ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

THERMODYNAMICS OF IRREVERSIBLE AGGREGATION

Themis Matsoukas

150 Fenske Laboratory, Department of Chemical Engineering, Pennsylvania State University,
University Park, PA 16802, United States

Keywords: Population balances, giant component, ensemble theory, thermodynamics, information theory.

Abstract. *The emergence of a giant component from a population of finite-size clusters is a problem in mathematical physics that is encountered in fields as diverse as polymer gelation, percolation of networks, and the spread of epidemics. But while such processes have long invited thermodynamic analogies, a formal connection to thermodynamics that goes beyond the qualitative has not been established. Here we develop a thermodynamic theory for a generic mathematical object, a population of individuals that cluster into groups. The theory views the distribution in the scaling limit as the one that is most probable among all distributions that satisfy the physical constraints of the problem. In this context, the emergence of the giant cluster is a phase transition that is governed by equilibrium criteria analogous to those in molecular systems.*

1 INTRODUCTION

Binary clustering is a process by which two elements (molecules, particles, aggregates or other) combine to form a new one. It is one of the most basic and ubiquitous growth mechanisms in nature, one that describes a wide range physical phenomena over a vast scale that encompasses molecular systems, social networks, and stars [1, 2]. It is industrially relevant in granulation, a unit operation designed to increase the size of granules and homogenize their composition through the aggregation of primary particles of the pure components. Roughly 80% of pharmaceutical formulations [3] and 60% of all industrial products [4] are processed at some point as solid materials. Despite such widespread application, the design, modelling and control of particulate processes has yet to achieve the same level of accuracy and predictability as other common industrial unit operations such as separations, for example. This unmet challenge in particle technology calls for new approaches to this problem. The main theoretical tool in the study of population balances is Smoluchowski's theory of aggregation, which formulates the governing equation for the size distribution. In the past 100 years since Smoluchowski's work, another view is emerging, one that treats populations from a probabilistic standpoint. This approach, which originated in the polymer literature [5, 2], has been migrating, albeit slowly, to particulate systems [6, 7]. We recently developed a new approach to the study of generic populations based on the notion of the cluster ensemble [8]. The theory considers the ensemble of all distributions that can be formed with a fixed number of primary particles placed into a fixed number of clusters and implements a selection rule to sample distributions from the ensemble. The macroscopic observable in the thermodynamic limit –the information accessible to the observer– is the most probable distribution in this ensemble. By proper construction of the selection rule, any conceivable distribution can be obtained as the most probable distribution of the ensemble. We derived the statistics of the ensemble and showed it is described by a mathematical calculus analogous that of statistical thermodynamics. Thus we have a general theory of populations formulated in the language of thermodynamics that is in principle applicable to generic populations. Here we propose to apply this theory to the study of population balances.

Why is this direction worth pursuing? First, the ensemble theory encompasses the Smoluchowski approach, but goes beyond it. For example, it is well known that the Smoluchowski equation fails with gelling kernels, a problem that has received great attention in the literature of mathematical physics [1, 2, 9, 10, 11]. Gelation in the cluster ensemble emerges naturally and coexistence conditions of the sol and the gel are obtained in direct analogy to those in phase equilibrium of molecular systems [12]. Second, The theory generates a new set of mathematical relationships between the primary inputs and outputs of the problem (aggregation rate, size distribution). These relationships are unavailable to the Smoluchowski theory and provide new ways to study and analyze population balance problems. And third, the theory makes available the language and toolbox of thermodynamics to the study of population balances. Many problems in this area could benefit from a thermodynamic insight but one that is particularly interesting and of practical significance is aggregative mixing. In granulation, we begin with a segregated state that consists of primary particles of the pure components, and end up with a population that is hopefully well-mixed at the granule level. How long does it take for components to reach an acceptable level of mixing? What if similar or dissimilar components tend to aggregate preferentially? More importantly, if it is found experimentally that granules remain poorly mixed, what does this tell us about the magnitude of unfavorable interactions between the components? The analogy to solution thermodynamics cannot be missed. Thermodynamics teaches us how quantify such interactions in solution using activity coefficients. Is there some

analogous metric that would allow us to interpret deviations from random mixing in granules in a similar manner? These are questions that we may now begin to address, not merely by qualitative analogy to thermodynamics, but by rigorous application of the cluster-ensemble theory [8, 12, 13].

2 THE CLUSTER ENSEMBLE

Suppose we have M identical particles ("primary particles" or monomers) which we distribute into N clusters, such that each cluster contains at least one particle. The cluster ensemble consists of all distributions that can be formed with fixed M and N. All distributions of the ensemble satisfy the conditions,

$$\sum_{i=1}^{\infty} n_i = N, \quad \sum_{i=1}^{\infty} i n_i = M, \tag{1}$$

where n_i is the number of clusters that contain i particles. We envision a stochastic process the pick distributions out of that ensemble. Given a distribution $\mathbf{n} = (n_1, n_, \cdots)$, we write its probability in the canonical form,

$$P(\mathbf{n}) = \mathbf{n}! \frac{W(\mathbf{n})}{\Omega_{M,N}}. \tag{2}$$

Here, $\mathbf{n}!$ is the multinomial coefficient of the list $(n_1, n_2, \cdots)$, $W(\mathbf{n})$ is the selection bias, and $\Omega_{M,N}$ is the partition function. These quantities are discussed in more detail below.

The multinomial coefficient is

$$\mathbf{n!} = \frac{N!}{n_1! n_2! \cdots},$$

and represents the natural multiplicity of the distribution in the ensemble. If we imagine that we sample a distribution one cluster at a time, $\mathbf{n}!$ is the number of different ordered sequences in which the clusters may appear. Taking the logarithm of the multinomial coefficient and applying the Stirling formula $x! \approx x \log x - x$, we obtain

$$\log \mathbf{n}! = -\sum_i n_i \log \frac{n_i}{N} = N S(\mathbf{p}), \tag{3}$$

where $\mathbf{p} = (p_1, p_2, \cdots) = (n_1/N, n_2/N, \cdots)$ is the probability of cluster size i in the distribution and $S[\mathbf{p}]$ is the Shannon entropy of the discrete probability distribution $\mathbf{p}$,

$$S = -\sum_i p_i \log p_i. \tag{4}$$

In other words, the log of the natural multiplicity is the Shannon entropy of the distribution.

