWITZ, M. and STEGUN, I. A., eds. (1992). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover Publications, New York. MR1225604

[2] ALDOUS, D. J. (1999). Deterministic and stochastic models for coalescence (aggregation and coagulation): A review of the mean-field theory for probabilists. *Bernoulli* **5** 3–48. MR1673235

[3] AMANN, H. (2000). Coagulation-fragmentation processes. *Arch. Ration. Mech. Anal.* **151** 339–366.

[4] ANDREWS, G. E. (1976). *The Theory of Partitions. Encyclopedia of Mathematics and Its Applications* **2**. Addison-Wesley, Reading, MA. MR0557013

[5] BALL, J. M. and CARR, J. (1990). The discrete coagulation-fragmentation equations: Existence, uniqueness, and density conservation. *J. Stat. Phys.* **61** 203–234. MR1084278

[6] BECKER, R. and DÖRING, W. (1935). Kinetische Behandlung der Keimbildung in übersättigten Dämpfen. *Ann. Phys.* **24** 719–752.

[7] CARR, J. and DA COSTA, F. P. (1994). Asymptotic behavior of solutions to the coagulation-fragmentation equations. II. Weak fragmentation. *J. Stat. Phys.* **77** 89–123. MR1300530

[8] CHANDRASEKAR, S. (1943). Stochastic problems in physics and astrophysics. *Rev. Modern Phys.* **15** 1–89.

[9] COLLET, J. F. (2004). Some modelling issues in the theory of fragmentation-coagulation systems. *Commun. Math. Sci.* **1** 35–54. MR2119872

[10] DA COSTA, F. P. (1998). Asymptotic behaviour of low density solutions to the generalized Becker–Döring equations. *NoDEA Nonlinear Differential Equations Appl.* **5** 23–37.

[11] DOERING, C. R. and BEN-AVRAHAM, D. (1988). Interparticle distribution functions and rate equations for diffusion-limited reactions. *Phys. Rev. A* **38** 3035.

[12] DURRETT, R., GRANOVSKY, B. L. and GUERON, S. (1999). The equilibrium behavior of reversible coagulation-fragmentation processes. *J. Theoret. Probab.* **12** 447–474. MR1684753

[13] GUERON, S. (1998). The steady-state distributions of coagulation-fragmentation processes. *J. Math. Biol.* **37** 1–27. MR1636636

[14] HOZE, N. and HOLCMAN, D. (2012). Coagulation–fragmentation for a finite number of particles and application to telomere clustering in the yeast nucleus. *Phys. Lett. A* **376** 845–849.

[15] HOZE, N. and HOLCMAN, D. (2014). Modeling capsid kinetics assembly from the steady state distribution of multi-sizes aggregates. *Phys. Lett. A* **378** 531–534.

[16] HOZE, N. and HOLCMAN, D. (2015). Kinetics of aggregation with a finite number of particles and application to viral capsid assembly. *J. Math. Biol.* **70** 1685–1705. MR3343937

[17] HOZE, N., RUAULT, M., AMORUSO, C., TADDEI, A. and HOLCMAN, D. (2013). Spatial telomere organization and clustering in yeast Saccharomyces cerevisiae nucleus is generated by a random dynamics of aggregation–dissociation. *MBoC* **24** 1791–1800.

[18] JACQUOT, S. (2009). A historical law of large numbers for the Marcus–Lushnikov process. *Electron. J. Probab.* **15** 605–635. MR2639735

[19] JONES, W. B. and THRON, W. J. (1980). *Continued Fractions: Theory and Applications. Encyclopedia of Mathematics and Its Applications* **11**. Addison-Wesley, Reading, MA. MR0595864

[20] KELLY, F. P. (1979). *Reversibility and Stochastic Networks*. Wiley, Chichester. MR0554920

[21] KRAPIVSKY, P. L., REDNER, S. and BEN-NAIM, E. (2010). *A Kinetic View of Statistical Physics*. Cambridge Univ. Press, Cambridge. MR2757286

[22] LIGGETT, T. M. (1985). *Interacting Particle Systems. Grundlehren der Mathematischen Wissenschaften* [*Fundamental Principles of Mathematical Sciences*] **276**. Springer, New York. MR0776231

[23] LUSHNIKOV, A. A. (1978). Coagulation in finite systems. *J. Colloid Interface Sci.* **65** 276–285.

[24] MARCUS, A. H. (1968). Stochastic coalescence. *Technometrics* **10** 133–143. MR0223151

[25] MEYER, C. D. JR. (1975). The role of the group generalized inverse in the theory of finite Markov chains. *SIAM Rev.* **17** 443–464. MR0383538

[26] NORRIS, J. R. (1999). Smoluchowski's coagulation equation: Uniqueness, nonuniqueness and a hydrodynamic limit for the stochastic coalescent. *Ann. Appl. Probab.* **9** 78–109. MR1682596

[27] ROTSTEIN, H. G. (2015). Cluster-size dynamics: A phenomenological model for the interaction between coagulation and fragmentation processes. *J. Chem. Phys.* **142** 224101.

[28] RUAULT, M., DE MEYER, A., LOĒODICE, I. and TADDEI, A. (2011). Clustering heterochromatin: Sir3 promotes telomere clustering independently of silencing in yeast. *J. Cell Biol.* **192** 417–431.

[29] SIMPSON, E. H. (1949). Measurement of diversity. *Nature* **163** 688.

[30] SZEGÖ, G. (1975). *Orthogonal Polynomials*, 4th edn. Amer. Math. Soc. Colloq. Publ., Providence, RI.

[31] THOMPSON, C. J. (1988). *Classical Equilibrium Statistical Mechanics*. Oxford Univ. Press, Oxford.

[32] THOMSON, B. R. (1989). Exact solution for a steady-state aggregation model in one dimension. *J. Phys. A* **22** 879–886. MR0989039

[33] VON SMOLUCHOWSKI, M. (1916). Drei Vorträge über Diffusion Brownsche Molekularbewegung und Koagulation von Kolloidteichen. *Z. Phys.* **17** 557–571.

[34] WATTIS, J. A. D. (2006). An introduction to mathematical models of coagulation-fragmentation processes: A discrete deterministic mean-field approach. *Phys. D* **222** 1–20. MR2265763

[35] YVINEC, R., D'ORSOGNA, M. R. and CHOU, T. (2012). First passage times in homogeneous nucleation and self-assembly. *J. Chem. Phys.* **137** 244107.

[36] ZLOTNICK, A. (2005). Theoretical aspects of virus capsid assembly. *JMR, J. Mol. Recognit.* **18** 479–490.

INSTITUTE OF INTEGRATIVE BIOLOGY
ETH ZURICH
8092 ZURICH
SWITZERLAND
E-MAIL: nathanael.hoze@env.ethz.ch

NEWTON INSTITUTE
DEPARTMENT OF APPLIED MATHEMATICS
AND THEORETICAL PHYSICS (DAMPT)
UNIVERSITY OF CAMBRIDGE
CB30DS
UNITED KINGDOM
AND
ECOLE NORMALE SUPÉRIEURE
75005 PARIS
FRANCE
E-MAIL: david.holcman@ens.fr