AIP | The Journal of Chemical Physics

First passage times in homogeneous nucleation: Dependence on the total number of particles

Romain Yvinec, Samuel Bernard, Erwan Hingant, and Laurent Pujo-Menjouet

Citation: The Journal of Chemical Physics **144**, 034106 (2016); doi: 10.1063/1.4940033
View online: http://dx.doi.org/10.1063/1.4940033
View Table of Contents: http://scitation.aip.org/content/aip/journal/jcp/144/3?ver=pdfcov
Published by the AIP Publishing

First passage times in homogeneous nucleation: Dependence on the total number of particles

Romain Yvinec,[1] Samuel Bernard,[2,3] Erwan Hingant,[4] and Laurent Pujo-Menjouet[2,3]
[1]*PRC INRA UMR85, CNRS UMR7247, Université François Rabelais de Tours, IFCE, F-37380 Nouzilly, France*
[2]*Université de Lyon, CNRS, Université Lyon 1, Institut Camille Jordan UMR5208, 69622 Villeurbanne, France*
[3]*INRIA Team Dracula, Inria Center Grenoble Rhône-Alpes, Grenoble, France*
[4]*Departamento de Matemática, Universidad Federal de Campina Grande, Campina Grande, PB, Brazil*

(Received 18 October 2015; accepted 5 January 2016; published online 20 January 2016)

Motivated by nucleation and molecular aggregation in physical, chemical, and biological settings, we present an extension to a thorough analysis of the stochastic self-assembly of a fixed number of identical particles in a finite volume. We study the statistics of times required for maximal clusters to be completed, starting from a pure-monomeric particle configuration. For finite volumes, we extend previous analytical approaches to the case of *arbitrary size-dependent* aggregation and fragmentation kinetic rates. For larger volumes, we develop a scaling framework to study the first assembly time behavior as a function of the total quantity of particles. We find that the mean time to first completion of a maximum-sized cluster may have a surprisingly weak dependence on the total number of particles. We highlight how higher statistics (variance, distribution) of the first passage time may nevertheless help to infer key parameters, such as the size of the maximum cluster. Finally, we present a framework to quantify formation of macroscopic sized clusters, which are (asymptotically) very unlikely and occur as a large deviation phenomenon from the mean-field limit. We argue that this framework is suitable to describe phase transition phenomena, as *inherent infrequent stochastic processes*, in contrast to classical nucleation theory. © *2016 AIP Publishing LLC*. [http://dx.doi.org/10.1063/1.4940033]

I. INTRODUCTION

Self-assembly of macromolecules and particles into clusters is a fundamental process in many physical, chemical, and biological systems. Although particle nucleation and assembly have been studied for many decades,[1,2] interest in this field has recently increased due to engineering, biotechnological, and imaging advances at nanoscale levels.[3–5] Applications range from material physics to cell physiology and virology (for a detailed list of examples, see Ref. 6 and references therein). Many of these applications involve a fixed "maximum" cluster size — of tens to hundreds of units — at which the process is completed or beyond which dynamics changes.[7,8] One example includes the rare and sporadic self-assembly of misfolded proteins into fibril aggregates at the origin of several neurodegenerative diseases (Alzheimer, Parkinson, Prion, etc.).[9,10] Developing a stochastic self-assembly model focusing on formation of a fixed "maximum" cluster size is thus important for our understanding of a large class of biological processes, and the quantification of experimental data[11–15] variability in order to find strategies to optimize processes for industrial applications or to prevent onsets in the case of neurodegenerative diseases.

Theoretical models for self-assembly have typically described mean-field concentrations of clusters of all possible sizes using the well-studied mass-action, Becker-Döring (BD) equations.[16–19] While master equations for the fully stochastic nucleation and growth problem have been derived, and initial analyses and simulations have been performed[20–24] (we compare our results with previous ones in Section V), there has been relatively less scientific contribution to the stochastic self-assembly problem. On the other hand, it has been recently shown that in finite systems, where the maximum cluster size is capped, results from mass-action equations are inaccurate and a discrete stochastic treatment is then necessary.[6,25] We consider here the BD model defined by the following biochemical reactions:

$$C_1 + C_k \underset{q_{k+1}}{\overset{p_k}{\rightleftharpoons}} C_{k+1}, \quad k \geq 1, \tag{1}$$

where C_k denotes the number of clusters of size k. Thus, the size of each cluster can increase or decrease by one, with an attachment or detachment of a single free particle (called monomer here). The mean-field (BD) model is described in Sec. II. In the stochastic Becker-Döring (SBD) version, the state-space of the system is discrete and finite (see Fig. 1), given by all possible combinations of cluster sizes that have a given fixed total number of particles (defined by M, given by the initial condition),

$$\mathcal{E} := \left\{(C_k)_{k\geq 1} \subset \mathbb{N} : \sum_{k\geq 1} kC_k = M\right\}. \tag{2}$$

The key modeling assumption of the SBD model is the Markovian hypothesis.[36] Indeed, clusters $(C_k(t))_{k\geq 1}$ evolve in continuous time by discrete jumps according to a Markovian description of reactions (1), with $C_k(t) \in \mathcal{E}$ for all $t \geq 0$. In previous examination of the first assembly time (FAT) in this model,[6] authors surprisingly found that a striking finite-size effect could arise in the limit of slow self-assembly. In particular, a *faster* detachment rate could lead to a

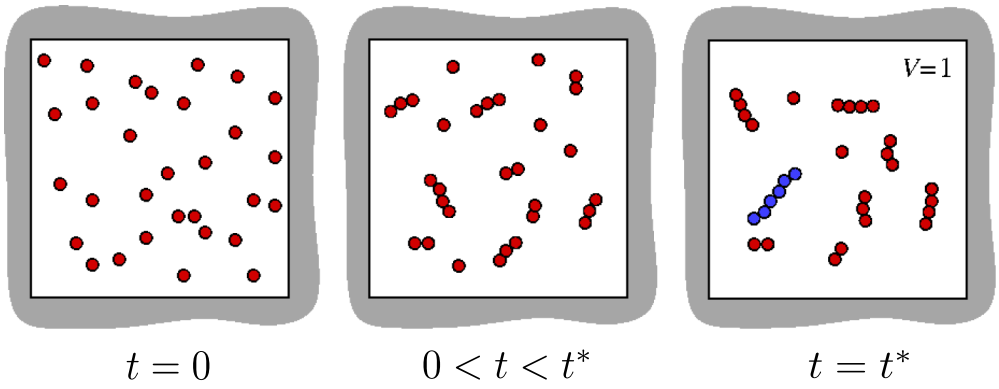

FIG. 1. Homogeneous self-assembly and growth in a closed unit volume initiated with $M = 30$ free monomers. At a specific intermediate time $0 < t < t^*$ in this depicted realization, there are six free monomers, four dimers, four trimers, and one cluster of size four. For each realization of this process, there will be a specific time t^* at which a maximum cluster of size $N = 6$ in this example is first formed (blue cluster). This time t^* is a realization of the first assembly time (FAT, see definition in (7)).

shorter assembly time. This unexpected effect was proven to occur when the finite-size system occupies some specific configurations named "traps," where no single particle is free and the maximal-size cluster completion can only be achieved through the detachment of single particles from a cluster. Discrepancies between mean-field mass-action approach and stochastic model were more pronounced in the strong binding limit.

Objectives of this paper. In this paper, we have the following:

1. We present a generalization of earlier results[6] on the statistics of the first assembly times towards completion of a full cluster, for arbitrary aggregation and fragmentation rates. Indeed, constant-size reaction rates were the main limitation of previous studies.[6] And it is known that, in general, both of physical and biological modeling require size-dependent attachment and detachment rates.[13,26]
2. Moreover, we focus here on how assembly times depend on the total initial number of monomers M, an aspect which was not treated in earlier studies.[6] We will show how statistics of the first assembly time as a function of the total number of monomers M may shed light on the biophysical properties of the newly formed critical aggregates.
3. We highlight discrepancies between the mean-field mass-action approach and our stochastic model. Even in the limit $M \to \infty$, we show that our SBD model can display a large variability in the first assembly times, with a non-vanishing normalized variance. Thus, this work gives a suitable theoretical framework to explain experimental variability in the *in vitro* self-assembly of misfolded proteins which are typically performed[14] with a large number of proteins, in the order of 10^{10}–10^{12} molecules.

Our work is organized as follows. In Sec. II, we review Becker-Döring mass-action equations for self-assembly and introduce our full stochastic problem. We derive stochastic equations for time-dependent cluster numbers and introduce assembly times as first passage time problems. In Section III, we explore two simplified models for which the first assembly time can be solved analytically and derive asymptotic expressions for the first assembly time in both large number of monomer limit and large cluster size limit. Results from kinetic Monte Carlo (or stochastic simulations algorithm) simulations are presented in Section IV and compared with our analytical estimates. Finally, we compare our results to the literature and discuss possible implications of our results and propose further extensions in Sec. V.

II. STOCHASTIC BECKER-DÖRING MODEL, FIRST ASSEMBLY TIMES DEFINITIONS

The classic deterministic mass-action description for spontaneous, homogeneous self-assembly is the BD model,[1] where concentrations $c_k(t)$ of clusters of size k obey an infinite (or truncated up to $k = N$) system of ordinary differential equations, given, for all $t \geq 0$, by

$$\begin{cases} \dfrac{d}{dt}c_1(t) = -2j_1(t) - \displaystyle\sum_{k\geq 2} j_k(t), \\ \dfrac{d}{dt}c_k(t) = j_{k-1}(t) - j_k(t), \quad k \geq 2, \end{cases} \tag{3}$$

with

$$j_k(t) = p_k c_1(t) c_k(t) - q_{k+1} c_{k+1}(t), \quad k \geq 1, \tag{4}$$

and initial condition $c_1(0) = M$ and $c_k(0) = 0$ for all $k \geq 2$. The rates p_k and q_k are, respectively, monomer attachment and detachment rates to and from a cluster of size k. These rates are limited to sub-linear function of k, with bounded increments, in order to fulfill the standard well-posedness criteria.[27,28] It has been previously shown that such equations provide a poor approximation of the expected number of clusters when the total mass M and the maximum cluster size N are comparable in magnitude.[25] Furthermore, such representations do not capture the randomness of the binding/unbinding events and cannot describe heterogeneity of cluster size distributions and of time-dependent properties such as first assembly times. *A stochastic treatment is thus necessary and is the subject of the remainder of this paper.*

Using a Markovian approach, we have previously derived[6] a forward master equation to describe the probability that the system is in any given admissible configuration at times $t \geq 0$. An equivalent formulation of this model is given by stochastic equations, driven by Poisson processes. This formulation is the

natural one for performing numerical simulations of sample paths and is more efficient for computing first assembly times than the master equation formulation. Moreover, this formulation leads to a natural comparison with deterministic systems when the total mass M is large. The equations are built in the following way. For each reaction (attachment, detachment of a monomer) involving a cluster of size k, we associate a counting process, $R_k^+(t)$ and $R_k^+(t)$, that counts the number of occurrences of that reaction between times 0 and t. The Markovian hypothesis implies that each counting process can be formulated as a random time-change of a standard (with a unit rate) Poisson process. When a reaction occurs, the number of each species is updated according to the stoichiometry of the reaction. For instance, if the aggregation reaction of two free particles ($k = 1$) occurs at time t, the number $C_1(t)$ is decreased by 2 and the number $C_2(t)$ is increased by 1. Given the quantity of free particles $C_1(t) = C_1$ at time t, the next increment of R_1^+ will occur after a random time given by an exponential law of parameter $\frac{p_1}{V}C_1(C_1 - 1)$, where V denotes the volume of the system, and p_1 the kinetic reaction rate constant. Thanks to the homogeneity property of the exponential law, one can represent the counting process R_1^+ as

$$R_1^+(t) = Y_1^+\Big(\int_0^t \frac{p_1}{V}C_1(s)(C_1(s) - 1)ds\Big),$$

where Y_1^+ is a unit rate Poisson process. All reactions in the system proceed similarly and independently of each other. Denoting by Y_k^+ (respectively, Y_k^-), the standard Poisson processes associated to the forward, aggregation (respectively, backward, fragmentation) reaction of clusters of size k, SBD equations for time evolution of the number of cluster of size k, $C_k(t)$, starting from a pure monomeric initial condition, are given for $t \geq 0$ by

$$\begin{cases} C_1(t) = M - 2J_1(t) - \sum_{k\geq 2} J_k(t), \\ C_k(t) = J_{k-1}(t) - J_k(t), \quad k \geq 2, \end{cases} \tag{5}$$

with

$$J_k(t) = Y_k^+\Big(\int_0^t \frac{p_k}{V}C_1(s)(C_k(s) - \delta_{1k})ds\Big) - Y_{k+1}^-\Big(\int_0^t q_{k+1}C_{k+1}(s)ds\Big), \quad k \geq 1, \tag{6}$$

where $\delta_{1k} = 1$ if $k = 1$ and $\delta_{1k} = 0$ if $k > 1$. Analogy between Eqs. (5) and (3) is clear. The number of clusters of size $k \geq 2$ evolves according to the differences between two (stochastic) cumulative counts J_{k-1} and J_k.