The selection bias $W(\mathbf{n}) = W(n_1, n_2, \cdots)$ is a functional of the distribution that biases its selection. With $W(\mathbf{n}) = 1$ we have the special case of the unbiased ensemble, namely an ensemble in which the probability of distribution depends only on its natural multiplicity. The selection bias is the quantity of the ensemble that embodies the physics of the problem that governs the evolution of the population. The bias will remain for the moment unspecified and the only requirement we impose is that its logarithm must be homogeneous in n_i with degree 1. Accordingly, $\log W$ satisfies the condition [8]

$$\log W(\mathbf{n}) = \sum n_i \left(\frac{\partial \log W(\mathbf{n})}{\partial n_i} \right)_{n_j} \equiv \sum n_i \log w_i, \tag{5}$$

which follows from Euler's theorem for homogeneous functions. Here, $\log w_i$ is the partial derivative of the log bias with respect to n_i. It represents the contributions cluster size i to the log bias and we refer to it as the cluster bias.

The partition function is the normalization factor of probabilities and satisfies the normalizing condition,

$$\Omega_{M,N} = \sum_{\mathbf{n}} \mathbf{n}! W(\mathbf{n}). \tag{6}$$

The summation goes over all distributions that can be formed by M monomers assembled into N clusters. Accordingly, the partition function depends on M and N. It also depends on the selection bias, but to the extent that the bias is determined by the physics that govern the population, for a given process Ω is a function of M and N only.

2.1 Most probable distribution

The premise of the ensemble theory is that for large M, N, ensemble reduces to its most probable distribution (mpd). This distribution is obtained by maximizing the probability $P(\mathbf{n})$, or equivalently its logarithm, under the constraints in Eq. (1). We perform this constrained maximization by the method of Lagrange multipliers. The objective function to be maximized is

$$\mathcal{F} = \log \mathbf{n}! + \log W(\mathbf{n}) - \alpha \left(\sum n_i - N \right) - \beta \left(\sum i n_i - M \right). \tag{7}$$

Using Eq. (3) for $\log \mathbf{n}!$ and setting the derivative of $\mathcal{F}$ with respect to n_i equal to zero we find

$$\log \frac{\tilde{n}_i}{N} = -(\alpha + 1) - \beta i + \left(\frac{\partial W(\tilde{\mathbf{n}})}{\partial \tilde{n}_i} \right) \tag{8}$$

where $\tilde{n}_i$ refers to the most probable distribution. Solving for $\tilde{n}_i$, the most probable distribution takes the form [8]

$$\frac{\tilde{n}_i}{N} = \tilde{w}_i e^{-\beta i}/q, \tag{9}$$

where $q = \exp(\alpha + 1)$ and $\log \tilde{w}_i$ are the partial derivatives of $\log \tilde{W} = \log W(\tilde{\mathbf{n}})$ evaluated at the most probable distribution.

2.2 The partition function in the thermodynamic limit

According to the maximum term method, as the ensemble converges to the most probable distribution, the log of the partition function converges to the log of the maximum term:

$$\log \Omega_{M,N} = \sum_{\mathbf{n}} \log \mathbf{n}! W(\mathbf{n}) \to \log \tilde{\mathbf{n}}! W(\tilde{\mathbf{n}}) = -\sum n_i \log \frac{n_i}{N} + \log \tilde{W}. \tag{10}$$

We use Eq. (9) to calculate the right-hand side of the above equation:

$$\log \Omega_{M,N} = \beta \sum i n_i + \log \sum n_i + \sum n_i \log \tilde{w}_i, \tag{11}$$

and this, by virtue of Eq. (1) and the homogeneity of the log bias in Eq. (5), is written in the more compact form [8],

$$\log \Omega_{M,N} = \beta M + (\log q) N. \tag{12}$$

In the thermodynamic limit $\tilde{n}_i/N$ is an intensive property. It follows from Eq. (9) that β, $\log q$ and $\tilde{w}_i$ are intensive, and then from Eq. (12) that $\log \Omega_{M,N}$ is extensive, i.e., it is homogeneous in M and N with degree 1. By Euler's theorem $\log \Omega$ is of the form

$$\log \Omega_{M,N} = M \left(\frac{\partial \log \Omega}{\partial M} \right)_N + N \left(\frac{\partial \log \Omega}{\partial N} \right)_M . \tag{13}$$

Direct comparison of this result with Eq. (12) leads to the identifications

$$\beta = \left(\frac{\partial \log \Omega}{\partial M} \right)_N , \quad \log q = \left(\frac{\partial \log \Omega}{\partial N} \right)_M . \tag{14}$$

Equations (9), (12) and (14) summarize the results of the cluster ensemble: the most probable distribution depends on three intensive variables: β and q, which are given by the partial derivatives of the partition function, and $\tilde{w}_i$, which are the partial derivatives of the selection bias.

