Some Open Problems – David Aldous

(May 2003)

Here are some open problems that I have (unsuccessfully) thought about in the past, but am not working on now. They are not intended to be "representative" or "the most important" or "the deepest" of all open problems in mathematical probability. The majority are (I think) my own invention and have not been discussed extensively elsewhere.

Contents

1 Existence of Limit Constants in Probabilistic Combinatorics

Often in probabilistic combinatorics one studies random variables X_n associated with size-n structures, and wishes to prove

$$\frac{EX_n}{s(n)} \to c \tag{1}$$

for some natural normalizing sequence $s(n)$ and some limit constant c. When direct methods (seeking to estimate EX_n via "concrete mathematics") fail, the are two more abstract general methods that sometimes work to show existence of a limit c in (1) without giving any explicit value for c.
(i) Subadditivity, used in e.g. the *longest common subsequence* problem ([16] section 6.6) and *first-passage percolation* ([16] section 6.7).
(ii) Weak convergence to a limit random infinite or continuous structure, so that c can be defined in terms of the limit structure. This method can be used for random trees and triangulations [1, 2], or problems such as "how long until random walk on the N-cycle visits every vertex N times?" which reduces to the corresponding question for local time for Brownian motion on the circle.

Indeed there are examples to which both techniques can be applied, such as the Euclidean minimal spanning tree [10, 30].

But there are examples (we list three below) where neither of these techniques seems applicable:

PROBLEM. Find a new general technique for proving existence of limits of the form (1).

Discussion
Here are the three examples we mentioned, in which existence of a limit c has not been proved. The first example is well-known.

(i) *Random* 3-*SAT*. Define a random subset $A_n \subset \{0,1\}^n$ as follows. Pick uniformly at random an increasing triple $1 \le i_1 < i_2 < i_3 \le n$ and a binary triple $(b_1, b_2, b_3) \in \{0,1\}^3$. Then let A_n be the complement of $\{(x_1, \dots, x_n) : x_{i_1} = b_1, x_{i_2} = b_2, x_{i_3} = b_3\}$. Now let $A_n(1), A_n(2), A_n(3) \dots$ be i.i.d. copies of A_n. Finally define

$$X_n = \min\{k : \cap_{i=1}^{k} A_n(i) \text{ is the empty set}\}.$$

A well-known conjecture is that

$$n^{-1}EX_n \to c$$

where Monte Carlo suggests $c \approx 4.2$. See [20] for discussion.

(ii) *Independent sets in sparse random graphs.* Fix $1 < \alpha < \infty$. Consider the random graph $\mathcal{G}(n, \alpha/n)$: so there are n vertices, and each possible edge is present with probability α/n. An *independent set* in a graph is a set of vertices, no two of which are linked by an edge. Let X_n be the maximal size of an independent set in $\mathcal{G}(n, \alpha_n)$. Then it is natural to believe

$$n^{-1}EX_n \to c(\alpha)$$

since by considering isolated vertices we have a lower bound $e^{-\alpha}$ for the limit. See [17] for bounds.

(iii) *Greedy tours.* Take n points i.i.d. uniform on the unit square. Let ξ_1 be one of these point (the rule for choosing ξ_1 shouldn't matter) and then define an ordering $\xi_2, \xi_3, \ldots, \xi_n$ of the remaining point as follows. Given $\xi_1, \ldots, \xi_i$, choose amongst the remaining $n - i$ points the one closest to ξ_i, and let that closest point be ξ_{i+1}. This gives a tour with some length $L_n = d(\xi_1, \xi_2) + d(\xi_2, \xi_3) + \ldots + d(\xi_n, \xi_1)$, where $d(\cdot)$ is Euclidean distance. It is natural to believe

$$n^{-1/2}EL_n \to c$$

for some constant c.

2 Largest common substructures in probabilistic combinatorics

Consider the following general setting. There is a set of n labeled elements $[n] := \{1, 2, \ldots, n\}$. There is an instance $\mathcal{S}$ of a "combinatorial structure" built over these elements. The type of structure is such that for any subset $A \subset [n]$ there is an induced substructure of the same type on A. Three examples of types:
• graphs on vertex-set $[n]$
• partial orders on the set $[n]$
• cladograms (leaf-labeled trees – see below) on leaf-set $[n]$.

