Stochastic Coalescence

David J. Aldous[1]

Abstract. Consider N particles, which merge into clusters according to the rule: a cluster of size x and a cluster of size y merge at (stochastic) rate $K(x,y)/N$, where K is a specified rate kernel. This Marcus-Lushnikov model of coalescence, and the underlying deterministic approximation provided by the Smoluchowski coagulation equations, have an extensive scientific literature. A recent reformulation is the *general stochastic coalescent*, whose state space is the infinite-dimensional simplex (the state $\mathbf{x} = (x_i, i \geq 1)$ represents unit mass split into clusters of masses x_i), and which evolves by clusters of masses x_i and x_j coalescing at rate $K(x_i, x_j)$. Existing mathematical literature (Kingman's coalescent, component sizes in random graphs, fragmentation of random trees) implicitly studies certain special cases. Recent work has uncovered deeper constructions of special cases of the stochastic coalescent in terms of Brownian-type processes. Rigorous study of general kernels has only just begun, and many challenging open problems remain.

1991 Mathematics Subject Classification: 60J25, 60K35
Keywords and Phrases: continuum tree, entrance boundary, fragmentation, gelation, random graph, random tree, Smoluchowski coagulation equation.

1 Introduction

Our topic centers around two closely related models. The *Marcus-Lushnikov process* is the following finite-state continuous-time Markov process [17, 16].

Fix an integer $N \geq 1$ and a *rate kernel* $K(x,y) \geq 0$. Imagine N particles, originally separate, which merge into clusters according to the rule

each pair of clusters, sizes $\{m_i, m_j\}$ say, coalesces into one cluster

of size $m_i + m_j$ at rate $K(m_i, m_j)/N$.

The *general stochastic coalescent* [10] is the continuous-time Markov process whose state space is the infinite-dimensional simplex $\Delta = \{\mathbf{x} = (x_i) : x_i \geq 0, \sum_i x_i = 1\}$,

[1]Research supported by N.S.F. Grant DMS96-22859

where we imagine a state $\mathbf{x}$ as a fragmentation of unit mass into clusters of masses x_i, and the process evolves according to the rule

each pair of clusters, masses $\{x_i, x_j\}$ say, coalesces into one cluster

of mass $x_i + x_j$ at rate $K(x_i, x_j)$.

Provided the rate kernel K is *homogeneous* with some exponent γ

$$K(cx, cy) = c^\gamma K(x, y),\ 0 < c, x, y < \infty \tag{1}$$

we see that the Marcus-Lushnikov process is a special case of the general stochastic coalescent, by taking each particle to have mass $1/N$ and rescaling time.

There is a large literature in various science disciplines (e.g. physical chemistry [9]) on deterministic equations (see section 3) for coalescence. A lengthy survey of deterministic and stochastic models appears in [2]. In particular, there are three special cases which are now well understood: $K(x,y) = 1$ (Kingman's coalescent), $K(x,y) = x + y$ (the additive coalescent), $K(x,y) = xy$ (the multiplicative coalescent). The next five sections focus on five open problems, whose discussion will illustrate some of the known results.

2 The Feller property of the general stochastic coalescent

In making precise the verbal description of the general stochastic coalescent, one would like to prove it is a Markov process with some regularity properties, specifically the *Feller property* that the distribution at time t be weakly continuous as a function of the initial state. Intuitively, this should be true under very weak assumptions, e.g. that $K(x,y)$ is continuous and strictly positive. But a proof is elusive. Evans and Pitman [10] give a proof under stronger assumptions of Lipschitz continuity.

3 Deterministic limits

Studying $t \to \infty$ time asymptotics in these models isn't interesting: the mass all coalesces into a single cluster. Our remaining problems concern different sorts of asymptotics. In the Marcus-Lushnikov process write $\mathbf{ML}^{(N)}(x,t)$ for the (random) number of mass-x clusters at time t. One expects a weak law of large numbers, saying that as $N \to \infty$

$$N^{-1}\ \mathbf{ML}^{(N)}(x,t)\ \ \to^p\ \ n(x,t),\ x \geq 1,\ t \geq 0 \tag{2}$$

where the deterministic limit $n(x,t)$ is the solution of the *Smoluchowski coagulation equation*