What we call FAT for stochastic discrete Becker-Döring equations is defined as a first passage time problem[29]

$$T_{1,0}^{N,M} := \inf\{t \geq 0 : C_N(t) = 1 \mid C_k(0) = M\delta_{1k}\}. \tag{7}$$

Hence, FAT is the first time to obtain a cluster of size N, starting with an M single particle initial state (see Fig. 1, for example). To link it with the macroscopic nucleation time definition, we also consider the generalized first assembly time (GFAT) problem

$$T_{\rho,h}^{N,M} := \inf\{t \geq 0 : C_N(t) \geq \rho M^h \mid C_k(0) = M\delta_{1k}\}, \tag{8}$$

for a given positive constant ρ and $0 \leq h \leq 1$. Superscripts M, N in Eqs. (7) and (8) and subscripts ρ, h in Eq. (8) are the key parameters of the first assembly times and are thus written explicitly. The first assembly times also depend on the reactions rates p_k, q_k. This will be mentioned further below. Subscripts 1 and 0 of FAT in Eq. (7) are consistent with $\rho = 1$ and $h = 0$ of GFAT in Eq. (8). For instance, the specific time t^* in Fig. 1 represents a particular realization of the random variable $T_{1,0}^{6,30}$.

Here, we want to analyze how the statistics (mean, variance, distribution) of $T_{\rho,h}^{N,M}$ depend on the total number of monomers M. We are interested in characterizing the asymptotic behavior of GFAT $T_{\rho,h}^{N,M}$, for $M \gg 1$: convergence, and speed of convergence, to 0, a positive value or infinity. A question may arise then here. Is the asymptotic limit random or deterministic? As $M \to \infty$, the maximal cluster size N is allowed to increase with M or stay constant, with different expected results. Two distinct cases are then considered: finite (small) maximal cluster size N and large maximal cluster size $N \gg 1$. The latter obviously requires that $M \geq N$. We see below that we need to specify more precisely the relationship between N and M. Influence of the other parameters ρ, h, p_k, q_k will also be highlighted.

One way of computing the distribution of first assembly times is to consider the Backward Kolmogorov equation (BKE) describing evolution of configuration probabilities as a function of local changes in initial configuration, as done previously.[6] It has the advantage to yield exact results for the full distribution of FAT, but it is strictly limited by the number of reactions, which grows exponentially with M. In this paper, we rely on exact calculations of simplified reduced models, limit theorems from Eq. (5) for large M and N, and extensive numerical simulations of these equations. We use also asymptotic approaches, when $M \to \infty$, for fixed N, and when both $M, N \to \infty$. The total number of monomers M can be expressed as the product of an initial concentration c^0 and volume V,

$$M = c^0 V. \tag{9}$$

We distinguish two situations: large monomer number limit $M \to \infty$ can correspond either to a large initial concentration $c_0 \to \infty$ in a fixed volume or to a large volume $V \to \infty$ with a fixed concentration c^0. As aggregation reactions naturally depend on the volume of system,[30] the two situations (large concentration or large volume) will yield distinct results. Experimentally, they also correspond to different protocols.

III. RESULTS AND ANALYSIS

Although state-space (2) of our SBD model (5) and (6) is finite, the first passage problem defined by Eq. (8) is, in general, a challenging problem. Two of the reasons for this difficulty rely on intrinsic non-linearity of the aggregation process, and the (very large) size of state-space (2). There are two distinct simplifications allowing our problem to be analytically tractable. Let us develop them in Subsections III A and III B. In Subsection III A, we consider a linear version of Eq. (5), and in Subsection III B we present a state-space

reduction to a one-dimensional space of cardinal N. Then, we come back to full SBD models (5) and (6) and present asymptotic results for large initial quantity of particles, $M \to \infty$, with either large initial concentration $c_0 \to \infty$ or large volume $V \to \infty$. Those results are developed in two subsections, depending on whether nucleus size is finite (Subsection III C) or infinite (Subsection III D). Our strategy is based on a re-scaling procedure of stochastic equation (5). Numerical illustrations and detailed discussion of our results are postponed to Section IV.

A. Constant monomer formulation

SBD model defined by Eq. (5) has the constant mass property

$$\sum_{k\geq 1} kC_k(t) \equiv \sum_{k\geq 1} kC_k(0) = M, \quad t \geq 0,$$

which implies a nonlinear relationship between the size of each cluster. This contrasts with the original formulation of the BD model, sometimes used in deterministic contexts,[17,27] where the total mass of the system is not preserved, but the quantity of free particles is kept constant. We refer to this formulation as constant monomer stochastic Becker-Döring (CMSBD) model. We can represent it by the following reactions:

$$\begin{cases} \emptyset & \underset{q_2}{\overset{\frac{p_1}{V}M(M-1)}{\rightleftharpoons}} & C_2, \\ C_k & \underset{q_{k+1}}{\overset{\frac{p_k}{V}M}{\rightleftharpoons}} & C_{k+1}, \quad i \geq 2. \end{cases} \tag{10}$$

In this formulation (10), $C_1(t) \equiv M$ is now a constant parameter. Note that we expect such model to be close to the original SBD (for small times, up to the FAT) in the limit of large number of particles M. The main advantages of constant monomer formulation are its linearity and the fact that all clusters are independent from each others. Hence, it is analytically solvable. Indeed, it is known that for linear population models,[31] the numbers of individuals in each subclass of a population (starting with no individuals at time 0), namely, here $C_2(t), \ldots, C_N(t) \ldots$, are independent Poisson random variables. Moreover, for CMSBD model (10), mean cluster sizes $c_2(t), \ldots, c_N(t) \ldots$ are solutions of a system of linear equations, given for all $t \geq 0$, by

$$\frac{d}{dt}c_k(t) = j_{k-1}(t) - j_k(t), \quad \forall k \geq 2, \tag{11}$$

with

$$\begin{cases} j_1(t) = \frac{p_1}{V}M(M-1) - q_2c_2(t), \\ j_k(t) = \frac{p_k}{V}Mc_k(t) - q_{k+1}c_{k+1}(t), \quad \forall k \geq 2, \end{cases} \tag{12}$$

and initial condition $c_k(0) = 0$ for all $k \geq 2$. Note that the last set of Eqs. (11) and (12) is very close to deterministic Becker-Döring models (3) and (4) taking $c_1 \equiv M$. To calculate the FAT $T_{1,0}^{N,M}$, we use the survival function

$$\begin{aligned} S_{1,0}^{N,M}(t) &:= \mathbb{P}\{T_{1,0}^{N,M} > t\} \\ &= \mathbb{P}\{C_N(s) = 0, s \leq t \mid C_k(0) = M\delta_{1k}\}. \end{aligned}$$

Then, using an absorbing boundary condition at $k = N$ ($q_N = p_N = 0$) together with initial condition entails that $C_N(t) = 0$ for some $t \geq 0$ if and only if $C_N(s) = 0$ for all $s \leq t$, so that

$$S_{1,0}^{N,M}(t) = \mathbb{P}\{C_N(t) = 0 \mid C_k(0) = M\delta_{1k}\}.$$

Finally, since $C_N(t)$ is Poisson distributed (linear system) with mean $c_N(t)$, we have

$$S_{1,0}^{N,M}(t) = e^{-c_N(t)}. \tag{13}$$

Equations (11) and (12) with absorbing boundary at $k = N$ can be rewritten as a linear system

$$\begin{cases} \dot{c} &= Ac + B, \\ \dot{c}_N(t) &= p_{N-1}Mc_{N-1}(t), \end{cases} \tag{14}$$

where c and B are vectors, with $c = (c_2, c_3, \ldots, c_{n-1})^T$, $B = (\frac{p_1}{V}M(M-1), 0, \ldots, 0)^T$ and A is a tridiagonal matrix with elements

$$\begin{cases} a_{k,k} &= -q_{k+1} - \frac{p_{k+1}}{V}M, \\ a_{k+1,k} &= \frac{p_{k+1}}{V}M, \\ a_{k,k+1} &= q_{k+1}. \end{cases}$$

Study of linear system (14) has been performed both for the infinite dimensional case[32] and for the truncated case.[33] See Section 1 of the supplementary material[49] for a general formula of solutions of (14). Asymptotic analysis for small times of system (14) gives that, for $t \ll 1$,

$$c_N(t) \approx_{t\ll 1} \frac{M^N}{V^{N-1}} \prod_{k=1}^{N-1} p_k \frac{t^{N-1}}{(N-1)!},$$

and Eq. (13) is thus the survival function of a Weibull distribution, of shape parameter $k = N-1$ and scale parameter $\lambda = V((N-1)!/(M^N \prod_{k=1}^{N-1} p_k))^{1/(N-1)}$. Hence, we get

$$\langle T_{1,0}^{N,M} \rangle \approx_{M\to\infty} V \frac{\Gamma(1 + 1/(N-1))}{\left(\prod_{k=1}^{N-1} p_k\right)^{1/(N-1)}} \frac{((N-1)!)^{1/(N-1)}}{M^{1+1/(N-1)}}. \tag{15}$$

From Eq. (15), we can distinguish the large concentration limit from the large volume limit. Recall that we defined $M = c^0V$. Thus, in the large concentration limit (taking $V = 1$),

$$\langle T_{1,0}^{N,M} \rangle \approx_{c^0\to\infty} \frac{\Gamma(1 + 1/(N-1))}{\left(\prod_{k=1}^{N-1} p_k\right)^{1/(N-1)}} \frac{((N-1)!)^{1/(N-1)}}{(c^0)^{1+1/(N-1)}}, \tag{16}$$

while for the large volume limit (taking $c^0 = 1$),

$$\langle T_{1,0}^{N,M} \rangle \approx_{V\to\infty} \frac{\Gamma(1 + 1/(N-1))}{\left(\prod_{k=1}^{N-1} p_k\right)^{1/(N-1)}} \frac{((N-1)!)^{1/(N-1)}}{V^{1/(N-1)}}. \tag{17}$$

Note that in both cases, Eqs. (16) and (17), the mean FAT converges to 0, at speeds $1 + 1/(N-1)$ and $1/(N-1)$, respectively.

Variance formula for the Weibull distribution yields the asymptotic coefficient of variation (CV, standard deviation over the mean)

$$cv_{T_{1,0}^{N,M}} \approx_{M\to\infty} \sqrt{2(N-1)\frac{\Gamma(2/(N-1)}{\Gamma(1/(N-1))^2} - 1}. \quad (18)$$

The coefficient of variation does not vanish in large population, it is independent of the particular aggregation rate shape and depends only on the size of the maximal cluster N. It is also independent of the particular limit, being a large concentration or a large volume limit.

For the GFAT, a time scale asymptotic on equations similar to Eq. (14) for mean gives the following expression:

$$\langle T_{\rho,h}^{N,M}\rangle \approx_{M\to\infty} V\frac{C(p,N)}{M}\frac{1}{M^{(1-h)/(N-1)}}, \quad (19)$$

where $C(p,N)$ is a constant that depends only on N and aggregation rates $p_k, k \le N$ (that can be made explicit if the full solution of Eq. (14) is known). Those asymptotic expressions are illustrated in Figure S1 of the supplementary material[49] where a perfect match is observed with numerical simulations.