The differential of $\log \Omega$ with respect to M and N is

$$d \log \Omega = \beta dM + (\log q) dN. \tag{15}$$

The result bears a direct analogy to the thermodynamic differential,

$$d \log \Omega = \beta dE + \beta P dV - \beta \mu dn, \tag{16}$$

where Ω is the microcanonical partition function, E is energy, V is volume and n is the number of particles. M and N in the cluster ensemble may be taken to be analogous to E and n (other analogies can be drawn as well [8]). Accordingly, β in the cluster ensemble is analogous to inverse temperature and $\log q$ to chemical potential. There is one difference: in the thermodynamic ensemble the log of the partition function is entropy; in the cluster ensemble the corresponding relationship is

$$\log \Omega_{M,N} = \tilde{S} + \log \tilde{W}.$$

In the special case $W = 1$ (unbiased ensemble), we obtain $\log \Omega_{M,N} = \tilde{S}$, as in molecular systems, and the most probable distribution is the familiar exponential distribution of the canonical ensemble. In fact, $W = 1$ represents the mathematical statement of the postulate of equal a priori probabilities, which forms the basis for the derivation of the canonical distribution in statistical mechanics. The cluster ensemble is a generalization to systems that are not limited to unbiased selection. Whereas the most probable distribution of of thermodynamic microstates is exponential, the most probable distribution in the cluster ensemble can be *any* distribution. Conversely, any distribution may be viewed as the most probable distribution of the cluster ensemble under appropriate selection bias.

3 LINEAR ENSEMBLES

The cluster bias w_i, defined in Eq. (5), is a functional of the distribution, i.e., it depends on all n_i. An important special case is the selection bias is of the form,

$$W(\mathbf{n}) = a_1^{n_1} a_2^{n_2} \cdots \tag{17}$$

where a_i functions of i. Its logarithm is

$$\log W(\mathbf{n}) = a_1 n_1 + a_2 n_2 + \cdots , \tag{18}$$

from which we obtain the cluster bias,

$$w_i = \left(\frac{\partial \log W}{\partial n_i} \right)_{n_j} = a_i. \tag{19}$$

In this special type the cluster bias w_i is an intrinsic function of i, i.e., it is independent of the distribution. We call this type of selection functional linear because its logarithm is a linear combination with fixed coefficients. The corresponding ensemble, which we call linear, has several interesting properties. Combining this selection bias with Eq. (6) we obtain

$$\Omega_{M,N} = N! \sum_{n_i} \frac{a_1^{n_1}}{n_1!} \frac{a_2^{n_2}}{n_2!} \cdots , \tag{20}$$

with the summation going over all clusters of the ensemble. The derivative with respect to a_k gives

$$\frac{\partial \Omega_{M,N}}{\partial a_k} = N \left((N-1) \sum_{n_i} \cdots \frac{a_k^{n_k-1}}{(n_k-1)!} \cdots \right). \tag{21}$$

Notice that the differentiation amounts to removing a cluster of mass k from the summation, therefore producing the partition function of the ensemble with mass $M-k$ and number of particles $N-1$:

$$\frac{\partial \Omega_{M,N}}{\partial n_i} = N \Omega_{M-i,N-1}. \tag{22}$$

An equivalent form to write Eq. (21) is

$$\frac{\partial \Omega_{M,N}}{\partial a_k} = \frac{\Omega_{M,N}}{a_k} \left(\frac{1}{\Omega_{M,N}} \sum_{n_i} n_k \cdots \frac{a_k^{n_k-1}}{n_k!} \cdots \right) = \frac{\Omega_{M,N}}{a_k} \langle n_k \rangle , \tag{23}$$

whose result on the right is obtained by recognizing the quantity in brackets as the ensemble average of the number of clusters with k particles. The ensemble average $\langle n_k \rangle$ then is

$$\langle n_k \rangle = N a_k \frac{\Omega_{M-k,N-1}}{\Omega_{M,N}}. \tag{24}$$

Therefore, the mean cluster distribution of the linear ensemble can be calculated directly from the partition function and the cluster bias a_k.

4 DISCRETE BINARY AGGREGATION

The theory of the cluster ensemble discussed above is very general and applies to any population. We will now demonstrate the theory by applying it to a physical process, binary aggregation. In binary aggregation two clusters with mass i and j, respectively, combine to form a new cluster with mass $i+j$. The process represented by the reaction

$$(i) + (j) \xrightarrow{k_{ij}} (1+j)$$

whose rate is characterized by the aggregation kernel, k_{ij}. Given a cluster distribution $\mathbf{n}$, the probability that cluster masses i and j combine into a cluster is proportional to the aggregation

kernel and the number of (i,j) pairs. In discrete finite systems the number of pairs is $n_i n_j$ if $i \neq i$, and $n_i(n_i-1)/2$ if $i=j$. Thus the general case can be expressed in the form,

$$P_{(i)+(j)\to(i+j)} = C_{\mathbf{n}} \frac{n_i(n_j-\delta_{i,j})}{1+\delta_{i,j}}\, k_{i,j}, \tag{25}$$

with

$$C_{\mathbf{n}} = \frac{2}{N(N-1)\bar{k}(\mathbf{n})}, \tag{26}$$

with the constant C_n is such that the sum of the transition probabilities are normalized. The factor $\bar{k}(\mathbf{n})$ on the right-hand side of (26) is the mean kernel within distribution $\mathbf{n}$ and is calculated by averaging over all possible aggregation events of the $N(N-1)/2$ pairs of clusters in the distribution:

$$\bar{k}(\mathbf{n}) = \frac{2}{N(N-1)} \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} \frac{n_i(n_j-\delta_{i,j})}{1+\delta_{i,j}}\, k_{i,j}. \tag{27}$$

The aggregation event between two cluster masses may be viewed as a reaction that converts a parent distribution into a new distribution that has the same total mass but one less particle. This parent-offspring relationship establishes a network of connections that link the distributions of the (M,N) ensemble to those of the $(M,N-1)$ ensemble. This network is shown in Fig. 1 for $M=7$. At the top of the network we have $M=7$, $N=7$, corresponding to a population of monomers that represents generation (fully dispersed state). In the next step, two monomers combine to form a dimer. In the third step a monomer may add to the dimer to produce a trimer, or two monomers combine to form a new dimer. From there on the possibilities increase, until at the end we reach the fully gelled system in which all monomers are part of the same cluster $(M=7,N=1)$. The transition between parents and offsprings can be conveniently tracked by the generation g, which we define as $g=M-N+1$. The completely dispersed state corresponds to generation $g=1$ and the completely gelled state to $g=N$. The distributions in generation g contain the complete ensemble of distributions that can be formed by M monomers into N clusters. For example, with $M=7$, $N=2$, the possible distributions are monomer + hexamer ($n_1=1$, $n_6=1$), dimer + pentamer ($n_2=1$, $n_5=1$) and trimer+tetramer ($n_3=1$, $n_4=1$). All are present in generation $g=6$. The aggregation network may them be viewed as a process that acts on the distributions of the (M,N) ensemble to produce the distributions of the $(M,N-1)$ ensemble. The problem now becomes to find the equations that govern this evolution of the ensemble.

