Some of my favorite open problems

David Aldous

19 September 2018

I will talk about a few of the 20 open problems posted on my web site at www.stat.berkeley.edu/∼aldous/Research/OP/index.html.

- Random Eulerian circuits.
- Topological properties of a random partition of the plane.
- The Lake Wobegon process
- Compactifications of finite reversible Markov chains.

The first 3 are easy-to-describe models, for whose properties we have heuristics partly supported by simulation but cannot do rigorous proofs. The 4th is more abstract – could likely be done by known functional analysis/metric space theory.

1. Random Eulerian circuits

A rather obvious observation in introductory graph theory is

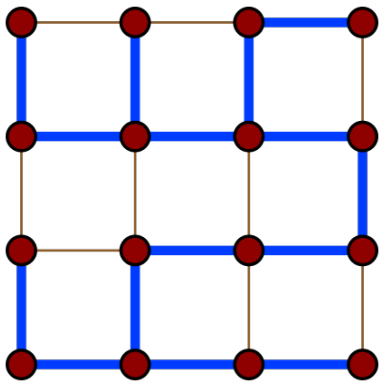
Lemma

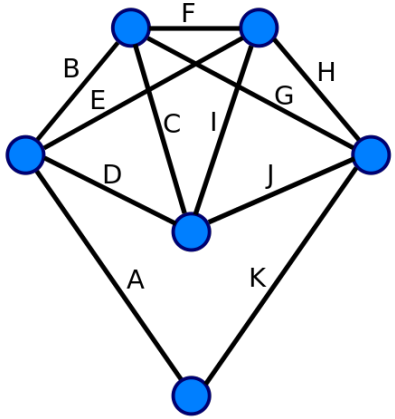
A finite connected undirected graph has a **spanning tree**, *that is a connected edge-subgraph which is a tree.*

Euler proved what's often regarded as "the first theorem in graph theory".

Theorem

A finite, strongly connected, directed graph which is balanced (each vertex has in-degree = out-degree) has an **Eulerian circuit** *using each edge exactly once.*





The field **probabilistic combinatorics** studies probability distributions over combinatorial objects. On the theory side, best known is the Erdős-Rényi model for random graphs. For instance, on the applied side, over the last 15 years there has been $>$ 20K papers on (allegedly) realistic models for random networks.

Fundamental math theory starts by studying the *uniform* distribution over a specified set of objects. So there is a well-defined concept of a *uniform* random spanning tree and a *uniform* random Eulerian circuit within a given graph.

It turns out that there is a large literature on uniform random spanning trees, because they relate to many other discrete structures – see Lyons - Peres monograph *Probability on Trees and Networks*.

In contrast, very little literature on uniform random Eulerian circuits – curious because there's a surprising connection between the two topics.

In a balanced directed graph, take any spanning tree, with directed edges toward an arbitrary root. From the root do an arbitrary walk, at each stage choosing an unused edge but saving the spanning-tree-edge until last. This always gives an Eulerian circuit [easy].

True (but not obvious) that with a uniform random spanning tree and uniform random walk-step choices we get a *uniform* random Eulerian circuit.

We need two more facts.

Fact 1: Simple random walk on the infinite lattice $\mathbb{Z}^d$ is recurrent (always returns sometime to starting point) in $d = 1, 2$ but not in $d \geq 3$.

Fact 2: It is *quite easy* to simulate a uniform random spanning tree.

Now as a simple example consider the discrete torus $\mathbb{Z}_N^d$. Replace each edge by 2 directed edges. So in-degree = out-degree = 2d. Any Eulerian circuit consists of 2d "loops" from the origin.

The facts above enable us to simulate a uniform random Eulerian circuit of this graph – done as a student project. Intuition (from random walk theory, Fact 1) suggests, and simulation supports, the following.