Given two distinct instances $\mathcal{S}_1, \mathcal{S}_2$ of the same type of structure on $[n]$, we can ask for each $A \subset [n]$ whether the two induced substructures on A are identical; and so we can define

$$c(\mathcal{S}_1, \mathcal{S}_2) = \max\{\#A : \text{ induced substructures are identical}\}$$

where $\#A$ denotes cardinality. Finally, given a probability distribution μ_n on the set of all structures of a particular type, we can consider the random variable

$$C_n = c(\mathcal{S}_1, \mathcal{S}_2) \text{ where } \mathcal{S}_1, \mathcal{S}_2 \text{ are independent picks from } \mu_n.$$

This general framework includes the following two well-known examples.

Example 1. Suppose the type is "graph" and the distribution μ_n is the usual random graph $G(n, p)$ in which possible edges are independently present with probability p. Given two instances G_1, G_2 of graphs we can define the "similarity" graph G to have an edge (i, j) iff both or neither of G_1, G_2 has the edge (i, j). Then

$$c(G_1, G_2) = \mathrm{cl}(G) := \text{ maximal clique size of } G.$$

Moreover if $\mathcal{G}_1, \mathcal{G}_2$ are independent picks from $G(n, p)$ then their "similarity" is distributed as $G(n, q)$ for $q = p^2 + (1 - p)^2$. Thus C_n is just the maximal clique size of a random graph, a well-understood quantity ([12] section 11.1).

Example 2. Suppose the type is "total order" and μ_n is the uniform distribution on all $n!$ total orders on $[n]$. A few moments thought shows that here C_n is distributed as the longest increasing subsequence of a (single) uniform random permutation. This is again a well-studied quantity, of recent interest because of its connection with extreme eigenvalues of random matrices [8, 11, 28].

Of course these two examples are atypical, in that "by symmetry" a problem about two independent random structures reduces to a problem about one random structure, but they suggest that investigation of other examples may be interesting. Here are two new examples.

Example 3. Figure 1 shows a *cladogram* on $[n]$ (rooted unordered binary tree with non-root leaves labeled by $[n]$) for $n = 11$, together with the sub-cladogram on $A = \{1, 2, 3, 4\}$.

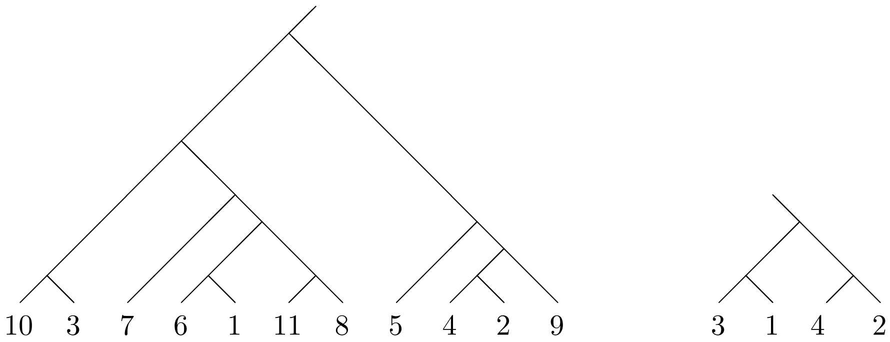

Figure 1. A cladogram on [11] and the induced sub-cladogram on [4].

There are two natural probability measures on n-cladograms:
(a) uniform on all $(2n - 3)!!$ cladograms;
(b) the *coalescent*, starting with n lineages and successively joining two randomly-chosen lineages into one lineage.
We conjecture that in both cases

$$EC_n = n^{\gamma+o(1)}$$

for different constants $\gamma_a, \gamma_b < 1/2$. We do not have conjectures for numerical values, but one can consider continuous limits of the relevant structures and seek to define candidate constants γ in terms of the limit random structures.