4.1 Parent-offspring relationship

There is a systematic way to construct all parents of $\mathbf{n}$. A cluster of mass $i \geq 2$ is formed by the aggregation of masses $i-j$ and j. Accordingly, if we break a cluster i of distribution $\mathbf{n}$ into fragments $i-j$ and j, the resulting distribution is a parent of $\mathbf{n}$, specifically, the $(i-j,j)$-parent. Using unprimed variables for the offspring and primed for the parent, the equations that define the parent of distribution $\mathbf{n}$ are

$$\begin{aligned} n'_i &= n_i - 1; \\ n'_{i-j} &= n_{i-j} + 1 + \delta_{i-j,j}; \\ n'_j &= n_j + 1 + \delta_{i-j,j}; \end{aligned} \tag{28}$$

Clearly, only clusters with $i \geq 2$ produce a parent (the monomer cannot be produced by aggregation of smaller particles), The complete set of the parents of $\mathbf{n}$ is obtained by letting i, j, span

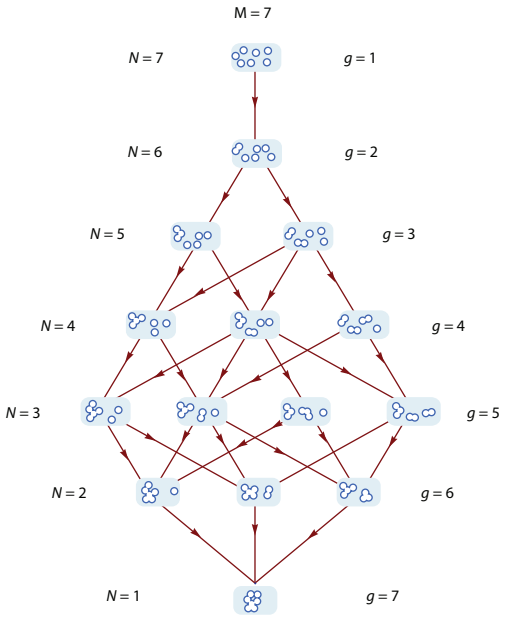

Figure 1: The network for discrete aggregation of $M = 7$ identical particles.

the range

$$i = 2, \cdots \infty; \quad j = 1, \cdots i/2. \tag{29}$$

where $i/2$ is understood as integer division. For example, the distribution that contains one monomer, one dimer and one tetramer has three parents: one corresponds to breaking the dimer into monomers (the result is three monomers and a trimer); the second corresponds to breaking the tetramer into a monomer and a trimer (the result is two monomers, a dimer and a trimer); and the third corresponds to breaking the tetramer into two dimers (the result is a monomer and three dimers).

4.2 The Master Equation

The probability of distribution propagates from parent to offspring via the Master Equation,

$$P(\mathbf{n}) = \sum_{\mathbf{n}'} P(\mathbf{n}') P_{\mathbf{n}' \to \mathbf{n}}. \tag{30}$$

Here, $P(\mathbf{n})$ is the probability of the offspring, $P(\mathbf{n}')$ the probability of the parent, and $P_{\mathbf{n}' \to \mathbf{n}}$ is the transition probability from parent to offspring. The transition probability from parent $\mathbf{n}'$ to offspring $\mathbf{n}$ is given by Eq. (31) with the understanding that the right-hand side must be expressed in terms of the parent. This amounts to replacing $\mathbf{n}$ by $\mathbf{n}'$ and N by $N + 1$. We now take $\mathbf{n}'$ to be the $(i - j, j)$-parent, namely, the parent that produces a new cluster of size i via the aggregation of cluster masses $(i - j)$ and (j). The transition probability then becomes

$$P_{\mathbf{n}' \to \mathbf{n}} = \frac{2}{N(N+1)} \frac{n_i'(n_j' - \delta_{i,j})}{1 + \delta_{i,j}} \frac{k_{i,j}}{\bar{k}(\mathbf{n}')}. \tag{31}$$

Finally, we express the probabilities $P(\mathbf{n})$, $P(\mathbf{n}')$ in the canonical form of Eq. (2),

$$P(\mathbf{n}) = \mathbf{n}! \frac{W(\mathbf{n})}{\Omega_{M,N}}; \quad P(\mathbf{n}') = \mathbf{n}'! \frac{W(\mathbf{n}')}{\Omega_{M,N+1}}, \tag{32}$$

and substitute these expressions in to the Master Equation. The resulting equation produces two separate recursions, one for the partition function and another for the selection bias. The recursion for the partition function is

$$\frac{\Omega_{M,N}}{\Omega_{M,N+1}} = \frac{M-N}{M}\frac{1}{\langle k_{M,N+1}\rangle}, \tag{33}$$

and is easily solved to produce the closed-form result

$$\Omega_{M,N} = \binom{M-1}{N-1}\left(\prod_{L=N+1}^{M}\langle k_{M,L}\rangle\right). \tag{34}$$

The recursion for the selection bias does not have a similar closed-form expression and is given by

$$W(\mathbf{n}) = \sum_{i=2}^{\infty}\frac{n_i(i-1)}{M-N}\left\{\frac{1}{i-1}\sum_{j=1}^{i-1}\frac{k_{i-j,j}\,\langle k_{M,N+1}\rangle}{\bar{k}(\mathbf{n}')}W(\mathbf{n}')\right\}. \tag{35}$$