B. Single cluster model

Another simplified model that can be analytically solved for our FAT problem is given by the assumption that only a single cluster can be formed at a time.[6,34] We expect such a model to be close to the original one when fragmentation dominates, so that formation of many (large) clusters is unlikely. In such model, called single-cluster stochastic Becker-Döring (SCSBD) model, we may represent only the size of the single cluster, so that our state space is now one dimensional, being simply

$$\mathcal{E}_1 := [1,\dots,N],$$

and possible reactions are given by (k denotes the size of the single cluster)

$$\begin{cases} k=1 \underset{q_2}{\overset{\frac{p_1}{V}M(M-1)}{\rightleftharpoons}} k=2, \\ k \underset{q_{k+1}}{\overset{\frac{p_k}{V}(M-k)}{\rightleftharpoons}} k+1, \quad k\ge 2. \end{cases} \quad (20)$$

In such a scenario, exact solution and classical first passage theory[30] gives (it is a one-dimensional discrete random walk)

$$\langle T_{1,0}^{N,M}\rangle = \sum_{i=1}^{N-1}\sum_{j=1}^{i}\frac{\prod_{k=j+1}^{i} q_k}{\prod_{k=j}^{i} p_k}\frac{V^{i-j+1}}{M^{\delta_{1j}}\prod_{k=j}^{i}(M-k)}. \quad (21)$$

In addition, general formulas for variance and cumulative distribution function are available.[35] Those theoretical expressions are illustrated in Figure S2 of the supplementary material[49] where a perfect match is observed with numerical simulations.

Although exact expressions such as Eq. (21) are valid, asymptotic expressions are still of interest and will illustrate the rescaling strategy we use for the full SBD model. Thus, we consider various different limits, including large fragmentation rate, large initial number of monomer M, and large maximal cluster size N.

First, in the unfavorable aggregation limit, i.e., $q_k = \frac{\overline{q}_k}{\varepsilon}$ and $\varepsilon \to 0$, the leading order of mean assembly time is

$$\langle T_{1,0}^{N,M}\rangle \approx_{\varepsilon\to 0} \frac{1}{\varepsilon^{N-2}}\frac{V^{N-1}\prod_{k=2}^{N-1}\overline{q}_k}{\prod_{k=1}^{N-1}p_k\prod_{k=0}^{N-1}(M-k)}.$$

Also, one can show that in the asymptotic $\varepsilon \to 0$, for large fragmentation rate, the FAT $T_{1,0}^{N,M}$ is an exponential distribution.[6]

Then, we consider the limit of large total number of monomers M. For the large volume scenario, taking $c^0 = 1$, we have

$$\langle T_{1,0}^{N,M}\rangle \approx_{V\to\infty} \sum_{i=2}^{N-1}\sum_{j=2}^{i}\frac{\prod_{k=j+1}^{i} q_k}{\prod_{k=j}^{i} p_k}, \quad (22)$$

which corresponds to the mean first passage time of a simple random walk between 2 and N, with rates p_k, q_k. For large concentration scenario, the leading order of the mean assembly time is

$$\langle T_{1,0}^{N,M}\rangle \approx_{c^0\to\infty} \frac{1}{c^0}\left(\sum_{k=2}^{N-1}\frac{1}{p_k(1-k/M)}\right), \quad (23)$$

and the normalized FAT, $MT_{1,0}^{N,M}$, is asymptotically a sum of $N-2$ exponential variables of parameter p_i, $2 \le i \le N-1$. Note that as in the linear CMSBD model, for very large concentration c^0, mean FAT is only dependent on the forward aggregation rates and is roughly inversely proportional to M.

Finally, to illustrate the case of very large nucleus size, let us consider N as a fraction $\alpha < 1$ of the total number of particles, i.e., $N = \lfloor \alpha M \rfloor$ in Eq. (21). Writing

$$p(x) = \sum_{k\ge 2} p_k \mathbf{1}_{[k/M,(k+1)/M)}(x),$$

the sum in Eq. (23) may be approximated by

$$\langle T_{1,0}^{\alpha M,M}\rangle \approx_{c^0\to\infty} \int_0^\alpha \frac{1}{p(x)(1-x)}dx, \quad (24)$$

which may be finite or infinite according to the behavior of p and α. Thus, in the large concentration limit (favorable aggregation limit), a simple criterion for formation of a very large cluster in a finite time is that integral (24) has to be finite. For large N and in the large volume limit, let us introduce a continuous rescaled size variable $x = k/N$, and define the rescaled kinetic rates

$$p(x) = \sum_{k\ge 2} p_k \mathbf{1}_{[k/N,(k+1)/N)}(x),$$
$$q(x) = \sum_{k\ge 2} q_k \mathbf{1}_{[k/N,(k+1)/N)}(x).$$

We have, for $N = \sqrt{M}$, taking $c^0 = 1$ and $V \to \infty$ (see Section 2.2 of the supplementary material[49] for detailed calculations),

$$\langle T_{1,0}^{\sqrt{M},M}\rangle \approx_{V\to\infty} V\int_0^1\int_0^y \frac{e^{(y^2-z^2)/2}}{q(y)}\cdot \exp\left[\sqrt{V}\int_z^y \ln\left(\frac{q(x)}{p(x)}\right)dx\right]dydz. \quad (25)$$

In particular, when $q(x) > p(x)$ on an interval of positive measure on $[0,1]$, last expression (23) implies that the mean FAT required to reach *macroscopic* size $x = 1$ ($k = N = \sqrt{M}$)

is exponentially large as $M \to \infty$. This case corresponds to the one studied numerically in earlier study.[20] As an example, suppose that kinetic rates are size-independent with $q > p$. Then, Eq. (23) becomes

$$\langle T_{1,0}^{\sqrt{M},M} \rangle \approx_{V\to\infty} \frac{V}{q} \int_0^1 \int_0^y e^{(y^2-z^2)/2} \left(\frac{q}{p}\right)^{\sqrt{V}(y-z)} dydz.$$

Those theoretical expressions are illustrated in Figures S3 and S4 of the supplementary material[49]. We could have derived the asymptotic expressions straightforwardly because of explicit formula (21). This is unfortunately not the case for the full SBD model. Note however that, for the single-cluster model, a different approach from the exact formula to derive asymptotic expressions for FAT is to link one-dimensional discrete random walk (20) to a continuous size one-dimensional ordinary or stochastic differential equation, and to use limit theorems and large deviation theory[34] (see Section 2.3 of the supplementary material[49] for a brief introduction). This scaling approach, and the link with a continuous size model when $N \to \infty$, will be used on the full SBD model in Sections III C and III D.

C. Full SBD, large *M*, finite *N*

In this section, we investigate the behavior of the SBD model and its FAT when the total number of particles M goes to infinity, while size N stays finite. We recall that writing $M = c^0 V$ leads to distinguishing two scenarios that yield distinct results. In the first one, concentration c^0 is large. As the overall aggregation propensities increase with concentration c^0, it is expected that the FAT will decrease to 0 as $M \to \infty$. The objective is to find a valid asymptotic expression, and its dependence with respect to other parameters, like the maximal cluster size N for instance. In the second case, volume V is large. This scaling is motivated by classical system size expansion of chemical reaction networks.[30] As volume increases, the total number of particles also increases, so that overall reaction propensities of aggregation reactions stay constant. In such a case, one expect to regain the deterministic first passage time of the classical deterministic BD model.

Before introducing our general rescaling strategy, let us consider an illustrative example. Consider the deterministic irreversible aggregation model with first passage time defined as follows:

$$\begin{cases} \dfrac{dc_k}{dt} = c_1(p_{k-1}c_{k-1} - p_k c_k), \quad k \geq 2, \\ c_1(t) = m - \displaystyle\sum_{k\geq 2} k c_i(t), \\ t^* = \inf\{t \geq 0 : c_N(t) = \rho m \mid c_1(0) = m\}. \end{cases}$$

Then, we remark that the transformation $\tilde{c}_i(\tau) = \frac{c_i(\tau)}{m}$, $\tau = t/m$ leads to

$$\begin{cases} \dfrac{d\tilde{c}_k}{d\tau} = \tilde{c}_1(p_{k-1}\tilde{c}_{k-1} - p_k\tilde{c}_k), \quad k \geq 2, \\ \tilde{c}_1(\tau) = 1 - \displaystyle\sum_{k\geq 2} k\tilde{c}_k(\tau), \\ \tau^* = \inf\{\tau \geq 0 : \tilde{c}_N(\tau) = \rho \mid \tilde{c}_1(0) = 1\}. \end{cases}$$

The right hand side in formulation of τ^* is independent of m (and is finite if ρ is small enough). Hence, if N, ρ, p_k are such that $\tau^* < \infty$, there exists a constant that depends only on N, ρ, p_i such that

$$t^* = \frac{C(N, \rho, p_i)}{m}.$$

We now use a similar strategy, but on the SBD model given by Eqs. (5) and (6). The number of clusters of size k, given by C_k, is rescaled into

$$D_k^M(t) = \frac{C_k(t/M^\gamma)}{M}, \tag{26}$$

with γ a scaling coefficient to be chosen later. Eq. (26) stands for the cluster size rescaled in time and in monomer size. From Eqs. (5) and (6), we obtain, by a simple change of variable, for any $t \geq 0$,

$$\begin{cases} D_1^M(t) = 1 - 2J_1^M(t) - \displaystyle\sum_{k\geq 2} J_k^M(t), \\ D_k^M(t) = J_{k-1}^M(t) - J_k^M(t), \quad k \geq 2, \end{cases} \tag{27}$$

with

$$J_k^M(t) = \frac{1}{M} Y_k^+ \Big(\int_0^t M^{2-\gamma} \frac{p_k}{V} D_1^M(s)(D_k^M(s) - \frac{\delta_{1k}}{M})ds \Big) - \frac{1}{M} Y_{k+1}^- \Big(\int_0^t M^{1-\gamma} q_{k+1} D_{k+1}^M(s)ds \Big), \quad k \geq 2. \tag{28}$$

We recall a standard convergence results for Poisson processes (a law of large numbers,[36] which will be useful in the following). A standard Poisson process at large times can be rescaled to obtain a deterministic process, that is,

$$\frac{1}{n} Y(nt) \to_{n\to\infty} t, \tag{29}$$

where convergence here means convergence in distribution. It is clear that the mean value of random variable $Y(nt)$ is nt. Fundamental result (29) states that fluctuations around this mean value are negligible compared to n, as $n \to \infty$. Such result can be generalized[36] for the solution of stochastic differential equations (27) and (28). Having the limiting model in our hands, we can deduce an approximation for the FAT and GFAT.

1. Large concentration c^0

Here, we set $V = 1$ and $c^0 = M \to \infty$. Using $\gamma = 1$ in Eq. (26), and the standard law of large numbers applied to Eqs. (27) and (28), can show[37] (see Section 3.1 of the supplementary material[49] for details) that the sequence of stochastic processes $(D_k^M(t))$ converges, as $M \to \infty$, in a rigorous way (in trajectory space) to the solution of the irreversible aggregation deterministic model (BD with $q_k = 0$), given, for all $t \geq 0$, by

$$\begin{cases} \dfrac{d}{dt} d_1 = -2j_1(t) - \displaystyle\sum_{k\geq 2} j_k(t), \\ \dfrac{d}{dt} d_k = j_{k-1}(t) - j_k(t), \qquad \text{for all } k \geq 2, \end{cases} \tag{30}$$

with

$$j_k(t) = p_k d_1 d_k(t), \text{ for all } k \geq 1, \tag{31}$$

and initial condition $d_1(0) = 1$ and $d_k(0) = 0$, for all $k \geq 2$. Intuitively, in the rescaled variable D_k^M, aggregation process is much more favorable compared to fragmentation because the number of free particles is large. By definition of the GFAT Eq. (8), with $h = 1$,

$$MT_{\rho,1}^{N,M} = \inf\{t \geq 0 : D_N^M(t) \geq \rho\}.$$

Then, using the convergence of $(D_k^M(t))$, we obtain the following asymptotic behavior of the GFAT for $h = 1$:

$$\lim_{M\to\infty} MT_{\rho,1}^{N,M} = \inf\{t \geq 0 : d_N(t) \geq \rho\}. \tag{32}$$

The right hand side in formulation of (32) is deterministic, and may be finite or infinite, according to the respective value of p_k, N and ρ. However, it is independent of M. If for deterministic models (30) and (31), p_k, N and ρ are such that $\inf\{t \geq 0 : d_N(t) \geq \rho\} < \infty$, then, for the GFAT of SBD models (5) and (6), we have

$$T_{\rho,1}^{N,M} \approx_{c^0=M\to\infty} \frac{C(p_k, N, \rho)}{M},$$

where $C(p_k, N, \rho)$ is a constant that depends only on p_k, N and ρ. Thus, $T_{\rho,1}^{N,M}$ is asymptotically inversely proportional to M and deterministic. This result is similar to the CMSBD model (Eq. (19)).