$$\tfrac{d}{dt} n(x,t) = \tfrac{1}{2} \sum_{y=1}^{x-1} K(y, x-y) n(y,t) n(x-y,t) - n(x,t) \sum_{y=1}^{\infty} K(x,y) n(y,t) \tag{3}$$

with $n(x,0) = 1_{(x=1)}$. It is these equations which have been studied most intensively in the scientific community [9]. From the verbal description of the model we expect solutions to have the property that mass density is preserved:

$$m_1(t) \equiv \sum_{x=1}^{\infty} x n(x,t) = 1 \ \forall t. \tag{4}$$

This is true [23] under the assumption

$$K(x+y) = O(1+x+y) \tag{5}$$

but in general there might be a phase transition called *gelation*: $\exists T_{\text{gel}} < \infty$ such that

$$m_1(t) = 1, \ t \leq T_{\text{gel}}; \quad m_1(t) < 1, \ t > T_{\text{gel}}.$$

The physical interpretation of gelation is that after the critical time, a strictly positive proportion of mass lies in infinite-mass clusters, the *gel*. Exact conditions on K for gelation or non-gelation are another open problem, but let us return to the question of proving (2), which provides conceptual justification for the deterministic approximation. Proving (2) for $t < T_{\text{gel}}$ is closely related to proving uniqueness of solutions of (3) up to T_{gel}. While this is not difficult under assumption (5), the gelling case seems intrinsically much harder, in that the natural techniques one tries to use would prove regularity of solutions for all time, whereas by definition a gelling kernel has solutions with a certain non-regularity property. Jeon [13] and Norris [18] contain the latest results on such questions.

4 The emerging giant cluster for a gelling kernel

The study of component sizes in the classical random graph process [7] is essentially the same as the study of the Marcus-Lushnikov process with $K(x,y) = xy$. It has long been known that $T_{\text{gel}} = 1$ and that the $N \to \infty$ behavior around the critical point is as follows. For large A, at time $1 - A/N^{1/3}$ the largest cluster has size $\delta N^{2/3}$ for some small δ, and there are many clusters of similar size; at time $1 + A/N^{1/3}$ the largest cluster has size $DN^{2/3}$ for some large D, and the second-largest cluster has size $\delta N^{2/3}$. In other words, a distinguished *giant cluster* emerges over the time interval $1 \pm O(N^{-1/3})$ and it has size $\Theta(N^{2/3})$. See [15] for an exhaustive analysis. Rescaling size and time near the critical point leads to a limit process, the *standard multiplicative coalescent* [1], which is the $K(x,y) = xy$ case of the general multiplicative coalescent, except that one has to enlarge the state space to the l_2 space $\{\mathbf{x} = (x_i) : x_i \geq 0, \sum_i x_i^2 < \infty\}$. Remarkably, the marginal distribution of the standard multiplicative coalescent at a fixed time can be expressed in terms of excursion-lengths of a certain inhomogeneous reflecting Brownian motion.

No other gelling kernel is understood, so it is a matter of speculation to what extent this behavior holds for general gelling kernels. Heuristic arguments of van Dongen [21] suggest that for exponent $1 < \gamma < 2$ there should be an emerging giant cluster of size $N^{2/(1+\gamma)}$, but the only rigorous theory is some weak results in [3].

5 Self-similarity and entrance boundary

Consider a non-gelling kernel K which is homogeneous with some exponent $\gamma \leq 1$. It is natural [11] to seek a solution of the Smoluchowski coagulation equation which is asymptotically *self-similar* (also called *self-preserving* or *scaling*), in the sense that, as $t \to \infty$,

$$n(x,t) = s^{-2}(t)(\psi(x/s(t)) + o(1)) \text{ uniformly in } x \tag{6}$$

where $\psi(x) \geq 0$ satisfies $\int_0^\infty x\psi(x)\ dx = 1$. As at (4), we want the mean density $\int xn(x,t)\ dx$ to be constant in time, which explains the s^{-2} term in (6). Of course, the interpretation of (6) is that cluster mass scales with time as $s(t)$. Moreover one expects

$$s(t) \propto t^{\frac{1}{1-\gamma}},\ -\infty < \gamma < 1 \quad s(t) \propto e^{wt} \text{ for some } w,\ \gamma = 1.$$