The ensemble average kernel that appears in these equation is the average of $\bar{k}(\mathbf{n})$ overall distributions,

$$\langle k_{M,N}\rangle = \sum_{\mathbf{n}\in(M,N)} P(\mathbf{n})\,\bar{k}(\mathbf{n}) \tag{36}$$

with the summation running over all distributions of the (M,N) ensemble. $\langle k_{M,N}\rangle$ and $\bar{k}(\mathbf{n})$ are generally different: $\bar{k}(\mathbf{n})$ is the mean kernel within distribution $\mathbf{n}$ and $\langle k\rangle$ is the average of these mean kernels over all distributions of the ensemble. Equation (35) is a recursion that gives the bias of distribution $\mathbf{n}$ in terms of the bias $W(\mathbf{n}')$ of its parents. The double summation goes over all parents of $\mathbf{n}$: the inner summation goes over all possible ways to form a cluster of size i through the aggregation of sizes $(i-j)$ and (j), and the outer summation goes over all $i>1$ (the monomer cannot be formed by aggregation).

Having obtained the partition function, we derive expressions for the parameters β and $\log q$ in the most probable distribution from Eq. (14),

$$\beta = \frac{\Omega_{M+1,L}}{\Omega_{M,L}} = \frac{M}{M-N+1}\prod_{L=N+1}^{M}\frac{\langle k_{M+1,N+l}\rangle}{\langle k_{M,N+l}\rangle} \tag{37}$$

$$\log q = \frac{\Omega_{M,N+1}}{\Omega_{M,N}} = \frac{M-N}{N}\frac{1}{\langle k_{M,N+1}\rangle}, \tag{38}$$

which follow by applying the discrete equivalent of the derivatives $\partial\Omega/\partial M$ and $\partial\Omega/\partial N$. Equations (34)–(38) are a closed set of equations for β, $\log q$ and W, which together allow us to calculate the most probable distribution. To obtain the final distributions closed form we must first solve the recursion for W. Certain special cases where this can be done analytically are discussed below.

5 EXACT SOLUTIONS

In general, the mean kernel in distribution, $\bar{k}(\mathbf{n})$, and the ensemble average kernel, $\langle k_{M,N}\rangle$, are not the same. A special class of kernels is when these two averages are the same:

$$\langle k_{M,N}\rangle = \bar{k}(\mathbf{n}); \quad \text{for all } \mathbf{n} \text{ in the } (M,N) \text{ ensemble.} \tag{39}$$

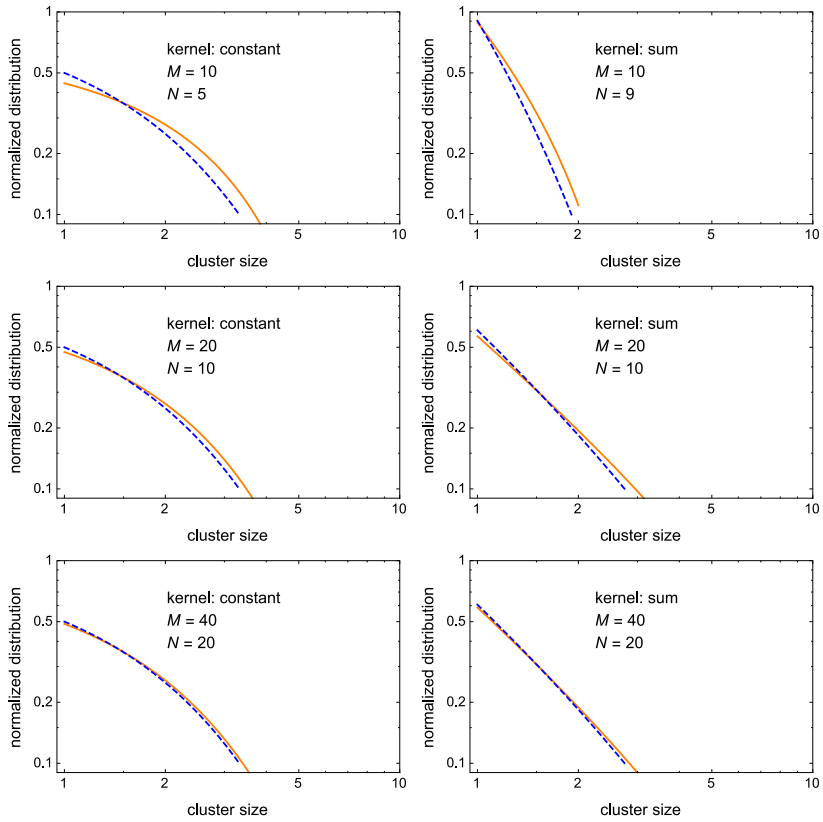

Figure 2: Distributions for the constant kernel (left column) and the sum kernel (right column) for various values of N, N, at $M/N = 2$. The solid line is the discrete solution (Eqs. 45 and 55) and the dashed line is the most probable distribution in the thermodynamic limit (Eqs. 48, 58). The two distributions converge as M and N increase.

If we insert this condition into Eq. (35) we obtain the following simplified recursion,

$$W(\mathbf{n}) = \sum_{i=2}^{\infty} \frac{n_i(i-1)}{M-N} \left\{ \frac{1}{i-1} \sum_{j=1}^{i-1} k_{ij} W(\mathbf{n}') \right\}, \tag{40}$$

whose solution is a linear bias with coefficients

$$a_i = \frac{1}{i-1} \sum_{j=1}^{i} a_{i-j} a_j k_{i-j,j}; \quad a_1 = 1. \tag{41}$$

The only kernels that satisfy Eq. (39) and produce linear ensembles are the constant kernel ($k_{ij} = 1$), the sum kernel ($k_{ij} = (i+j)/2$) and their linear combinations. Below we derive the results for the constant and sum kernels.