Limit models (30) and (31) do not capture the FAT and the GFAT $T_{\rho,h}^{N,M}$ for $h < 1$ (such an event is reached for time $t = 0^+$). However, as the initial number of monomers is large, we can derive a simple criterion to know the order of magnitude of the GFAT and to understand its variability. The following criterion will be confirmed by numerical simulations in Section IV.

First, limiting models (30) and (31) show that our SBD asymptotically behaves as a pure coagulation BD model. Such models (30) and (31) have been extensively studied,[38] where exact time-dependent solutions for $p_k = pk$ are given, and time asymptotic behaviors are given for power law coefficient $p_k = pk^\lambda$, $0 \leq \lambda \leq 1$. We restrict the following discussion to the constant rate case, $\lambda = 0$, for simplicity (results are analogous in the power law case). In such case, the stationary state of the pure coagulation BD model[17,38] (30) and (31) is $d_1^* = 0$ and

$$d_k^* = \frac{k-1}{ek!}, \quad k \geq 2. \tag{33}$$

Although the rescaled threshold ρM^{h-1} will be reached by d_N (and hence by D_N^M) for any ρ and $h < 1$ for large enough M (as $d_N^* > 0$), one can see that for "intermediate" M, we may have $Md_N^* \ll 1$, so that the threshold may not be reached before the free particles have vanished ($Md_1^* \approx 0$). In such a case, it is necessary to take into account the small but crucial contribution of the aggregate shortening. To this end, let us consider as a further approximation of Eqs. (27) and (28) the following deterministic model, given, for all $t \geq 0$, by Eq. (30) and fluxes defined as

$$j_k(t) = pd_1(t)d_k(t) - \frac{1}{M}qd_{k+1}(t), \quad k \geq 1, \tag{34}$$

where M is large enough such that $1/M$ is a small parameter. We detail in Section 3.2 of the supplementary material[49] the successive relevant time scales of deterministic models (30)-(34) which are also illustrated with Fig. 2 (see also Figures S7 and S8 of the supplementary material[49]). In particular, it is known[17] that under the favorable aggregation limit $q/M \ll p$, our deterministic BD model, Eqs. (30)-(34),

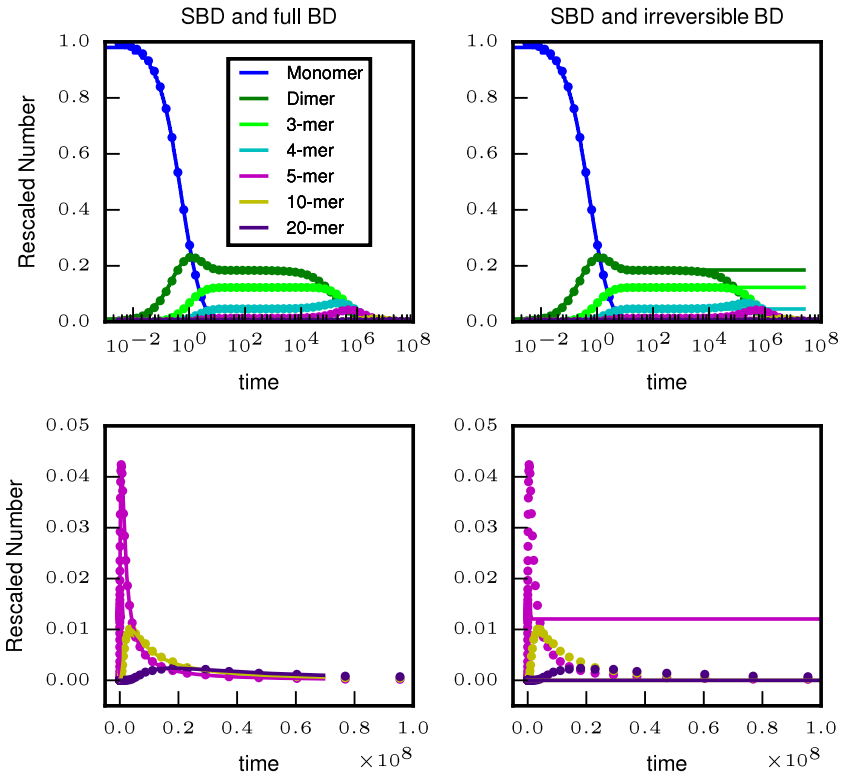

FIG. 2. SBD and BD trajectories. For the SBD, we simulate rescaled equations (27) and (28), with $M = 10^{5.5}$, $V = 1$, and kinetic rates are $p_k \equiv 1$ for all $k \geq 1$ and $q_k \equiv 1$ for all $k \geq 2$. For the BD model, we simulate on the left columns full BD Eqs. (30) and (34) and on the right columns irreversible BD Eqs. (30) and (31). The rescaled SBD trajectories are plotted with filled circles, together with the corresponding BD trajectories in plain lines, for the monomer and i-cluster, $i = 2, 3, 4, 5, 10, 20$, according to the legend. The lower panel corresponds to the same numerical simulation of the upper panel, with a zoom on the y-axis to improve the visualization of the $i-cluster$ for $i = 5, 10, 20$. It is immediate to see that full BD Eqs. (30) and (34) agree perfectly with rescaled SBD equations (27) and (28) for all time, while irreversible BD Eqs. (30) and (31) match only up to a time scale of order M.

has a metastable state, reached in a time scale of order $\log(M)$, in which concentrations of each species of size $k \geq 2$ are nearly constant, equal to d_k^*, the equilibrium state (Eq. (33)) of the irreversible aggregation BD model, Eqs. (30) and (31). The concentration $d_k(t)$ stays roughly constant for a time of order $log(M)$, and relax to the steady-state values of the full BD Eqs. (30)-(34) in a time of order M only. During the metastable period, monomer concentration is also nearly constant, given by Ref. 17 (see Section 3.2 of the supplementary material for detailed calculations and Figure S10 of the supplementary material[49] for illustration),

$$d_1^* = \frac{q}{M}\frac{d_2^* + \sum_{k\geq 2} d_k^*}{p\sum_{k\geq 2} d_k^*} = \frac{3}{2}\frac{q}{pM}.$$

Thus, when looking at the original variable, the cluster number $C_N(t)$ given by the SBD model will reach the metastable value $c_N^* = Md_N^*$ in a very short time $(\log(M)/M)$. If M is large enough, the metastable state will be large enough to reach ρM^h. In that case, the GFAT $T_{\rho,h}^{N,M}$ is found to behave as linear CMSBD model (10) with $C_1 = M$ (see numerical section). If M takes intermediate values, such that $Md_N^* < \rho M^h$, however, one needs to wait longer for aggregate shortening to produce more critical clusters. As the initial pure-aggregation phase is short, we can neglect it and use the metastable values $c_k^* = Md_k^*$ as initial values for a linear CMSBD model (10), where the monomer number is now equal to $C_1 \equiv c_1^*$ given by

$$c_1^* = Md_1^* = \frac{3}{2}\frac{q}{p}.$$

The metastable state c_1^* is independent of the initial number of monomers M and is of order q/p. Thus, the GFAT depends on M only through the initial condition $c_k^* = Md_k^*$, $k \geq 2$, and is found to be (see numerical results, Section IV A) almost independent of M over several orders of magnitude for $N \geq 15$. Finally, note that there is always a probability for the threshold to be reached before the metastable period, which is responsible for the bimodal behavior of $T_{\rho,h}^{N,M}$ seen in numerical results. For values of d_k^* and a summary of the different cases, see Tables I and II.

TABLE I. Normalized metastable values d_k^* for the deterministic BD model ((30)-(34)) in the favorable aggregation case $pM \gg q$. In this table, we compute the numerical values of the normalized metastable values d_i^* for the deterministic BD model ((30)-(34)) with constant kinetic rate $p_k \equiv p$ and $q_k \equiv q$ in the favorable case $pM \gg q$. Such values represent the level that each variable reach during the metastable period after the pure-aggregation period. It is given by the equilibrium value of the irreversible BD model ((30) and (31)), see Eq. (33). See text in Subsection III C 1.

Size	Value	Size	Value
d_2^*	0.1839	d_{10}^*	9.1240×10^{-7}
d_3^*	0.1226	d_{11}^*	9.2162×10^{-8}
d_4^*	0.0460	d_{12}^*	8.4481×10^{-9}
d_5^*	0.0123	d_{13}^*	7.0894×10^{-10}
d_6^*	0.0026	d_{14}^*	5.4858×10^{-11}
d_7^*	4.3795×10^{-4}	d_{15}^*	3.9385×10^{-12}
d_8^*	6.3868×10^{-5}	d_{20}^*	2.8730×10^{-18}
d_9^*	8.1102×10^{-6}	d_{50}^*	5.9269×10^{-64}

2. Large volume V

We now deal with the case where c^0 is set to 1 and $V = M \to \infty$. We recall that in this case the limit $M \to \infty$ is to be understood as a volume expansion, and the reaction rates must be scaled with the volume according to their respective order. In particular, it is classical[30] to assume the first-order reaction rates to be independent of the volume, and the second-order reaction rates to be inversely proportional to the volume. With $\gamma = 0$, the re-scaled variable $D_k^M(t) = C_k(t)/M$ now converges to the BD system given, for all $t \geq 0$, by Eq. (30) and flux definition

$$j_k(t) = p_k d_1 d_k(t) - q_k d_{k+1}(t), \quad k \geq 1. \tag{35}$$

TABLE II. Summary of the First Assembly Time (FAT) and Generalized First Assembly Time (GFAT) findings in this paper. Analytical and numerical results. In this table, we sum up the different analytical findings on the FAT, for the full Stochastic Becker-Döring (SBD), Eqs. (5) and (6) and its two simplifications with constant monomer (CMSBD), Eq. (10) and single cluster (SCSBD), Eq. (20). The first column denotes which model is considered with which scaling. The second column provides in which asymptotic the results are valid (we recall that $M = c^0 V$). The third column gives the slope of the log-log dependence of the GFAT with respect to M (except for the SCSBD and SBD with $N = \sqrt{M}$ where exponential large deviation occurs). The last column gives the full distribution of the FAT (if known). See text for more details.

Model	Condition	M (log-log) dependence	Distribution
CMSBD	$c^0 \to \infty$	$-(1+(1-h)/(N-1))$	Weibull
CMSBD	$V \to \infty$	$-(1-h)/(N-1)$	Weibull
SCSBD	$q \to \infty$	$-N$	Exponential
SCSBD	$c^0 \to \infty$	-1	
SCSBD	$V \to \infty$	0 (finite sum)	
SCSBD	$N = \sqrt{M}, V \to \infty, q_k > p_k$	$Me^{\sqrt{M}}$	Expo. Large deviation
SBD	$c^0 \to \infty$	$-(1+(1-h)/(N-1))$	Weibull
SBD	$c^0 \gg 1, Md_N^* \ll 1$	~ 0	Bimodal
SBD	$V \to \infty$	$-(1-h)/(N-1)$	Weibull
SBD	$N = \sqrt{M}, c^0 \to \infty$	$-1/2$	
SBD	$N = \sqrt{M}, V \to \infty$	$1/2$	
SBD	$N = \sqrt{M}, V \to \infty, q_k > p_k$	$Me^{\sqrt{M}}$	Expo. Large deviation

As before, using the convergence of $(D_k^M(t))$, we obtain the following asymptotic behavior of the GFAT for $h = 1$:

$$\lim_{M\to\infty} T_{\rho,1}^{N,M} = \inf\{t \geq 0 : d_N(t) \geq \rho\}.$$

Once again, the latter quantity is deterministic, and may be finite or infinite, according to the respective value of q_k, p_k, N, and ρ. Thus, in this scenario, if for the deterministic model (Eqs. (30)-(35)), q_k, p_k, N, and ρ are such that $\inf\{t \geq 0 : d_N(t) \geq \rho\} < \infty$, the GFAT $T_{\rho,1}^{N,M}$ is asymptotically independent of M.

As in the first scenario, for very large M, the GFAT $T_{\rho,h}^{N,M}$ with $h < 1$ behaves asymptotically as the GFAT of linear CMSBD model (10) with $C_1 \equiv M = V$. Thus,

$$\langle T_{\rho,h}^{N,M}\rangle \approx_{M=V\to\infty} C(p,N)\frac{1}{M^{(1-h)/(N-1)}}, \tag{36}$$

where $C(p,n)$ is a constant that depends only on N and p_k, $k \leq N$.