Outside the special cases, little is known, though under extra conditions it is probably not hard to prove a "tightness" condition weaker than existence of a self-similar limit (cf. [8] for this result in a slightly different model). For the Marcus-Lushnikov process, this question relates to the time period when typical clusters have size $o(N)$ but not $O(1)$. Reformulating in terms of the stochastic coalescent, we are interested in the time period when typical clusters have mass $o(1)$. When $\gamma < 1$ one expects there to be a unique version of the stochastic coalescent on $0 < t < \infty$ such that the maximum cluster size $\to 0$ as $t \to 0$, and that in this version the empirical distribution of cluster sizes tends (as $t \to 0$, and after rescaling by $s(t)$) to the self-similar distribution ψ. In Markov chain jargon, this is a question about the *entrance boundary*, and is easy to verify ([2] section 4.2) in the case $K(x,y) = 1$. Establishing it more generally seems difficult.

Paradoxically, the two other special cases (kernels xy and $x+y$) seem atypical, in that they have rich entrance boundaries, i.e. there are many different ways to start the process with all the mass in infinitesimally small clusters. See [4, 6] for detailed studies.

6 Connections with d-dimensional models

Our models are "mean-field", in that they do not track positions and velocities of particles in d-dimensional space. This does not mean the models are completely divorced from physical reality. Rather, the details of the physical process under study are used to calculate the form of the rate kernel $K(x,y)$. Perhaps the most interesting case to a probabilist is the original 1916 setting of Smoluchowski [22], who considered particles diffusing under thermal noise, i.e. performing physical Brownian motion in three dimensions, and coalescing upon contact. In this case the appropriate kernel in the mean-field model turns out to be

$$K(x,y) = (x^{1/3} + y^{1/3})(x^{-1/3} + y^{-1/3}).$$

The second term reflects the faster diffusion of smaller particles, the first term reflects their smaller cross-section and hence smaller chance of touching. It is natural to conjecture that, in the full model of spherical masses diffusing by Brownian motion in 3 dimensions and coalescing upon contact (and relaxing to spheres), as time goes to infinity the distribution approximates a Poisson spatial distribution with rescaled mass distribution following the law ψ at (6) predicted by the mean field theory. This has apparently never been studied rigorously, though a less realistic model is treated by Lang and Nguyen [14].

7 The standard additive coalescent

Lest talking about open problems makes it seem little is known, let me end by mentioning a positive result. Cayley's formula says there are N^{N-2} trees on N labeled vertices. Pick such a tree T_∞ at random. To the edges e of T_∞ attach independent exponential(1) r.v.'s ξ_e. Write $F(t)$ for the forest obtained from T_∞ by retaining only the edges e with $\xi_e \leq t$. Write $\mathbf{Y}^{(N)}(t)$ for the vector of sizes of the trees comprising $F(t)$. It can be shown that $(\mathbf{Y}^{(N)}(t); 0 \leq t < \infty)$ is the Marcus-Lushnikov process associated with the additive kernel $K(x, y) = x + y$. This construction was apparently first explicitly given by Pitman [20], although various formulas associated with it had previously been developed in combinatorics [19, 24] and statistical physics [12]. What is remarkable is that one can take $N \to \infty$ limits in this construction. The (rescaled) limit of the discrete tree is the *continuum random tree* (CRT); the analog of cutting edges is placing marks according to a Poisson process of intensity e^{-t} along the skeleton of the CRT. Cutting the mass-1 CRT at these marks splits it into subtrees of finite mass; write $\mathbf{X}(t)$ for the vector of masses of these subtrees at time t. Then (as we expect by analogy with the discrete case above) the process $(\mathbf{X}(t), -\infty < t < \infty)$ evolves as the stochastic coalescent for $K(x,y) = x+y$. This process, the *standard additive coalescent*, is studied in detail in [5]. The CRT itself can be constructed from Brownian excursion, so ultimately the construction of the standard additive coalescent uses only Brownian and Poisson ingredients.

Acknowledgements. This project owes much to joint work and ongoing discussions with Jim Pitman, Steve Evans and Vlada Limic.

References

[1] D. Aldous. Brownian excursions, critical random graphs and the multiplicative coalescent. *Ann. Probab.*, 25:812–854, 1997.