5.1 Constant kernel

The constant kernel, $k_{ij} = 1$, satisfies the condition in Eq. (39) trivially,

$$\langle k_{ij} \rangle_{M,N} = \bar{k}(\mathbf{n}) = k_{ij} = 1; \quad \text{for all } M, N, i, j, \mathbf{n} \tag{42}$$

and with this result Eq. (41) gives

$$a_i = 1; \quad \text{for all } i. \tag{43}$$

The corresponding cluster bias is $W(\mathbf{n}) = 1$ for all $\mathbf{n}$, which means that the constant kernel produces an unbiased ensemble, one in which all distributions are equally probable. Various properties of the ensemble can now be calculated. First, the partition function, which follows from Eq. (34):

$$\Omega_{M,N} = \binom{M-1}{N-1}. \tag{44}$$

Next, the mean distribution from Eq. (24):

$$\frac{\langle n_k \rangle}{N} = \binom{M-k-1}{N-2} \Big/ \binom{M-1}{N-1}. \tag{45}$$

This result is exact for any finite M and N. For the most probable distribution we return to Eq. (9). Before we proceed we recall that the derivation treats the most probable distribution as a continuous function, which implicitly assumes large M and N. For this reason we will derive the results in the limit $M, N \to \infty$ at fixed $M/N = \bar{x}$ (thermodynamic limit). Using Eq. (14), the parameters β and q are

$$\beta = \frac{\Omega_{M+1,N}}{\Omega_{M,N}} = \log \frac{M}{M-N} \to \log \frac{\bar{x}}{\bar{x}-1}, \tag{46}$$

$$\log q = \frac{\Omega_{M,N+1}}{\Omega_{M,N}} = \log \frac{M-N}{N} \to \log(\bar{x}-1). \tag{47}$$

The most probable distribution is obtained by substituting these results in Eq. (9),

$$\frac{\tilde{n}_k}{N} = \frac{(1-1/\bar{x})^k}{\bar{x}-1} \to \frac{e^{-k/\bar{x}}}{\bar{x}}. \tag{48}$$

The limiting form on the right-hand side is obtained for $\bar{x} \gg 1$. In the unbiased ensemble the most probable distribution is the distribution that maximizes entropy, and as is well known, of all distribution with the same mean $\bar{x}$, the distribution that maximizes entropy is exponential. This is also the well-known solution of the Smoluchowski equation for the constant kernels, therefore we have also made contact with the classical literature on population balances.

5.2 Sum kernel

The sum kernel is proportional to the sum of the cluster masses,

$$k_{ij} = \frac{i+j}{2}. \tag{49}$$

In this case the mean aggregation kernel in any distribution of the (M, N) ensemble is

$$\langle k_{12} \rangle_{M,N} = \frac{M}{N}, \tag{50}$$

from which it follows that $\langle k \rangle_{M,N} = M/N$. The product of kernels that appears in Eq. (34) is

$$\prod_{l=N+1}^{M} \langle k_{12} \rangle_{M,l} = \frac{M}{M} \cdot \frac{M}{M-1} \cdots \frac{M}{N+1} = \frac{N!}{M!} M^{M-N}, \tag{51}$$

and the partition function becomes

$$\Omega_{M,N} = \frac{M^{M-N}}{M!}\binom{M-1}{N-1} \tag{52}$$

The cluster bias requires inversion of the recursion

$$a_i = \frac{i}{2(i-1)}\sum_{j=1}^{i-1} a_{i-j}a_j, \quad a_1 = 1. \tag{53}$$

The same recursion was obtained by Spouge [14] in a combinatorial study of aggregation, and also in [1] in a treatment of the same problem based on Smoluchowski equation; its solution is [1]

$$a_i = \frac{i^{i-1}}{i!}. \tag{54}$$

By application of Eq. (24), the mean distribution is

$$\frac{\langle n_k\rangle_{M,N}}{N} = \frac{(M-k)^{M-N-k}}{(M-N-k+1)!}\cdot\frac{(N-1)(M-N)!}{M^{M-N-1}} \tag{55}$$

The parameters β and $\log q$ are

$$\beta = \frac{\Omega_{M+1,N}}{\Omega_{M,N}} \to \frac{x-1}{x} - \log\left(\frac{x-1}{x}\right) \tag{56}$$

$$\log q = \frac{\Omega_{M,N+1}}{\Omega_{M,N}} \to \log\left(\frac{\bar{x}-1}{\bar{x}}\right) \tag{57}$$

and the most probable distribution is

$$\frac{\tilde{n}_k}{N} = \frac{k^{k-1}}{k!}\left(\frac{x-1}{x}\right)^{k-1} e^{-k(x-1)/x}. \tag{58}$$

In contrast to Eq. (55), which is valid for any M, N, Eq. (58) is appropriate in the thermodynamic limit. The classical solution for the number concentration c_i as a function of time, t, is [1]

$$\frac{n_k}{N} = \frac{k^{k-1}}{k!}\left(1-e^{-t}\right)^{k-1} e^{-k(1-e^{-t})}. \tag{59}$$

where t is time. With the substitution $e^t = \bar{x}$, this reverts to (58).

Figure 2 shows graphs of the mean and the most probable distribution at fixed $/N = 2$ for $M = 10, 20, 40$. The mean distribution is exact for all M and N. The most probable distribution approaches the mean as M increases. The convergence between the mean and the most probable distribution demonstrates the basic premise of the cluster ensemble: in the thermodynamic limit the ensemble converges into a single distribution. In this limit the mean and the most probable distributions are identical.

6 THE GIANT COMPONENT

The convergence of the mean distribution and the most probable distribution is a trait of single-phase systems. This equality is destroyed when multiple phases are present. Since each phase converges to its own mean and most probable distribution, the two-phase system does not consist of a single distribution but is a linear combination of the distributions in each phase. Is it possible for an irreversible system of aggregating clusters to exhibit phase equilibrium? Yes. The classical example is the product kernel,

$$k_{ij} = ij, \tag{60}$$

which is known to result in gelation [1, 10]. The physical manifestation of gelation comes to us from polymer science and is manifested by the formation of a giant network, a single "cluster" that contains a finite fraction of the total particles in the system. This network coexists with a phase of dispersed finite-size clusters (the sol). This process has been long viewed as a phase transition by qualitative analogy to vapor/liquid systems. We now have the tools to address this problem in rigorous thermodynamic terms.