Example 4. Amongst several models for random partial orders [13], consider the random two-dimensional partial order on $[n]$. This is the partial order obtained by taking n points $(x_i, y_i),\ 1 \le i \le n$ uniformly randomly in the unit square $[0, 1]^2$ and using the induced "coordinatewise" partial order [29]. Here the natural conjecture is

$$EC_n \sim cn^{1/3}, \text{ for some } 0 < c < \infty. \tag{2}$$

Remarkably, there are two quite different ways to obtain subsets $A \subset [n]$ of size $\approx n^{1/3}$ such that the partial orders agree on A.

(i) Partition $[0,1]^2$ into subsquares of side $n^{-1/3}$. Take B as the set of i such that the i'th point in both processes falls into the same subsquare, so $E\#B = n \times n^{-2/3} = n^{1/3}$. Then take A as a maximal subset of B such that no two of the corresponding subsquares are in the same row or column.
(ii) Take C as the set of i such that in both processes the i'th point is within $n^{-1/3}$ of the reverse diagonal in $[0,1]^2$. Again $\#C$ is order $n^{1/3}$. And one can choose $A \subset C$ with $\#A/\#C$ non-vanishing such that each partial order on A is the trivial partial order.
It is not hard (Graham Brightwell, personal communication) to prove an $O(n^{1/3})$ upper bound using the first moment method. But establishing a value for, or existence of, the presumed limit constant c in (2) may be genuinely hard.

3 Percolation of Averages, in the Mean-Field Setting

We use the following standard setting for "mean-field" (i.e. without d-dimensional geometry) models involving distances between random points. Take n vertices, and for each pair (i,j) let the *distance* $d(i,j) = d(j,i)$ be random with exponential (mean n) distribution, independently over the $\binom{n}{2}$ pairs. For any path $\sigma = v_0, v_1, v_2, \ldots, v_l$ of any length l, write $A_\sigma := l^{-1}\sum_{i=1}^{l} d(v_{i-1}, v_i)$ for the average edge-length. Now for each $c > 0$ define

$$M(n,c) := \max\{l : \exists \text{ some path } \sigma \text{ of length } l \text{ with } A_\sigma \leq c\}.$$

It is fairly easy to see that, as $n \to \infty$ for c fixed,

$$\begin{aligned} M(n,c) &= o(\log n) \text{ if } c < e^{-1} \\ &= \Omega(n) \text{ if } c > e^{-1}. \end{aligned}$$

PROBLEM. Give more details of the behavior near $c = e^{-1}$. In particular, do there exist *scaling exponents* α, β such that

$$n^{-\alpha} M(n, e^{-1} + xn^{-\beta}) \to m(x) \text{ in probability}$$

for some deterministic function $m(x)$ satisfying

$$\lim_{x\to\infty} m(x) = \infty, \quad \lim_{x\to-\infty} m(x) = 0.$$

Discussion

See [7] for references to other problems in this mean-field model of distance. The fact that the first-order critical value equals e^{-1} is mentioned at the end of [4], where the analogous first-order problem for spanning trees is treated in detail. Our problem here asks for second-order behavior, or *finite-size scaling* in the language of statistical physics. We regard the model as a mean-field analog of *first passage percolation*, though the $n^{1/3}$ scaling there (long conjectured and recently proved in various related models [21]) seems unlikely to have a direct parallel in our problem. The corresponding questions for ordinary percolation in this mean-field setting is just a rephrasing of questions concerning emergence of the giant component in the Erdős - Rényi random graph process, where the corresponding critical value is 1 and the scaling exponents are $\alpha = 2/3, \beta = 1/3$ [19, 3]. Returning to our definition of $M(n,c)$, it is natural to conjecture that a broader description of first-order behavior is given by

$$n^{-1}M(n,c) \to m_1(c) \text{ in probability}$$

where the limit function has

$$m_1(c) = 0 \text{ iff } c \leq e^{-1}; \quad m_1(c) = 1 \text{ iff } c \geq c^* \text{ (for some } c^*)$$

Convincing heuristics going back to [22] (see [7] section 5.2) for the mean-field *traveling salesman problem* assert a formula for c^*, numerically $c^* = 2.04$.