D. Full SBD, large *M*, and large *N*

In this section, we investigate the behavior of the SBD and its FAT when the size N of the maximal cluster is large, and scales with the total number of particles M. As in section B, we will use the natural rescaling variable $x = k/N$. We distinguish again two scenarios, which yield distinct results. In the first one, the volume V is fixed and the concentration c^0 is large. In the second one, the concentration c^0 is fixed and the volume V is large. In both cases, a rescaling of the solution is found to be solution of a deterministic continuous size model, namely, the Lifschitz-Slyozov model (LS). The LS model is a partial differential equation of transport type, which arises naturally in the study of BD model,[40,41] when cluster sizes change in small steps. Indeed, we have detailed in a companion paper how to choose a proper scaling and how to derive the limit equation for that rescaled solution.[39] We show here the consistency of this scaling with the behavior of the GFAT, which will be confirmed in Section IV by numerical simulations.

We will look at the case $N = \sqrt{M}$. We define the rescaled measure on $\mathbb{R}^+$,

$$\mu^M(t,dx) = \sum_{k\geq 2} \frac{C_k(t/M^\gamma)}{\sqrt{M}}\delta_{k/\sqrt{M}}(dx), \tag{37}$$

and $C_1^M(t) = C_1(t/M^\gamma)/M$, where $\delta_x(\cdot)$ is the Dirac measure at x. The GFAT $T_{\rho,h}^{\sqrt{M},M}$ involves an increasing maximal size $\sqrt{M}$, which is rescaled to the macroscopic size $x = 1$ by the definition of the measure μ^M in Eq. (37). We also need to define corresponding macroscopic aggregation and fragmentation rates, using

$$\begin{cases} p^M(x) = \sum_{k\geq 2} p_k \mathbf{1}_{[k/\sqrt{M},(k+1)/\sqrt{M})}(x), \\ q^M(x) = \sum_{k\geq 2} q_k \mathbf{1}_{[k/\sqrt{M},(k+1)/\sqrt{M})}(x), \end{cases} \tag{38}$$

where $\mathbf{1}_I(\cdot)$ is the characteristic function that is equal to 1 in set I and 0 outside.

1. $N \to \infty$, large concentration c^0

Using $\gamma = 1/2$, a fixed volume $V = 1$, and a rescaled[50] nucleation rate $p_1 = \frac{\overline{p}_1}{c^0}$, we have shown that the measure μ^M satisfies

$$\lim_{M\to\infty} \mu^M(t,dx) = f(t,x)dx,$$

where f is a density, solution of the irreversible LS coagulation model[39] (see Section 4 of the supplementary material[49] for detailed calculations). The LS model is given for all $t \geq 0$, by

$$\begin{cases} \frac{\partial}{\partial t} f(t,x) + \frac{\partial}{\partial x}(p(x)c_1(t)f(t,x)) = 0, \quad \forall x > 0, \\ c_1(t) + \int_0^\infty x f(t,x)dx = 1, \\ \lim_{x\to 0^+}(p(x)f(t,x)) = \overline{p}_1 c_1(t), \end{cases} \tag{39}$$

with initial condition $c_1(0) = 1$ and $f(0,\cdot) = 0$, and where $p(x)$ is the limit of the macroscopic coagulation rate $p^M(x)$ defined in Eq. (38). Eq. (39) is a transport partial differential equation with ingoing characteristics at $x = 0^+$ and is well defined if a boundary condition at $x = 0$ is given. We refer to the paper[39] for the choice of the boundary condition (that depends on the scaling used in Eq. (37) and the scaling of the reaction rates). The large cluster $C_k(t)$ for $k = \sqrt{M}$ is thus approximated by $f(\sqrt{M}t, 1)$, which yields

$$\sqrt{M}T_{\rho,h}^{\sqrt{M},M} \approx_{M=c^0\to\infty} \inf\{t \geq 0 : f(t,1) \geq \rho M^{h-1/2}\}. \tag{40}$$

For any $h \leq 1/2$, the right hand side in Eq. (40) is deterministic, and may be finite or infinite, according to the macroscopic coagulation rate p and the threshold ρ.

2. $N \to \infty$, Large volume V

Finally, if we consider a fixed concentration $c_0 = 1$ and large volume $V = M \to \infty$, and a rescaled nucleation rate $p_1 = \frac{\overline{p}_1}{V}$. Then, using $\gamma = -1/2$, we have shown in Ref. 39 that the measure μ^M satisfies

$$\lim_{M\to\infty} \mu^M(t,dx) = f(t,x)dx,$$

where f is a density, solution of the LS coagulation-fragmentation model given, for all $t \geq 0$, by

$$\begin{cases} \frac{\partial}{\partial t} f(t,x) + \frac{\partial}{\partial x}\left[(p(x)c_1(t) - q(x))f(t,x)\right] = 0, \\ c_1(t) + \int_0^\infty x f(t,x)dx = 1, \\ \lim_{x\to 0^+}(x^r f(t,x)) = c_1(t), \end{cases} \tag{41}$$

with initial condition $c_1(0) = 1$ and $f(0,\cdot) = 0$, and where $p(x), q(x)$ are the limits of the macroscopic rate $p^M(x), q^M(x)$ defined in Eq. (38), and $r \in [0,1]$ is determined through the relation $p(x) \approx_{x\to 0} x^r$. Again, such Eq. (41) is well-defined if a boundary condition at $x = 0$ is given when the characteristics are ingoing at $x = 0^+$. The large cluster $C_k(t)$ for $k = \sqrt{M}$ is now approximated by $f(t/\sqrt{M}, 1)$, so that

$$\frac{1}{\sqrt{M}}T_{\rho,h}^{\sqrt{M},M} \approx_{M=V\to\infty} \inf\{t \geq 0 : f(t,1) \geq \rho M^{h-1/2}\}. \tag{42}$$

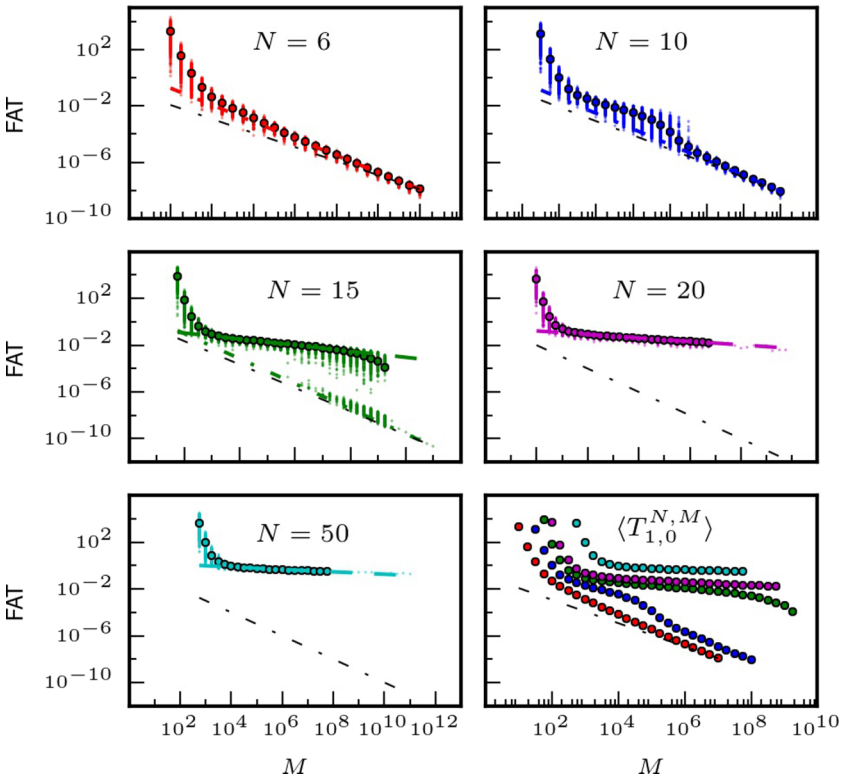

FIG. 3. First assembly time $T_{1,0}^{N,M}$ for the original SBD (Section III C 1) as a function of the total mass M (in log-log scale) for five different maximal cluster sizes $N \in \{6, 10, 15, 20, 50\}$, and $V = 1$. Each color light dot is a single realization of the FAT. For each condition, large circles represent the statistical mean over 1000 samples (a few condition are sampled only once, namely, for $N = 15, 20, 50$ and large M, for which the mean is not shown). Black dashed-dotted lines are straight lines of slope -1, color dashed-dotted lines are straight lines of slope $-(1+1/(N-1))$ (as in Eq. (15)). And for $N = 15, 20, 50$ we plot additionally dashed lines of slope, respectively, $-0.26, -0.15$, and -0.10. The last panel in bottom-right represents the 5 mean FATs on the same scale. Kinetic rates are $p_1 = 0.5$, $p_k \equiv 1$, and $q_k \equiv 100$ for all $k \geq 2$.

Again, for any $h \leq 1/2$, the latter quantity in Eq. (42) is deterministic, and may be finite or infinite, according to the macroscopic rates p, q, and ρ.

The results of Subsections III D 1 and III D 2 are illustrated below in Section IV with the help of numerical simulations. Note that for particular choice of rates p and q, one is able to obtain analytically time-dependent solution of Eqs. (39) and (41) (see Section 4.1 of the supplementary material[49] for detailed calculations).

IV. SIMULATIONS AND ANALYSIS

In this section, we present results derived from simulations of the SBD model associated to stochastic equations (5) and (6), for various values of its key parameters $\{M, N, p_k, q_k\}$ and volume V. Specifically, we use an exact stochastic simulation algorithm (kinetic Monte Carlo, KMC) to compute first assembly times.[42–44] For each set of $\{M, N, V, p_k, q_k\}$, we sample 10^3 trajectories (except for few cases where sampling so many trajectories was out of reach in terms of computational time) and follow the time evolution of the clusters until the threshold is reached, at which point simulation is stopped and the FAT/GFAT recorded. We compare and contrast our numerical results with the theoretical findings of Sec. III. The following is divided into four subsections that correspond to four main results on the FAT and the GFAT. In all figures from Figs. 3 and 9, we represent each realization of the FAT (respectively, GFAT) in light dot together with its empirical mean in large dot. We superpose on top to it the relevant analytical curves to illustrate the consistency with the theoretical findings.

A. The first assembly time can be weakly dependent on the total number of monomer *M*, and highly variable even in large population

We begin with the analysis of the FAT as a function of the total number of monomers M, when maximal cluster size N and volume V are fixed. For $N = 6, 10, 15$, prediction of the asymptotic behavior of $T_{1,0}^{N,M}$, time needed for a single maximal cluster to be formed, is verified: the mean FAT decreases linearly in log-log scale as M increases, with a slope equal to $-(1 + 1/(N-1))$, as in the linear CMSBD model (Fig. 3, upper panels), see Eq. (15). The CV that measures variability of the FAT is close to 1 for small M and close to the predicted value by Eq. (18) (Fig. 4, and Figure S1 of the supplementary material[49] for the CMSBD) for very large M. This fact is consistent with a transition

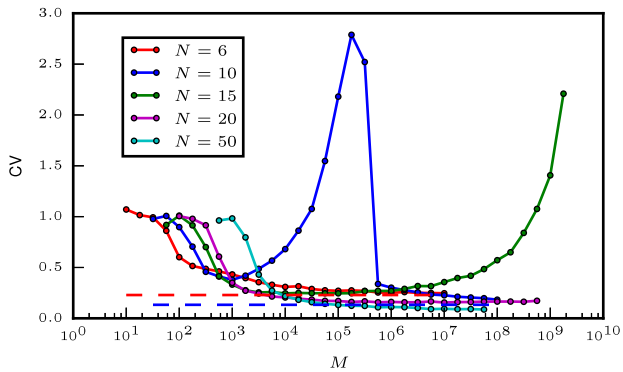

FIG. 4. Coefficient of Variation (CV) for the first assembly time $T_{1,0}^{N,M}$ as a function of the total mass M corresponding to the realizations of Fig. 3. For $N = 6, 10$, we plot additionally horizontal dashed lines at the value predicted by the Weibull distribution, see Eq. (18).