The product kernel asymptotically gives

$$\bar{k}(\mathbf{n}) \approx \left(\frac{M}{N}\right)^2, \tag{61}$$

a result that we obtain by replacing the diagonal elements $k_{ii}n_i(n_i-1)/2$ with $k_{11}n_i^2$. Except for distributions that contain a sizeable fraction of mass in very large clusters, this replacement is inconsequential. This approximation implicitly constraints the above relationship to the sol phase only, but as we will see, this will not prevent us from applying the theory to the two-phase system. It only means that the above expression can be used only for the properties of the sol, whereas the properties of the gel must be calculated from the the total mass balance. Since $\bar{k}(\mathbf{n})$ is (asymptotically) the same in all distributions of the ensemble, the product kernel satisfies Eq. (39) and the theory of linear ensembles applies to the sol distribution.

6.1 Sol distribution

The product of kernels in Eq. (34) is easily obtained in closed form

$$\prod_{L=N+1}^{M} \langle k_{ij}\rangle_{M,L} = \left(\frac{M}{M}\right)^2 \cdots \left(\frac{M}{N+1}\right)^2 = \left(\frac{N!}{M!}M^{M-N}\right)^2, \tag{62}$$

and leads to the following expression for the partition function [12]

$$\Omega_{M,N} = \left(\frac{N!}{M!}M^{M-N}\right)^2 \binom{M-1}{N-1} \tag{63}$$

The recursion for a_i in Eq. (53) becomes

$$a_i = \frac{1}{i-1}\sum_{j=1}^{i-1}(i-j)ja_{i-j}a_j$$

and its solution is

$$a_k = 2\frac{(2i)^{i-2}}{i!}. \tag{64}$$

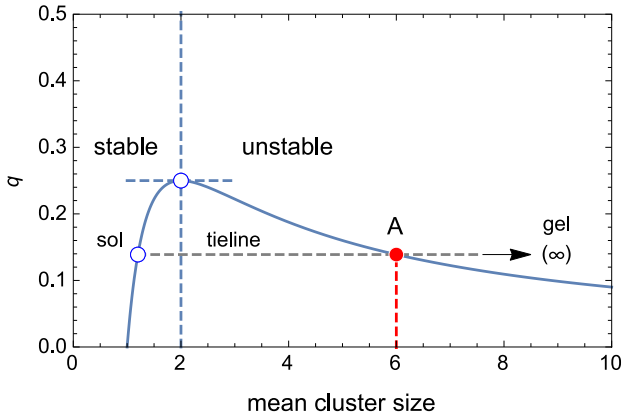

Figure 3: The parameter q for the product kernel. Stability requires $dq/d\bar{x} > 0$. Accordingly, the sol can exist only in the region $1 < \bar{x}_{\text{sol}} < 2$.

For the parameters β and q we find [12]

$$\beta = \frac{2(M-N)}{M} - \log\frac{M-N}{M} \to \frac{2(\bar{x}-1)}{\bar{x}} - \log\left(\frac{\bar{x}-1}{\bar{x}}\right)$$
$$\log q = \log\frac{N(M-N)}{M^2} \to \log\left(\frac{x-1}{x^2}\right).$$

With these results, the most probable distribution of the sol phase is

$$\frac{\tilde{n}_k}{N} = \frac{2^{k-1}k^{k-2}\bar{x}}{k!}\left(\frac{\bar{x}-1}{\bar{x}}\right)^{k-1} e^{-2k(x-1)/\bar{x}}. \tag{65}$$

6.2 Sol distribution

Let us return to the fundamental equation of the cluster ensemble, Eq. (12). The most probable distribution maximizes the partition function ($d\Omega = 0$) with respect to all distributions $\mathbf{n}$ with fixed M and N. Accordingly, Ω (as well as $\log\Omega$) is concave in M and N, which further implies that the second derivatives of $\log\Omega$ in M and N are both negative:

$$\left(\frac{\partial^2\Omega}{\partial M^2}\right) = \left(\frac{\partial\beta}{\partial M}\right)_N < 0, \quad \left(\frac{\partial^2\Omega}{\partial N^2}\right) = \left(\frac{\partial\log q}{\partial N}\right)_M < 0. \tag{66}$$

Using $\bar{x} = M/N$, the derivative with respect to N is

$$\left(\frac{\partial\log q}{\partial N}\right)_M = \left(\frac{\partial\log q}{\partial\bar{x}}\right)_M\left(-\frac{\bar{x}}{M}\right) < 0, \tag{67}$$

which requires the derivative $(\partial\log q/\partial\bar{x})$ to be positive. In aggregation M is constant, therefore, stability requires q to be a monotonically increasing function of $\bar{x}$. As we see in Fig. 3, q plotted against $\bar{x}$ has a maximum at $\bar{x} = x_* = 2$ that divides the domain of $\bar{x}$ into two regions: a stable region to the left of x_* and an unstable region to the right. If $M/N < 2$, the entire system exists as a stable single-phase sol with $\bar{x}_{\text{sol}} = M/N$. When $M/N > 0$, the single sol is unstable and the system splits into two phases, a sol with $x_{\text{sol}} < 2$ and a gel with $\bar{x}_{\text{gel}} > 2$. To

construct the tie line, we recall Eq. (33), which governs the evolution of the partition function during aggregation. This result states that the $\log q$ of the sol satisfies

$$\log q_{\text{sol}} = \left. \frac{M-N}{N} \frac{1}{\langle k_{M,N+1} \rangle} \right|_{\text{sol}}, \tag{68}$$