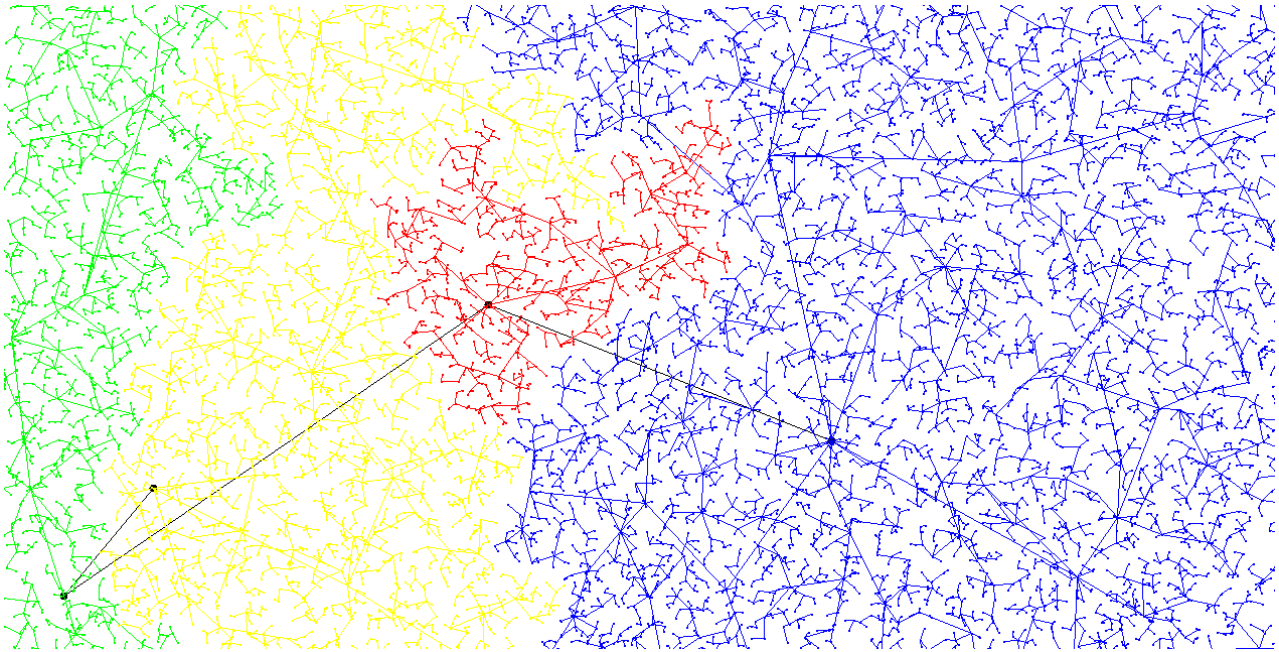
Open problem. *Prove (in fixed $d \geq 3$) that of the $2d$ loops at the origin, some have length $O(1)$, all others have length of order N^d as $N \to \infty$.*

Never been studied – no idea how to prove – can't do theoretical analysis of algorithm output.

2. Topological properties of a random partition of the plane.

Choose $k \geq 2$ distinct points $z_1, \ldots, z_k$ in the unit square, and assign to point z_i the color i from a palette of k colors. Take i.i.d. uniform random points $U_{k+1}, U_{k+2}, \ldots$ in the unit square, and inductively, for $j \geq k+1$,

> *give point U_j the color of the closest point to U_j amongst $U_1, \ldots, U_{j-1}$ where we interpret $U_i = z_i, 1 \leq i \leq k$.*