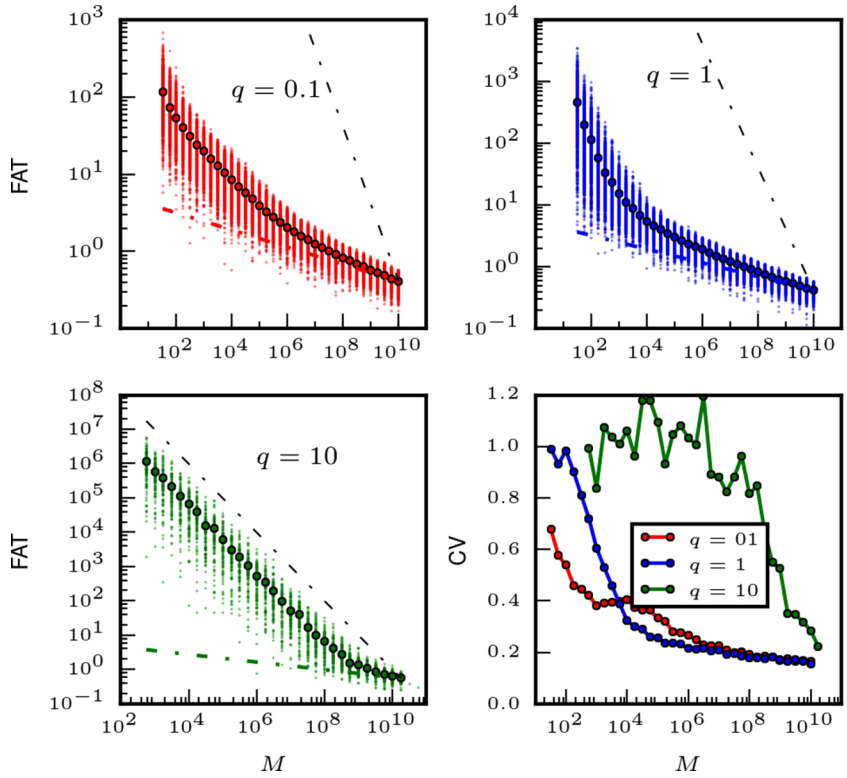

FIG. 5. First assembly time $T_{1,0}^{N,M}$ for the SBD as a function of the total mass M (in log-log scale) when the concentration $c^0 = 1$ is fixed and $M = V$ (Section III C 2) for three different detachment rates $q \in \{0.1, 1, 10\}$, and $N = 10$. Kinetic rates are $p_1 = 0.5$, and $p_k \equiv 1$ and $q_k \equiv q$ for all $k \geq 2$. Each color light dot is a single realization of the FAT. For each condition, large circles represent the statistical mean over 1000 samples. Black dashed-dotted lines are straight lines of slope -1, color dashed-dotted lines are straight lines of slope $-1/(N-1)$. Finally, the panel in bottom right represent the Coefficient of Variation (CV) as a function of the total mass M corresponding to the realizations of the first three panels (top and bottom left).

from an exponential distribution to a Weibull distribution as M gets large. However, the CV for the FAT for $N = 10$ is non-monotonic and has a large peak at intermediate values of M (Fig. 4). The same behavior is suspected for $N = 15$ (and $N = 20, 50$) but could not be verified due to numerical cost. Corresponding to this peak for the cv, one can observe very clearly for $N = 10, 15$ the bimodal behavior predicted for large but intermediate M values (Fig. 3, second and third panels). For instance, for $N = 15$ and M ranging from 10^6 to 10^{10}, the sampled FAT segregates between two groups separated by several orders of magnitude (one group below 10^{-6}, one group around $10^{-2}, 10^{-3}$). The higher values of the sampled FAT correspond to trajectories that went through the threshold $C_N = 1$ after the metastable period described in Sec. III C 1. For $N = 20$ and $N = 50$, we could simulate in a reasonable computation time (several weeks) only up to $M = 10^{13}$ and $M = 10^{11}$, respectively. Below these values, the metastable states computed in Table I predict that the threshold will be mostly reached after the metastable period, which explains the large "plateau" observed for the FAT up to M^{13} (respectively, M^{11}): the FAT is nearly independent of M over a broad range of values (Fig. 3, the slopes for $N = 15, 20, 50$ are, respectively, approximatively $-0.26, -0.15, -0.10$). Trajectories of the number of cluster as a function of time help to visualize the different phases. We illustrate in Figures S7 and S8 of the supplementary material,[49] stochastic trajectories of the SBD model together with the favorable aggregation limit of the deterministic BD model, in order to clearly identify the metastable period. In Figures S9 and S10 of the supplementary material,[49] we exhibit two trajectories of the stochastic SBD model that results in two FATs that differ from several logs of order of

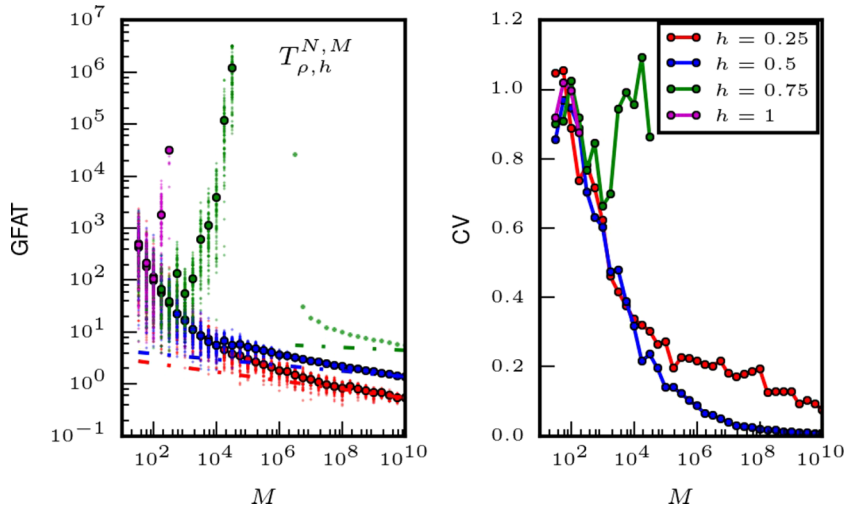

FIG. 6. (Left) Generalized first assembly time $T_{\rho,h}^{N,M}$ for the SBD as a function of the total mass M (in log-log scale) when the conce ntration $c^0 = 1$ is fixed and $M = V$ (Section III C 2), for $h \in \{0.25, 0.5, 0.75, 1\}$, and $N = 10$. Kinetic rates are $p_1 = 0.5$, $p_k \equiv 1$, and $q_k \equiv 1$ for all $k \geq 2$. Each color light dot is a single realization of the GFAT. For each condition, large circles represent the statistical mean over 1000 samples (a few conditions are sampled only once, namely, for $h = 0.75$ and large M, for which the mean is not shown). Color dashed-dotted lines are straight lines of slope $-(1-h)/(N-1)$. (Right) Coefficient of Variation (CV) as a function of the total mass M corresponding to the realizations of the left panel.

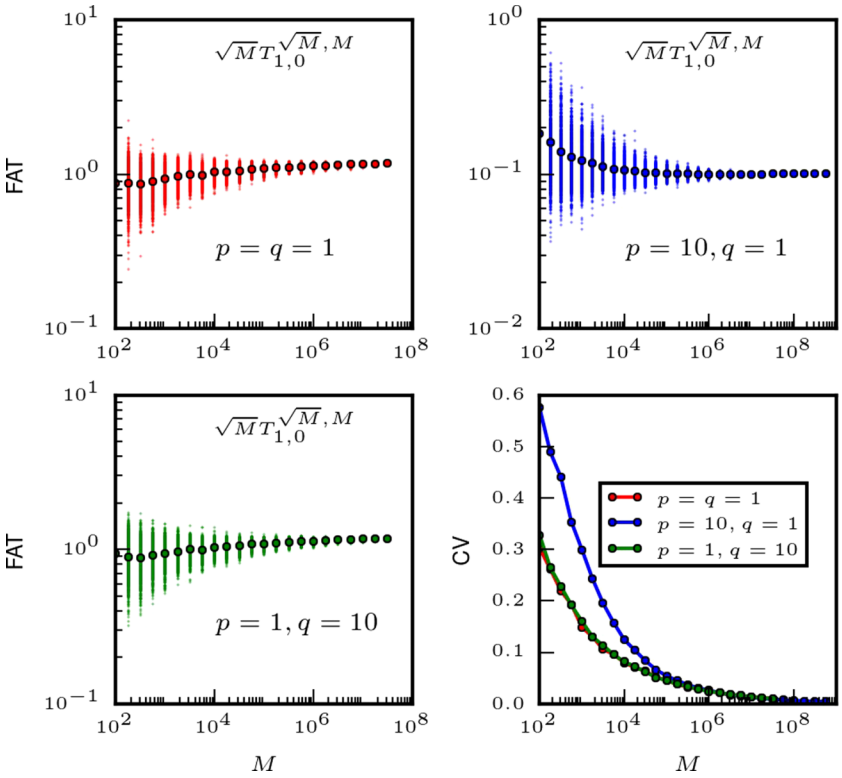

FIG. 7. First assembly time $T_{1,0}^{\sqrt{M},M}$ for the original SBD and large maximal cluster size of order $N = \sqrt{M}$ as a function of the total mass M (in log-log scale) when the volume $V = 1$ is fixed (Section III D 1) for three different kinetic rates $(p_k, q_k) \in \{(1,1);(10,1);(1,10)\}$. The FAT is multiplied by $\sqrt{M}$ to verify the scaling in Eq. (40). Finally, the panel in bottom right represents the Coefficient of Variation (CV) as a function of the total mass M corresponding to the realizations of the first three panels (top and bottom left).

magnitude, due to the metastable period. We also point out the accuracy of the approximation by a linear model that has metastable state for initial condition. Finally, the transition from an exponential distribution to a Weibull distribution as M increases, through an intermediate bimodal distribution, is illustrated on histograms of the FAT over 10^3 realization in Figure S11 of the supplementary material.[49]

Similar results are obtained for the GFAT $T_{\rho,h}^{N,M}$, where the linear log-dependence with a slope $-(1 + (1-h)/(N-1))$ (see Eq. (19) for the CMSBD model) is found to be perfectly satisfied for $N = 3, 5$ and $h = 0.25, 0.5, 0.75$ and $h = 1$ (Figure S5 of the supplementary material,[49] upper panels). Bimodal behavior and nearly flat log-dependence of the GFAT $T_{\rho,h}^{N,M}$ as a function of M on a broad range of M values are also observed for $N = 10, 20$ (Figure S5 of the supplementary material,[49] lower panels). The size of the bimodal region is found to be increased with increasing h (and N). For $N = 10, 20$ and $h = 1$, the mean FAT is increasing to ∞ as the deterministic limit given by Eq. (32) is infinite. The CV is non-monotonic with respect to M with a peak corresponding to the bimodal behavior (Figure S6 of the supplementary material[49]). We show that the GFAT has a lower variability as h increases, and vanishes for $h = 1$ and large M, in agreement with the deterministic limit in Eq. (32).

B. The first assembly time is non-monotonic with respect to the detachment rate

We verify in Figures S12-S14 at Ref. 49 the previously published[6] dependence of the FAT on the detachment rate. We confirm that the bimodal behavior is observed for *small* detachment rate, and that the mean FAT (and the cv) is a non-monotonic function of the detachment rate.

C. The generalized first assembly time may increase with M for large volume

When the concentration c^0 is fixed, and the volume V increases together with the total number of monomers M (see Section III C 2), the FAT to reach a maximal cluster of fixed size N decreases monotonically with M, and asymptotically with a linear log-dependence with a slope $1/(N-1)$ (Fig. 5), as predicted by Eq. (36). The same observation is valid for the GFAT $T_{\rho,h}^{N,M}$, for $h < 1$, with a slope $(1-h)/(N-1)$ (Fig. 6). However, for $h = 1$, if the threshold ρ is too large, the GFAT is never reached by the deterministic BD model ((30)-(35)). Thus, for the finite SBD, the GFAT for $h = 1$ increases to ∞ as M increases to ∞. For $h = 0.75$, we also found that the GFAT is *non-monotonic* with respect to the total number of monomers, even though it converges to 0 for (very) large volume and number of monomers.