4 Mixing Times for the Branch-Rotation Chain on Cladograms (or the Triangulation Walk)

This concerns two problems which are similar (intuitively, at least). Consider the regular n-gon. There are a finite number $C_n = \frac{2n-4)!}{(n-1)!(n-2)!}$ of *triangulations*, that is ways to draw diagonals which partition the n-gon into triangular regions (C_n is a Catalan number: see [27] pages 219–229). One can define a discrete time Markov chain on the space of triangulations of the n-gon as follows. In each step

> pick uniformly at random a diagonal line; delete it, to leave a quadrilateral; then insert the opposite diagonal of that quadrilateral to get a new triangulation.

A different combinatorial set is the set of n-*cladograms*. Such a cladogram, illustrated in figure 1, has leaves labeled $1, 2, \ldots, n$, an unlabeled root (at the top) and binary splits, where we do not distinguish left and right subtrees. (Cladograms are one formalization of *phylogenetic trees* from biological systematics, indicating evolutionary relationships between species. The number of n-cladograms equals $\frac{(2n-2)!}{2^{n-1}(n-1)!}$.) One can define several Markov chains on the set of n-cladograms (see [6, 26] for a more easily-analyzed chain), but the following type of chain seems most interesting.

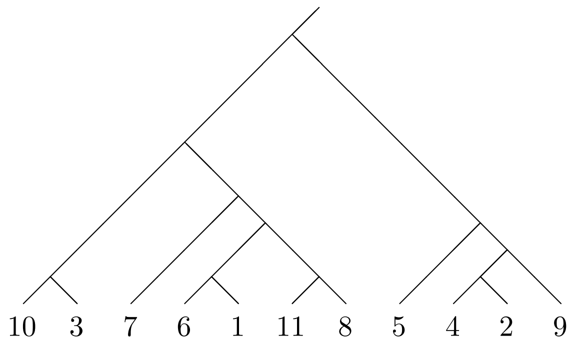

Figure 1. A cladogram on 11 species.

A n-cladogram has $2n-1$ edges. Pick one edge (not the edge at the root) uniformly at random; in figure 1, say we pick the edge upwards from the common ancestor of $\{11, 8\}$. Cut this edge at its top, thus separating the 3-edge subcladogram on $\{11, 8\}$ and making the two other edges at the cutpoint merge into a single edge e from the common ancestor of $\{6, 1\}$ to the common ancestor of $\{7, 6, 1\}$. Now there are exactly 4 edges adjacent to e, viz the edges leading upwards from

$$6, \quad 1, \quad 7, \quad \text{the common ancestor of } \{7, 6, 1\}.$$

Pick each of these 4 edges with chance $1/4$, and reattach the subcladogram to the middle of that edge. If we picked the edge upwards from the common ancestor of $\{7, 6, 1\}$, then we would obtain the cladogram in figure 2. (in general e might have less than 4 adjacent edges, in which case with the remaining probability we make no change).

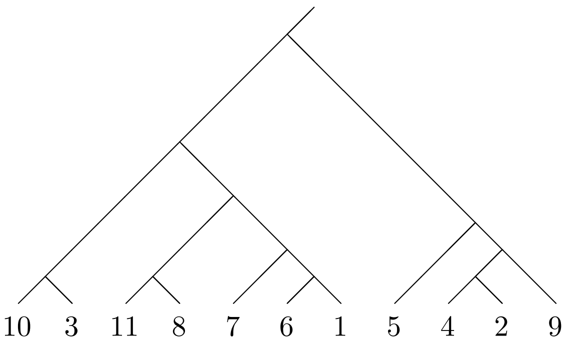

Figure 2. A step of the chain from figure 1.

Being reversible, each chain has largest eigenvalues $1 = \lambda_1 > \lambda_2$, and the *relaxation time* defined as $1/(1 - \lambda_2)$ has an interpretation as a mixing time parameter. In each chain it is easy to show (by the usual technique of applying the variational characterization of λ_2 to a suitable test function) that the relaxation is at least order $n^{3/2}$ as $n \to \infty$.