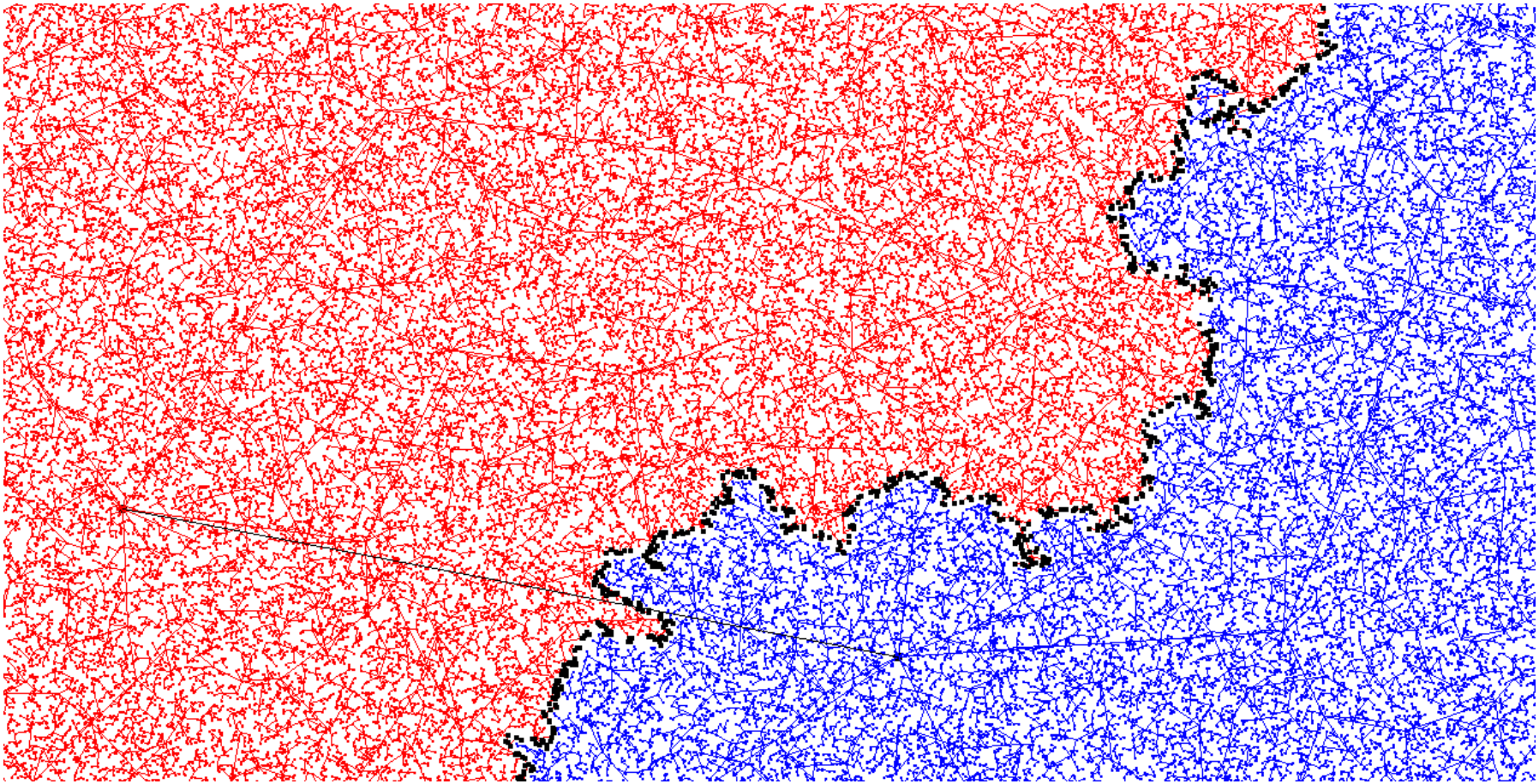
Simulations and intuition strongly suggest that there is (in some sense) **convergence** to an $n \to \infty$ limit which is a random partition of the square into k colored regions. This "coloring model" was considered independently by several people over last 10 years.

At first sight the convergence assertion seems easy.

First consider Voronoi regions. Intuitively, the area of the Voronoi region of a given color should behave almost as a martingale, because a new particle near the boundary seems equally likely to make the area larger or smaller. If one could bound the martingale approximation well enough to establish a.s. convergence of such areas, the convergence theorem would follow rather trivially. But doing so seems to require detailed knowledge of the geometry of the boundary.

So what does the boundary look like?

Simulations suggest that the boundaries between these limit regions should be fractal, in some sense.



Aldous (2018) proves (by different methods) that the limit random partition exists in a certain weak sense. With Preater (2009) it follows that the boundaries have zero area.

Open problem. Does the boundary have fractal dimension $= 1$ or > 1 ?

Why care about this particular "interface" model (intense deep current study of quite different interface models – SLE and KPZ).

It has a surprising connection to another model, "empires". Think of the coloring points arriving infinitely quickly to instantly create the random partition from given discrete "seeds" (the initial colored points), and envisage countries with capital cities. Introduce a "time" parameter $-\infty < t < \infty$ and take the "seeds" at time t as a Poisson point process ("completely random") on the infinite plane with mean e^t seeds per unit area. We now have a process in which the "countries" split into two countries (and the new country gets a capital city). This process has no simple description in "forwards" time, but it does in reversed time.

[show simulation]

What's the background story here? Many years ago I tried to study models of randomly coalescing partitions of the plane, but too difficult to prove anything. The time-reversed process above evolves according to a simple rule:

Each country is liable to be absorbed at rate 1 *per unit time; if so it is absorbed by the country whose capital city is closest to its capital city.*

From the explicit construction, this process automatically has an exact self-similarity property (from self-similarity of the space-time Poisson process).

Open problem. For other models of randomly coalescing partitions of the plane, we expect either asymptotic self-similarity or appearance of one infinite country (as in bond percolation theory). Can any other model be analyzed? [The Math Rule of 3].