[2] D. J. Aldous. Deterministic and stochastic models for coalescence: a review of the mean-field theory for probabilists. To appear in *Bernoulli*. Available via homepage http://www.stat.berkeley.edu/users/aldous, 1997.

[3] D. J. Aldous. Emergence of the giant component in special Marcus-Lushnikov processes. *Random Structures and Algorithms*, 12:179–196, 1998.

[4] D. J. Aldous and V. Limic. The entrance boundary of the multiplicative coalescent. *Electron. J. Probab.*, 3:1–59, 1998.

[5] D. J. Aldous and J. Pitman. The standard additive coalescent. Technical Report 489, Statistics Dept, U.C. Berkeley, 1997. To appear in Ann. Probab.

[6] D. J. Aldous and J. Pitman. Inhomogeneous continuum random trees and the entrance boundary of the additive coalescent. In Preparation, 1998.

[7] B. Bollobás. *Random Graphs.* Academic Press, London, 1985.

[8] J. M. C. Clark and V. Katsouros. Stable growth of a coarsening turbulent froth. To appear in J. Appl. Probab., 1998.

[9] R. L. Drake. A general mathematical survey of the coagulation equation. In G.M. Hidy and J.R. Brock, editors, *Topics in Current Aerosol Research (Part 2)*, volume 3 of *International Reviews in Aerosol Physics and Chemistry*, pages 201–376. Pergammon, 1972.

[10] S. N. Evans and J. Pitman. Construction of Markovian coalescents. Technical Report 465, Dept. Statistics, U.C. Berkeley, 1996. Revised May 1997. To appear in *Ann. Inst. Henri Poincaré.*

[11] S. K. Friedlander and C. S. Wang. The self-preserving particle size distribution for coagulation by Brownian motion. *J. Colloid Interface Sci.*, 22:126–132, 1966.

[12] E. M. Hendriks, J. L. Spouge, M. Eibl, and M. Shreckenberg. Exact solutions for random coagulation processes. *Z. Phys. B - Condensed Matter*, 58:219–227, 1985.

[13] I. Jeon. Existence of gelling solutions for coagulation fragmentation equations. *Commun. Math. Phys.*, to appear, 1998.

[14] R. Lang and X. X. Nguyen. Smoluchowski's theory of coagulation in colloids holds rigorously in the Boltzmann-Grad limit. *Z. Wahrsch. Verw. Gebeite*, 54:227–280, 1980.

[15] Svante Janson ; Donald E. Knuth ; Tomasz Łuczak and B. Pittel. The birth of the giant component. *Random Structures and Algorithms*, 4:233 – 358, 1993.

[16] A. A. Lushnikov. Certain new aspects of the coagulation theory. *Izv. Atmos. Ocean Phys.*, 14:738–743, 1978.

[17] A. H. Marcus. Stochastic coalescence. *Technometrics*, 10:133–143, 1968.

[18] J. R. Norris. Smoluchowski's coagulation equation: Uniqueness, non-uniqueness and a hydrodynamic limit for the stochastic coalescent. Statistical Lab., Cambridge. To appear in Adv. Appl. Probab., 1998.

[19] Yu. L. Pavlov. Limit theorems for the number of trees of a given size in a random forest. *Math. USSR Subornik*, 32:335–345, 1977.

[20] J. Pitman. Coalescent random forests. Technical Report 457, Dept. Statistics, U.C. Berkeley, 1996. Available via http://www.stat.berkeley.edu/users/pitman.

[21] P. G. J. van Dongen. Fluctuations in coagulating systems II. *J. Statist. Phys.*, 49:927–975, 1987.

[22] M. von Smoluchowski. Drei Vorträge über Diffusion, Brownsche Bewegung und Koagulation von Kolloidteilchen. *Physik. Z.*, 17:557–585, 1916.

[23] W. H. White. A global existence theorem for Smoluchowski's coagulation equation. *Proc. Amer. Math. Soc.*, 80:273–276, 1980.

[24] A. Yao. On the average behavior of set merging algorithms. In *Proc. 8th ACM Symp. Theory of Computing*, pages 192–195, 1976.

David J. Aldous
University of California
Department of Statistics
367 Evans Hall
Berkeley CA 94720-3860
U.S.A.
aldous@stat.berkeley.edu