PROBLEM. For each chain, show that the relaxation time is at most order $n^{3/2}$.

Discussion
There are good heuristic reasons (too lengthy to explain here) for expecting these two chains to have similar behavior. The chain on triangulations is discussed in [24, 25], who obtain an $O(n^4)$ upper bound. Over the last 20 years, techniques has been developed which enable one to find the correct order of magnitude of the mixing times for natural random walks on familiar combinatorial structures; this "triangulation walk" example is perhaps the simplest structure for which correct order is unproved. The "cladograms" chain has a semi-applied story. Reconstructing phylogenetic trees from actual biological data is a large-scale academic activity; it involves algorithmically hard optimization problems which in practice are attacked via

heuristic "local search" methods, exploring the space of cladograms to find a "best fit" to data. One class of algorithms uses MCMC (Markov chain Monte Carlo), built over a "base chain" like ours. The data-dependent chains which arise in practice are so complicated that rigorous theoretical analysis seems hopeless, but understanding the base "no data" chain is a natural first step.

5 Spectral Gap for the Interchange (Exclusion) Process on a Finite Graph

Consider a n-vertex graph – assume connected, undirected. Take n particles labeled $1, 2, \ldots, n$. In a *configuration*, there is one particle at each vertex. The *interchange process* is the following continuous-time Markov chain on configurations. For each edge (i, j), at rate 1 the particles at vertex i and vertex j are interchanged.

The interchange process is a reversible, and its stationary distribution is uniform on all $n!$ configurations. There is a *spectral gap* $\lambda_{IP}(G) > 0$, which is the smallest non-zero eigenvalue of the transition rate matrix. If instead we just watch a single particle, it performs a continuous-time random walk on G, which is also reversible and hence has a spectral gap $\lambda_{RW}(G) > 0$. Simple arguments (the *contraction principle [9]*) show $\lambda_{IP}(G) \leq \lambda_{RW}(G)$.

PROBLEM. Prove $\lambda_{IP}(G) = \lambda_{RW}(G)$ for all G.

Discussion

Fix $m < n$ and color particles $1, 2, \ldots, m$ red. Then the red particles in the interchange process behave as the usual *exclusion process* [23]. But in the finite setting, the interchange process seems more natural.

The problem arose in conversation with Persi Diaconis (see e.g. [15]). It has been proved in various special cases, such as trees [18].

6 Self-similarity for Coalescing Regions of R^2

Consider the two-dimensional plane partitioned in regions. At time zero the region are unit squares. As time increases, two regions which are adjacent (that is, have a common boundary segment) may merge into one region (the common boundary line disappears). For adjacent regions A, B, this merger rate $r(A, B)$ depends on the geometry of A and B (that is, some function invariant under Euclidean transformations, e.g. dependent on the areas of A and B and the length of their common boundary).

PROBLEM. Find a rate function $r(\cdot)$ for which one can prove this process is asymptotically self-similar, i.e. as $t \to \infty$ the configuration, with space rescaled by some deterministic $s(t)$, converges to some limit random partition of the plane.

Discussion
This arises from my interest in mean-field models of coalescence [5]. But apparently no problem of quite this type has been studied in the $d > 1$-dimensional setting. In the opposite process of fragmentation in d dimensions, it is easy to see (e.g. [14]) that some simple models where each region splits independently in some way lead to asymptotically self-similar processes. However, the limit processes have some long-range dependence which one doesn't expect with our "locally-specified" coalescence process.

References

[1] D.J. Aldous. The continuum random tree II: an overview. In M.T. Barlow and N.H. Bingham, editors, *Stochastic Analysis*, pages 23–70. Cambridge University Press, 1991.

[2] D.J. Aldous. Recursive self-similarity for random trees, random triangulations and Brownian excursion. *Ann. Probab.*, 22:527–545, 1994.

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

[4] D.J. Aldous. On the critical value for percolation of minimum-weight trees in the mean-field distance model. *Combin. Probab. Comput.*, 7:1–10, 1998.