3. The Lake Wobegon process

There are two quite well known models for "sequentially randomly placing articles into piles" which turn out to be parts of two different larger areas of probabilistic combinatorics. Both models have their own Wikipedia page. One is the **Chinese restaurant process** relating to the Poisson-Dirichlet distribution, population genetics models and Bayes priors for categorical data. The other is **patience sorting**, relating to the longest increasing subsequence of a random permutation and thence to Young tableaux.

In patience sorting, we imagine cards with i.i.d. random real values from a continuous distribution. Piles are labeled $1, 2, \ldots$ left-to-right in order of creation, and at any time the values on the top card in each pile are in increasing order. The rule for placing the current card is

place on top of the first (leftmost) pile whose top card has higher value than the current card

and if there is no such pile then start a new pile at the right end.

Are there other interesting processes of this type? As rather different background, recall that in the (fictional) *Lake Wobegon*

> *all the women are strong, all the men are good looking, and all the children are above average.*

In a finite list of numbers, we cannot arrange that each item is greater than the average of the whole list, but we can arrange that each item is greater than the average of the *previous* items.

So let us invent the **Lake Wobegon process** which is a variant of the patience sorting process. At any time the **average** of each pile is visible, and these averages are in increasing order of the piles. So the rule for placing the current card is

place on top of the first (leftmost) pile whose average value is higher than the current card

and if there is no such pile then start a new pile at the right end.

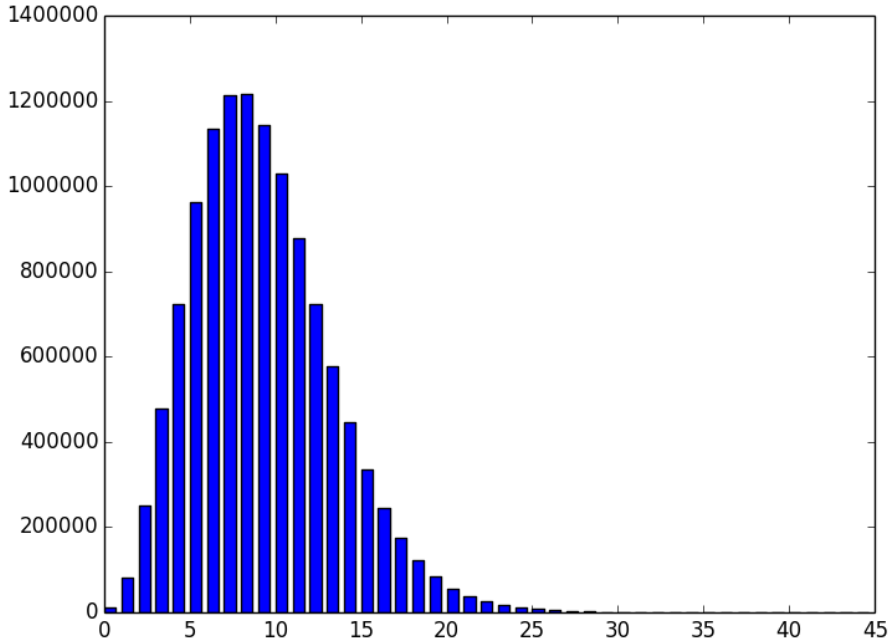
This rule looks pretty arbitrary, but here is a made-up story.

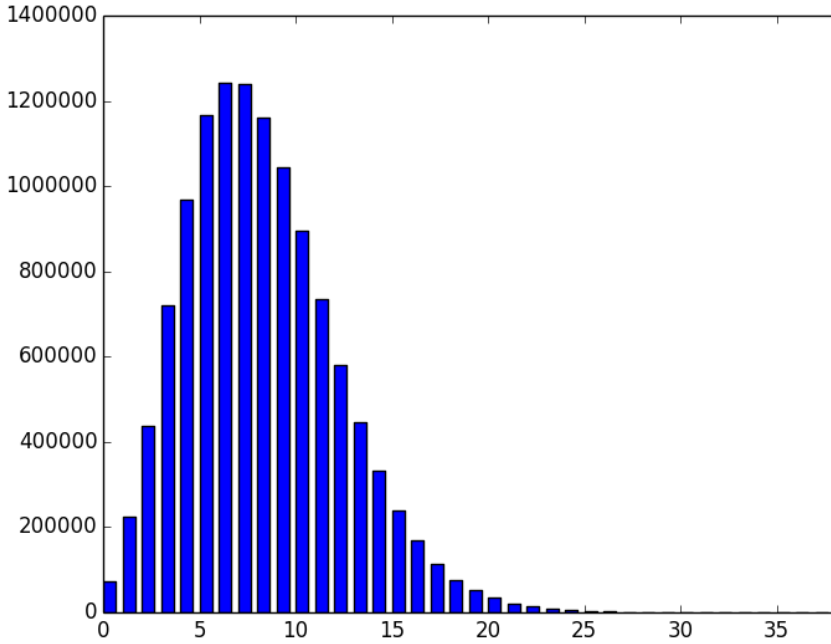
We envisage this process in terms of mathematics journals. Submitted papers have a quality, uniform random on $[0, 1]$, with 0 the best quality. Each journal only accepts papers whose quality is greater than the average of their previous accepted papers So the journals are ranked as 1st, 2nd, 3rd, . . . in decreasing order of prestige; authors then arrange to publish in the most prestigious journal that will accept.