D. Exponentially Large FAT for large maximal cluster size N and phase-transition phenomena

Finally, for large maximal size N, of order $\sqrt{M}$, we illustrate the validity of the two scalings in Eqs. (40)-(42). Specifically, in Fig. 7, we see that for $M > 10^6$, the FAT is nearly deterministic and can then be predicted by limit model Eq. (39). The same threshold is empirically observed in Fig. 8 for the GFAT as well. However, considering $p(x) = x$ and $q(x) = 1$, in Fig. 9, we show that exponential large deviation in the large volume limit may occur if the aggregation is not favorable compared to the fragmentation, as in SCSBD model (20) (Figure S4 of the supplementary material[49]). Indeed, in such a case, deterministic limit (41) predicts that the FAT is never reached (is infinite) as the drift of the transport equation is negative for small (macroscopic) size x.

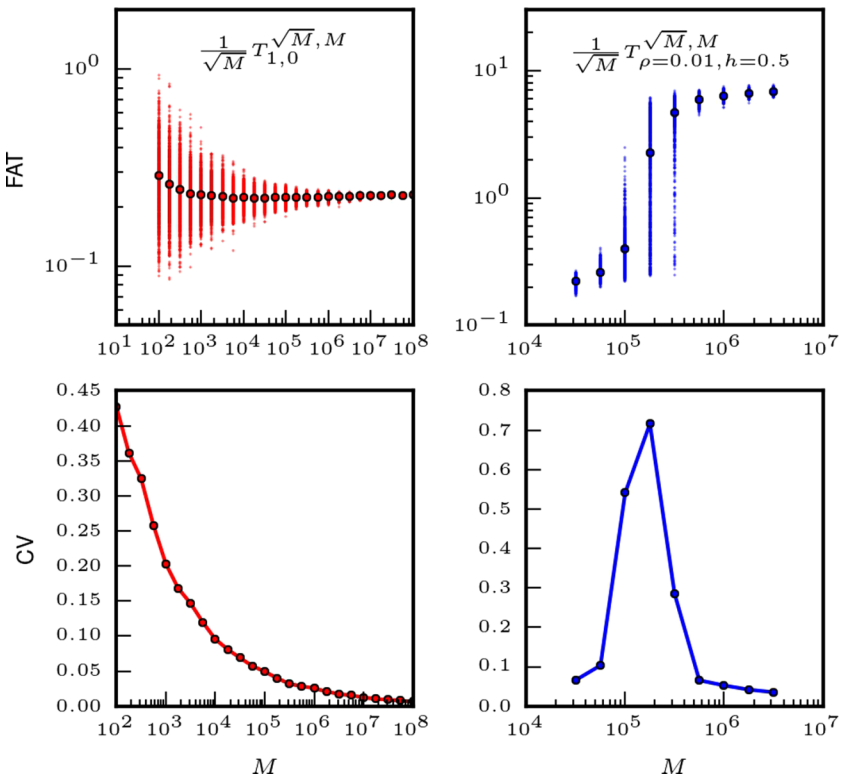

FIG. 8. First assembly time $T_{1,0}^{\sqrt{M},M}$ (top left) and generalized first assembly time $T_{\rho,h}^{\sqrt{M},M}$ (top right) for the SBD and large maximal cluster size of order $N=\sqrt{M}$ as a function of the total mass M (in log-log scale) when the concentration $c^0=1$ is fixed and $M=V$ (Section III D 2). Kinetics rates are $p(x)\equiv 5$ and $q(x)=x$. Both the FAT and the GFAT are divided by $\sqrt{M}$ to verify the scaling in Eq. (42). Finally, the panels in bottom left and right represent the Coefficient of Variation (CV) as a function of the total mass M corresponding to the realizations of the upper panels.

For the finite system, the FAT grows exponentially fast with M, in agreement with Eq. (23). On the right panels in Fig. 9, we show few time-dependent trajectories that are representative of a phase-transition phenomenon, with a very abrupt change of phase, occurring at a widely variable time (the CV is near 1). Although the deterministic limit predicts that the aggregation will *not* take place (and the monomer number will not decrease), in the SBD model the aggregation is *always* complete (no monomer at the end), but at very large time values as volume V and initial number of monomers M increase. The phase-transition is associated with the emergence of a single cluster that aggregates at the expense of smaller clusters, giving a stochastic description of the Ostwald ripening theory.

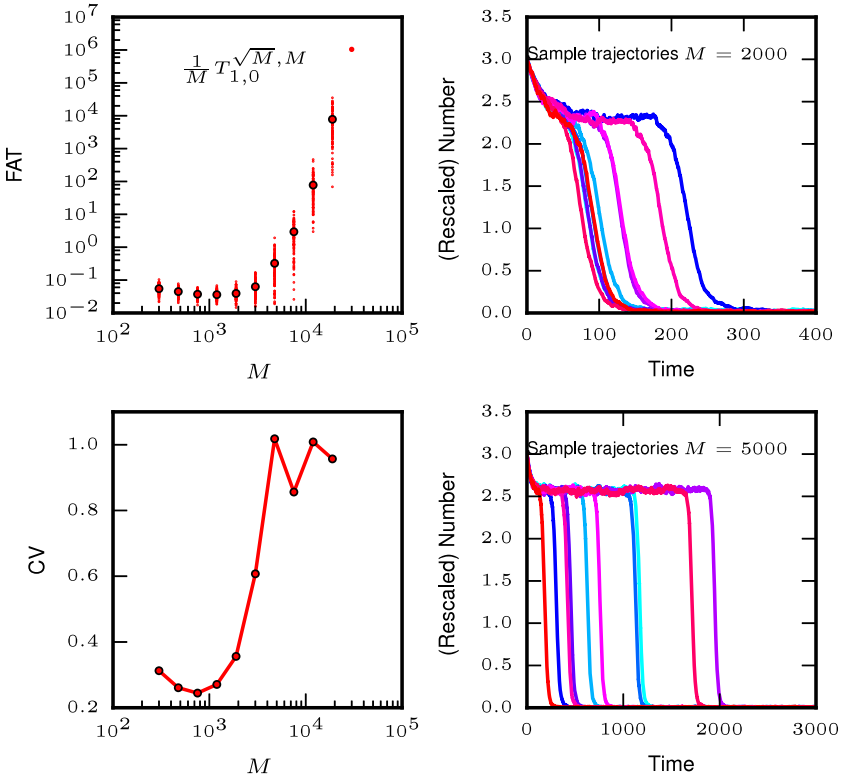

FIG. 9. (Top left) First assembly time $T_{1,0}^{\sqrt{M},M}$ for the rescaled SBD and large maximal cluster size of order $N=\sqrt{M}$ as a function of the total mass M (in log-log scale) when the concentration $c^0=1$ is fixed and $M=V$ (Section III D 2). Kinetics rates are $p(x)=x$ and $q(x)=1$. (Bottom Left) Coefficient of Variation (CV) as a function of the total mass M corresponding to the realizations of the upper left panel. (Top and bottom right) Time-dependent trajectories of the rescaled number of monomers $c_1(t)=C_1(t)/M$, for $M=2000$ (top) and $M=5000$ (down). Each color line represents a single realization with the same initial condition and kinetic parameter.

V. SUMMARY AND CONCLUSIONS

We have studied the problem of determining the FAT of a cluster of a pre-determined size N from an initial pool of M independent monomers characterized by size-dependent attachment and detachment rates p_k and q_k, respectively. We have developed a full stochastic approach, based on the SBD equations.

Up to our knowledge, the early work[20] paves the way to study fluctuations of the time-dependent cluster distributions and first passage time in finite system nucleation models. In particular, authors of this works estimate rates of individual cluster growth and shrinkage with physics arguments, they show numerical simulations for a special case, and heuristically derive a moment closure through deterministic approximation of the master equation governing the cluster distribution evolution. They also investigated reaction rates of the form $p_k \approx k^{2/3}$ and $q_k \approx k^{2/3}e^{-k^{1/3}}$, for which a critical cluster size exists, as in the Ostwald ripening theory. This case has also been considered in the present work and developed in detail in Sections IV B and IV D (where we quantified the mean FAT and explained the observed variability). It is important to note that the authors of this previous work noticed that the first passage time was subjected to large fluctuations. Later on, this approach was extended[23] by exploring numerical ways to solve the master equation, and by being focused on discrepancies between the classical deterministic nucleation rate (expressed as a particular flux j_{k^*}, see Eq. (4)) and its stochastic analog using moment equations. Note that if first passage time and nucleation rate are clearly linked, their precise relationship is not trivial. Also, we mention that Chapter 8 of Ref. 21 contains useful results on the equilibrium distribution for general clustering process, proven with the use of a queueing network theory approach. Finally, the recent work[6] presents in detail the first passage time $T_{1,0}^{N,M}$ behavior for constant kinetic rates $p_k \equiv p$, $q_k \equiv q$, and finite N, M, as a function of the detachment rate q.

In this study, we started by two simplified models and were able to find exact results for the FAT statistics for general values of M, N and p_k, q_k.

The first simplification was to consider that the number of monomers stayed constant over time (linear CMSBD model). The mean FAT was found to be a decreasing function of the total number of monomers M, with an asymptotic log-linear dependence with respect to M, with negative slope equal to $-(1 + 1/(N-1))$ (see Eq. (15)). Importantly, the coefficient of variation of the FAT was found to be asymptotically a positive value depending only on N (see Eq. (18)). Finally, the full distribution of the FAT is known for large M and is given by a Weibull distribution. We generalized our results for the mean GFAT, which was also found to be a decreasing function of M, with a log-linear dependence, with negative slope equal to $-(1 + (1-h)/(N-1))$, for any $h \in [0,1]$, and any N.

The second simplification was to consider that only a single cluster could be formed at a time (SCSBD). Here, an analytical formula for the mean FAT (and higher moments) is available, thanks to first passage time theory on one-dimensional random walk. The mean FAT depends on sums of products of ratio $q_{k+1}/((M-k)p_k)$ (see Eq. (21)). From this analytical formula, we gave several asymptotic ones that are important to understand the full SBD model. First, the mean FAT increases to infinity as (q_k) increases to infinity, with leading order given by $1/(p_1M^2)\prod_{k=2}^{N-1} q_k/(p_k(M-k))$, for any N. In such a case (for large detachment rate q_k), the full FAT distribution was found to be asymptotically an exponential distribution. In the opposite asymptotically favorable aggregation case, $M \to \infty$, for large concentration and fixed volume, the mean FAT is roughly inversely proportional to M, and the normalized FAT distribution given by a sum of $N-2$ exponential random variables of parameter p_i, $2 \le i \le N-1$. Finally, we studied the FAT of a very large cluster of size $N \to \infty$ for the SCSBD model. We transformed the discrete sum formula (see Eq. (21)) into continuous integrals, in order to get simple and easily computed asymptotic formulas. For $N = \alpha M$, $\alpha < 1$, in the limit of large concentration, the required mean time to form a cluster of size N becomes independent of M and is given by a simple continuous integral that depends only on the forward aggregation rate and α (see Eq. (24)). For the large volume scenario, for fixed concentrations, we found in contrast that for unfavorable aggregation ($q_k > p_k$), the mean time of formation of a large cluster takes an exponentially large time as the volume V increases to infinity.

With the analytical results on the two simplified models in mind, we analyzed the behavior of the FAT for the full SBD model. Using a rescaling strategy, as the total number of monomers M increases to infinity, we found asymptotic expression of the mean FAT and GFAT as a function of a first passage time associated to deterministic models, namely, the discrete-size BD model and the continuous-size LS model. This way we are able to find quickly the order of magnitude of the FAT (respectively, GFAT) with the help of a single (fast) numerical simulation of a deterministic model (rather than by extensive numerical simulation of the full SBD model). With the help of a careful time scale analysis, and with extensive numerical simulation, we also pointed out surprising deviations from the mean field deterministic model. Hence, in the limit of large concentrations, as expected, the time to form a macroscopic quantity (a positive fraction ρ of the total mass M) of clusters of size N is asymptotically deterministic and linked with a corresponding first passage time of the deterministic BD model. However, the time to obtain any "smaller" quantity (ρM^h, $h < 1$) is not well captured by the deterministic BD model. Indeed, the FAT $T_{1,0}^{N,M}$ and the GFAT $T_{\rho,h}^{N,M}$ for $h < 1$ decays to 0 at a speed faster than M (given asymptotically by $1/M^{1+(1-h)/(N-1)}$) and has a non-vanishing coefficient of variation as $M \to \infty$. This was explained by analogy with the simpler linear CMSBD model. Importantly, for moderately large maximal cluster size ($N \ge 15$), the mean FAT is found to be only weakly dependent on the total number of monomers M, and so for several orders of magnitude of intermediate values of M (from 10^6 to 10^{13} in our simulations). The coefficient of variation is much larger than 1 on this parameter region and the full distribution of the FAT is bimodal. We explained and gave a practical criterion (given by the comparison of d_N^* and ρM^{h-1}, see Eq. (33)) for this phenomenon to occur by a careful inspection of the metastable state of the favorable

aggregation limit for the deterministic BD model. Also, for the case corresponding to large volume experiments, and fixed concentrations, we exhibited in numerical simulations an example where the GFAT is non-monotonic with respect to the total mass parameter M (yet with vanishing limit), reflecting the intrinsic non-linearity of the full SBD model.