which establishes the condition that defines the tie line. Consider for example state A corresponding to an overall mean cluster size $\bar{x} = 6$. This state cannot exist as a single sol but splits into a stable sol (point S), and a gel cluster, whose size cannot be represented on this graph, which we recall is based on equations that apply only to the sol phase. The size of the gel cluster can be calculated if we know the M and N values of the overall system. The two phase system consists of a single gel cluster [8] with mass m_{gel} and $N-1$ sol clusters with average size $\bar{x}_{\text{sol}}$. By mass balance,

$$\bar{x}_{\text{sol}}(N-1) + m_{\text{gel}} = M. \tag{69}$$

The gel fraction $\phi_{\text{gel}} = m_{\text{gel}}/M$ is

$$\phi_{\text{gel}} = 1 - \bar{x}_{\text{sol}} \frac{N-1}{M}. \tag{70}$$

For large M and N this becomes

$$\phi_{\text{gel}} = 1 - \frac{\bar{x}_{\text{sol}}}{\bar{x}}. \tag{71}$$

Thus we have the gel fraction in terms of he overall mean cluster size $\bar{x}$ and the equilibrium sol size $\bar{x}_{\text{sol}}$, determined graphically from the tie line.

7 CONCLUSIONS

Here is what we have accomplished. We have formulated the cluster ensemble, a new theory for generic populations. The theory builds on Gibbs's notion of ensemble, with an added feature, a selection functional that establishes the selection rule among distributions of the ensemble. With the only requirement of homogeneous behavior, the ensemble is seen to obey regular thermodynamics. By "thermodynamics" we mean that it possesses the following properties:

1. It is governed by a partition function whose maximization with respect to the primitive stochastic variable of the ensemble (distribution $\mathbf{n}$) as well as with respect to its extensive independent variables (M and N) determines the equilibrium state.

2. It is equilibrated when its distribution relaxes to the most probable distribution.

3. The partition function satisfies the homogeneous condition

$$\log \Omega_{M,N} = M \left(\frac{\partial \log \Omega}{\partial M} \right)_N + N \left(\frac{\partial \log \Omega}{\partial N} \right)_M.$$

and the stability conditions

$$\left(\frac{\partial^2 \Omega}{\partial M^2} \right), \left(\frac{\partial^2 \Omega}{\partial N^2} \right) < 0.$$

This formalism describes *any* population. More generally, it describes the probability density function of any stochastic variable. Thus we recognize thermodynamics not as a physical theory, but as a probabilistic calculus whose applicability extends beyond molecular systems. The crux of the mathematical problem is the determination of the selection functional. In some cases this may be given as part of the model [8]. In more typical situations, this selection bias must be determined from the laws that govern the population. The case of binary aggregation discussed here is an example of how this would work in the general case. Start with the Master Equation of the problem, obtain the transition probabilities from the governing laws, express the probabilities of distribution in canonical form, and work out a recursive solution for the partition function and the selection bias.

All of this can be done much more easily using the Smoluchowski equation. Why venture into the complexities of the ensemble? The Smoluchowski equation is fine as long as (a) the only information we seek is the average distribution (in most cases adequate) and (b) the population consists of a single phase. When the second requirement is not met, the Smoluchowski equation breaks down. There are several examples, apart from gelation, where this is the case: Network connectivity, spread of epidemics, financial mergers and the emergence of monopolies, species extinction and dominance, economic inequality, are all systems in which a single element of the population, infinitesimal in terms of number fraction, possesses a finite fraction of the whole. For these types of problems, the cluster ensemble provides a rigorous tool, one whose success is well established in physics and chemistry, and whose applicability, we now recognize, is far more general.

REFERENCES

[1] F. Leyvraz, "Scaling theory and exactly solved models in the kinetics of irreversible aggregation," *Physics Reports*, vol. 383, no. 2-3, pp. 95–212, 2003.

[2] W. H. Stockmayer, "Theory of molecular size distribution and gel formation in branched-chain polymers," *The Journal of Chemical Physics*, vol. 11, no. 2, pp. 45–55, 1943.

[3] CSDD, "Tufts center for the study of drug development impact report," tech. rep., Tufts, 2012.

[4] S. M. Iveson, J. D. Litster, K. Hapgood, and B. J. Ennis, "Nucleation, growth and breakage phenomena in agitated wet granulation processes: a review," *Powder Technology*, vol. 117, no. 1-2, pp. 3 – 39, 2001.

[5] P. J. Flory, "Molecular size distribution in three dimensional polymers. II. trifunctional branching units," *Journal of the American Chemical Society*, vol. 63, no. 11, pp. 3091–3096, 1941.

[6] D. Ramkrishna, "Analysis of population balance—iv: The precise connection between Monte Carlo simulation and population balances," *Chemical Engineering Science*, vol. 36, no. 7, pp. 1203 – 1209, 1981.

[7] D. Ramkrishna, *Population Balances*. San Diego: Academic Press, 2000.

[8] T. Matsoukas, "Statistical thermodynamics of clustered populations," *Phys. Rev. E*, vol. 90, p. 022113, Aug 2014.

[9] E. M. Hendriks, M. H. Ernst, and R. M. Ziff, "Coagulation equations with gelation," *J. Stat. Phys.*, vol. 31, pp. 519–563, 1983.

[10] A. A. Lushnikov, "Exact kinetics of the sol-gel transition," *Phys. Rev. E*, vol. 71, p. 046129, Apr 2005.

[11] A. A. Lushnikov, "Gelation in coagulating systems," *Physica D: Nonlinear Phenomena*, vol. 222, pp. 37–53, 10 2006.

[12] T. Matsoukas, "Statistical thermodynamics of irreversible aggregation: The sol-gel transition," *Sci. Rep.*, vol. 5, p. 8855, 2015.

[13] T. Matsoukas, "Abrupt percolation in small equilibrated networks," *Phys. Rev. E*, vol. 91, p. 052105, May 2015.

[14] J. L. Spouge, "Equilibrium polymer size distributions," *Macromolecules*, vol. 16, no. 1, pp. 121–127, 1983.