[show simulation]

Here's what happens with 12 million papers.

X_n = journal sizes after n papers.
n = 12 million.





Open problem. *Describe the limiting behavior of the distribution of X_n as $n \to \infty$.*
Convincing heuristics say $\mathbb{E}X_n$ grows as $\log n$ but it is not apparent how the variance behaves.

4. Compactifications of finite reversible Markov chains.

Background. Two metric spaces are **isometric** if there exists a bijection which preserves distance (therefore continuous in both directions). Different-looking metric spaces like $\mathbb{R}^1$ and $\mathbb{R}^2$ are usually not isometric.

Two probability spaces are **isomorphic** if there exists a bijection which is measurable in both directions and preserves the probability measure. Counter-intuitively, it is roughly true (omitting lots of details in this section) that **all** interesting probability spaces are isomorphic to $([0,1], \mathrm{Leb})$.

Markov chains. A discrete-time Markov chain on n states $\{i, j, \ldots\}$ is specified by a transition matrix (p_{ij}) and has some time-t distribution

$$p_n(i,j,t) = \mathbb{P}(X(t) = j | X(0) = i).$$

Let's suppose (p_{ij}) is symmetric; then the chain's stationary distribution is uniform on the n states.

The $n \to \infty$ limit processes we study will typically have continuous state space and continuous time $0 \le t < \infty$. Working measure-theoretically we can take state space and stationary distribution to be $([0,1], \mathrm{Leb})$.

A "compactification" result conjectured by me and proved in a weak form by Henry Towsner (Limits of sequences of Markov chains, *Electron. J. Probab.* 2015).

Theorem

An arbitrary sequence of symmetric Markov chains with $n \to \infty$ has a subsequence in which (after time-scaling) the Markov chain either

- *has the L^2 cutoff property*
- *or converges (in a certain subtle sense) to a limit Markov process of the form described below.*

The form of a limit process

Consider measurable functions $p^\infty(x, y, t)$ for $x, y \in [0, 1]$ and $t > 0$ such that

- $p^\infty(x, y, t) \equiv p^\infty(y, x, t)$.
- $y \to p^\infty(x, y, t)$ is a probability density function.
- $p^\infty(x, z, t + s) = \int p^\infty(x, y, t) p^\infty(y, z, s) dy$. (Chapman-Kolmogorov)
- some $t \downarrow 0$ pinning.

This specifies the finite-dimensional distributions of a symmetric Markov process on $[0, 1]$ started at x.

Our intuition is that any "natural" sequence of Markov chains will have some "natural" limit process on some nice topological space. To prove the general theorem we needed a trick to map states to $[0, 1]$ – informally, to prove convergence of a sequence of objects one needs them to be the same type of object.

So can we recover this topology from the limit analytic description?

Conjecture. There is some natural way to define a metric, for instance

$$d(x_1, x_2) := \sqrt{\int \int (p^\infty(x_1, y, t) - p^\infty(x_2, y, t))^2 e^{-t} dy dt}$$

which makes $[0, 1]$ into a complete separable metric space and makes the Markov process have the Feller property.

This sounds, and is, very abstract, but actually has some real world interest. In some of the 20K papers on random/complex networks, an "edge" signifies existence of a given relationship. But in most cases it is not a 0 - 1 relationship but there is a quantitative "strength of relationship" better modeled as an edge-weighted graph. And an edge-weighted graph is essentially the same as a symmetric continuous-time Markov chain. So a "mathematically natural" metric d above would also serve as a natural distance function on vertices of an edge-weighted graph.