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

[6] D.J. Aldous. Mixing time for a Markov chain on cladograms. *Combin. Probab. Comput.*, 9:191–204, 2000.

[7] D.J. Aldous. The $\zeta(2)$ limit in the random assignment problem. *Random Structures Algorithms*, 18:381–418, 2001.

[8] D.J. Aldous and P. Diaconis. Longest increasing subsequences: From patience sorting to the Baik-Deift-Johansson theorem. *Bull. Amer. Math. Soc.*, 36:413–432, 1999.

[9] D.J. Aldous and J.A. Fill. Reversible Markov chains and random walks on graphs. Book in preparation, 2001.

[10] D.J. Aldous and J.M. Steele. Asymptotics for Euclidean minimal spanning trees on random points. *Probab. Th. Rel. Fields*, 92:247–258, 1992.

[11] J. Baik, P.A. Deift, and K. Johansson. On the distribution of the length of the longest increasing subsequence of random permutations. *J. Amer. Math. Soc.*, 12:1119–1178, 1999.

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

[13] G. Brightwell. Models of random partial orders. In *Surveys in Combinatorics, 1993 (Keele)*, pages 53–83. London Math. Soc., 1993.

[14] F.K.C. Chen and R. Cowan. Invariant distributions for shapes in sequences of randomly-divided rectangles. *Adv. in Appl. Probab.*, 31:1–14, 1999.

[15] P. Diaconis and L. Saloff-Coste. Comparison theorems for reversible Markov chains. *Ann. Appl. Probab.*, 3:696–730, 1993.

[16] R. Durrett. *Probability: Theory and Examples*. Wadsworth, Pacific Grove CA, 1991.

[17] A. Frieze. On the independence numberof random graphs. *Discrete Math.*, 81:171–175, 1990.

[18] S. Handjani and D. Jungreis. Rate of convergence for shuffling cards by transpositions. *J. Theoretical Probab.*, 9:983–93, 1996.

[19] S. Janson, D. E. Knuth, T. Łuczak, and B. Pittel. The birth of the giant component. *Random Structures Algorithms*, 4:233–358, 1993.

[20] S. Janson, Y.C. Stamatiou, and M. Vamvakari. Bounding the unsatisfiability threshold of random 3-SAT. *Random Structures Algorithms*, 17:103–116, 2000.

[21] K. Johansson. Transversal fluctuations for increasing subsequences in the plane. *Probab. Th. Rel. Fields*, 116:445–456, 2000.

[22] W. Krauth and M. Mézard. The cavity method and the travelling-salesman problem. *Europhys. Lett.*, 8:213–218, 1987.

[23] T.M. Liggett. *Interacting Particle Systems*. Springer–Verlag, 1985.

[24] L. McShine and P. Tetali. On the mixing time of the triangulation walk and other Catalan structures. In *Randomization Methods in Algorithm Design (Princeton NJ 1997)*, number 43 in DIMACS Ser. Discrete Math. Theoret. Comput. Sci., pages 147–160. Amer. Math. Soc., 1999.

[25] M. Molloy, B. Reed, and W. Steiger. On the mixing time of the triangulation walk. In *Randomization Methods in Algorithm Design (Princeton NJ 1997)*, number 43 in DIMACS Ser. Discrete Math. Theoret. Comput. Sci., pages 179–190. Amer. Math. Soc., 1999.

[26] J. Schweinsberg. An $O(n^2)$ bound for the relaxation time of a Markov chain on cladograms. Technical Report 572, Statistics Dept., U.C. Berkeley, 2000.

[27] R.P. Stanley. *Enumerative Combinatorics*, volume 2. Cambrdige University Press, 1999.

[28] C.A. Tracy and H. Widom. Random unitary matrices, permutations and Painlevé. *Comm. Math. Phys.*, 207:665–685, 1999.

[29] P. Winkler. Random orders. *Order*, 1:317–331, 1985.

[30] J.E. Yukich. *Probability Theory of Classical Euclidean Optimization Problems*. Number 1675 in Lecture Notes in Math. Springer, 1998.