Finally, for large maximal cluster size $N \to \infty$, we found that an appropriate rescaling of the FAT (and the GFAT) was asymptotically deterministic if some corresponding first passage time was finite for the continuous-size transport equation LS model. In the opposite, we showed that for large volume and unfavorable aggregation kinetic ($q(x) > p(x)$), the mean FAT is exponentially large as M increases to infinity for the full SBD model, and the coefficient of variability close to 1 (as in the SCSBD model). We linked this behavior with phase-transition phenomenon, when the number of monomers drastically drops to 0 in a very short time, compared to the FAT. This phase-transition phenomenon occurs as a large deviation from the macroscopic deterministic model, which predicts that the number of monomers remains constant (no aggregation takes place).

This study has generalized previous studies on the first passage time on the stochastic Becker-Döring model.[6,20,23] To our knowledge, this study is the first one to capture the behavior of the FAT and its generalization for arbitrary kinetic rates, and to explore systematically its dependence with respect to the total number of monomers and the size of the maximal cluster. In particular, our study sheds lights on the variability of various first passage times that arise even in the large population limit. Also, taking into account size-dependent kinetic rate is important in practice, as monomer binding and unbinding usually depend on the available surface area of the cluster (for the spherical shape, $p_k \sim k^{2/3}$). This study may have several important applications. One of these is the explanation of the nucleation time observed in *in vitro* polymerization assay of misfolded proteins linked to neurodegenerative diseases.[11–15] Typical experiments performed in this field are able to record the nucleation time (defined as the time for which the polymerization starts) for various initial quantity of proteins. Some experiments have described a very weak dependence with respect to this initial quantity, where traditional nucleation theory could not explain this fact. Our stochastic approach points out several new behaviors that may explain the observations. Furthermore, we argue that having a model that is able to take into account the observed variability on the nucleation time will be important for parameter inference from experimental data (see also the recent preprint[45]). Indeed, even though the mean FAT may be weakly dependent on the maximal cluster size N (consider the slope of $1 + 1/(N-1)$ for large M), having the observation of the full distribution will facilitate the inference of the maximal cluster size (the shape parameter of the Weibull distribution is $k = N - 1$). Finally, on a more theoretical side, the phase-transition phenomenon of the SBD model for unfavorable aggregation and large cluster size seems to be described here for the first time. This gives a possible different definition of the nucleation rate, as an inherent infrequent stochastic process, in contrast to classical nucleation theory. It remains in the future to link this work with studies on gelation phenomenon, that is, when a fraction of the mass is concentrated in a giant particle (N is of order of M). Such studies have been performed mostly, in general, Smoluchowski coagulation models.[37,46,47]

A number of generalization of this model could be considered and will be relevant to tackle new biophysical problems. One could generalize this study to allow general coagulation-fragmentation between any two clusters.[48] This extension as well as the treatment of heterogeneous nucleation and secondary pathways will be considered in a future work.

ACKNOWLEDGMENTS

This work has been supported by ANR grant MADCOW No. ANR-08-JCJC-0135-01 (France) and Association France-Alzheimer, SM 2014. E.H. was supported by CAPES/IMPA. We thank the reviewers for valuable comments.

[1]R. Becker and W. Döring, "Kinetische behandlung der keimbildung in übersättigten dämpfen," Ann. Phys. **24**, 719-752 (1935).
[2]J. Kuipers, "Theory and simulation of nucleation," Ph.D. dissertation, Utrecht University Repository, 2009.
[3]G. M. Whitesides and M. Boncheva, "Beyond molecules: Self-assembly of mesoscopic and macroscopic components," Proc. Natl. Acad. Sci. U. S. A. **99**, 4769-4774 (2002).
[4]G. M. Whitesides and B. Grzybowski, "Self-assembly at all scales," Science **295**, 2418-2421 (2002).
[5]R. Groß and M. Dorigo, "Self-assembly at the macroscopic scale," Proc. IEEE **96**, 1490-1508 (2008).
[6]R. Yvinec, M. R. D'Orsogna, and T. Chou, "First passage times in homogeneous nucleation and self-assembly," J. Chem. Phys. **137**, 24 (2012).
[7]M. Gibbons, T. Chou, and M. R. D'Orsogna, "Diffusion-dependent mechanisms of receptor engagement and viral entry," J. Phys. Chem. B **114**, 15403-15412 (2010).
[8]N. Hoze and D. Holcman, "Kinetics of aggregation with a finite number of particles and application to viral capsid assembly," J. Math. Biol. **70**(7), 1685-1705 (2015).
[9]C. Soto, "Unfolding the role of protein misfolding in neurodegenerative diseases," Nat. Rev. Neurosci. **4**, 49-60 (2003).
[10]J. Masela, V. A. A. Jansena, and M. A. Nowak, "Quantifying the kinetic parameters of prion replication," Biophys. Chem. **77**, 139-152 (1999).
[11]E. T. Powers and D. L. Powers, "The kinetics of nucleated polymerizations at high concentrations: Amyloid fibril formation near and above the supercritical concentration," Biophys. J. **91**, 122-132 (2006).
[12]R. Yvinec, "Probabilistic modelisation in molecular and cellular biology," Ph.D. thesis, Université Lyon 1, tel-00749633, 2012.
[13]E. Hingant, "Contributions á la modélisation mathématique et numérique de problémes issus de la biologie—Applications aux Prions et ála maladie d'Alzheimer," Ph.D. thesis, Université Lyon 1, tel-00763444, 2012.
[14]W.-F. Xue, S. W. Homans, and S. E. Radford, "Systematic analysis of nucleation-dependent polymerization reveals new insights into the mechanism of amyloid self-assembly," Proc. Natl. Acad. Sci. U. S. A. **105**(26), 8926-8931 (2008).
[15]T. P. J. Knowles *et al.*, "An analytical solution to the kinetics of breakable filament assembly," Science **326**(5959), 1533-1537 (2009).
[16]O. Penrose, "The Becker-Döring equations at large times and their connection with the LSW theory of coarsening," J. Stat. Phys. **89**, 305-320 (1997).
[17]J. A. D. Wattis and J. R. King, "Asymptotic solutions of the Becker-Döring equations," J. Phys. A: Math. Gen. **31**, 7169-7189 (1998).
[18]P. Smereka, "Long time behavior of a modified Becker-Döring system," J. Stat. Phys. **132**, 519-533 (2008).
[19]T. Chou and M. R. D'Orsogna, "Coarsening and accelerated equilibration in mass-conserving heterogeneous nucleation," Phys. Rev. E **84**, 011608 (2011).
[20]F. Schweitzer, L. Schimansky-Geier, W. Ebeling, and H. Ulbricht, "A stochastic approach to nucleation in finite systems: Theory and computer simulations," Physica A **150**, 261-279 (1988).
[21]F. P. Kelly, *Reversibility and Stochastic Networks* (Cambridge Mathematical Library, 1979).
[22]A. H. Marcus, "Stochastic coalescence," Technometrics **10**, 133-143 (1968).
[23]J. S. Bhatt and I. J. Ford, "Kinetics of heterogeneous nucleation for low mean cluster populations," J. Chem. Phys. **118**, 3166-3176 (2003).

[24]A. A. Lushnikov, "Coagulation in finite systems," J. Colloid Interface Sci. **65**, 276-285 (1978).
[25]M. R. D'Orsogna, G. Lakatos, and T. Chou, "Stochastic self-assembly of incommensurate clusters," J. Chem. Phys. **136**, 084110 (2012).
[26]V. Calvez, N. Lenuzza, M. Doumic, J.-P. Deslys, F. Mouthon, and B. Perthame, "Prion dynamics with size dependency-strain phenomena," J. Biol. Dyn. **4**, 1751-3766 (2010).
[27]J. M. Ball, J. Carr, and O. Penrose, "The Becker–Döring cluster equations: Basic properties and asymptotic behaviour of solutions," Commun. Math. Phys. **104**, 4 (1986).
[28]P. Laurençot and S. Mischler, "From the Becker-Döring to the Lifshitz-Slyozov-Wagner equations," J. Stat. Phys. **106**(5-6), 957 (2002).
[29]S. Redner, *A Guide to First Passage Processes* (Cambridge University Press, 2001).
[30]N. Van Kampen, *Stochastic Processes in Physics and Chemistry*, 3rd ed. (North Holland, 2007).
[31]J. F. C. Kingman, "Markov population processes," J. Appl. Probab. **6**, 1-18 (1969).
[32]M. Kreer, "Classical Becker–Döring cluster equations: Rigorous results on metastability and longtime behaviour," Ann. Phys. **2**, 398-417 (1993).
[33]D. B. Duncan and R. M. Dunwell, "Metastability in the classical, truncated Becker–Döring equations," Proc. Edinburgh Math. Soc. **45**, 701-716 (2002).
[34]O. Penrose, "Nucleation and droplet growth as a stochastic process," in *Analysis and Stochastics of Growth Processes and Interface Models* (Oxford University Press, 2008).
[35]D. T. Gillespie, "Transition time statistics in simple bi-stable chemical systems," Physica A **101**, 2 (1980).
[36]D. F. Anderson and T. G. Kurtz, *Models of Biochemical Reaction Systems in Stochastic Analysis of Biochemical Systems* (Springer International Publishing, 2015).
[37]I. Jeon, "Existence of gelling solutions for coagulation-fragmentation equations," Commun. Math. Phys. **567**, 541-567 (1998).
[38]N. V. Brilliantov and P. L. Krapivsky, "Nonscaling and source-induced scaling behaviour in aggregation model of movable monomers and immovable clusters," J. Phys. A: Math. Gen. **24**, 4789 (1991).
[39]J. Deschamps, E. Hingant, and R. Yvinec, "Boundary value for a nonlinear transport equation emerging from a stochastic coagulation-fragmentation type model," e-print arXiv:1412.5025 (2015).
[40]A. Vasseur, F. Poupaud, J.-F. Collet, and T. Goudon, "The Beker–Döring system and its Lifshitz–Slyozov limit," SIAM J. Appl. Math. **62**, 5 (2002).
[41]J.-F. Collet, "Some modelling issues in the theory of fragmentation-coagulation systems," Commun. Math. Sci. **1**, 35-54 (2004).
[42]A. B. Bortz, M. H. Kalos, and J. L. Lebowitz, "A new algorithm for Monte Carlo simulation of Ising spin systems," J. Comput. Phys. **17**, 10-18 (1975).
[43]D. T. Gillespie, "Exact stochastic simulation of coupled chemical reactions," J. Phys. Chem. **81**, 2340-2361 (1977).
[44]M. A. Gibson and J. Bruck, "Efficient exact stochastic simulation of chemical systems with many species and many channels," J. Phys. Chem. A **104**(9), 1876-1889 (2000).
[45]S. Eugene, W.-F. Xue, P. Robert, and M. Doumic-Jauffret, "Insights into the variability of nucleated amyloid polymerization by a minimalistic model of stochastic protein assembly," www.hal-01205549 (2015).
[46]F. Rezakhanlou, "Gelation for Marcus–Lushnikov process," Ann. Probab. **41**(3B), 1806 (2013).
[47]N. Fournier and P. Laurençot, "Marcus-Lushnikov processes, Smoluchowski's and Flory's models," Stochastic Processes Appl. **119**, 1 (2009).
[48]M. R. D'Orsogna, Q. Lei, and T. Chou, "First assembly times and equilibration in stochastic coagulation-fragmentation," J. Chem. Phys. **143**(1), 014112 (2015).
[49]See supplementary material at http://dx.doi.org/10.1063/1.4940033 for figures.
[50]The fact that the first aggregation rate needs to be rescaled differently from the other aggregation rate comes from the special role played by the monomer in the BD and LS models. For a detailed discussion on the modeling point of view, see Refs. 40 and 41.