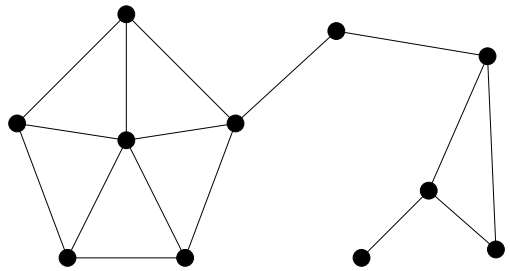
Chapter 6

Random Networks and Algorithms

Most of this course has involved developing the combinatorial tools, techniques and structures needed for discrete mathematical modeling. Our final topic is an active area of current research, and builds upon some of the "average-case" problems we have been considering.

Let us begin with the following basic question. Under an appropriate random model, how can we solve "typical" cases of a combinatorial decision or optimization problem? The study of such questions turns out to have close connections both with physics (phase transitions) and with computer science (algorithmic performance on typical problem instances).

Example 6.1. As motivation, let us consider the graph coloring problem that we have seen several times this semester. Given a simple unweighted graph $G = (V, E)$, is $\chi(G) \leq 3$? We know that if G contains certain identifiable structures, we can quickly answer the question. For instance, suppose we are given the following graph:



The presence of a wheel (cycle C_5 with each vertex connected to a hub) within G rules out any possibility of coloring it with only 3 colors, so the answer must be no. But we may not always be lucky enough to find structures of this sort. For general graphs, we have studied a depth-first search method for answering the chromatic number question, and have seen that in the worst case, that method could take exponential time to answer.

Now, however, suppose that the graph is randomly generated. What can we say, in the *typical case*, about the question of whether $\chi(G) \leq 3$?

- If $|E| \ll |V|$, the answer is almost certainly yes. The graph is sufficiently sparse, so there are few constraints. We can very likely color the graph with only 3 colors without violating any constraints.

- If $|E| \gg |V|$, the answer is almost certainly no. The graph is dense, so there are many constraints. It is quite likely that we will need to violate one of these, if we only use 3 colors.

The hard cases are those that lie in between these two extremes. There, we have no easy means of predicting whether a valid coloring exists. The only way is to check, and in general, that is slow.

Let us keep this scenario in mind, since it will turn up again soon, and we will be able to refine considerably our statement about computational hardness in the different graph density regimes.

6.1 Random graphs

In one of the homework assignments, we saw the $\mathcal{G}(n,p)$ random graph ensemble, proposed by Paul Erdős and Alfréd Rényi in 1959.

Definition 6.2. *If $G=(V,E)$ is drawn from the graph ensemble $\mathcal{G}(n,p)$, then $|V|=n$, and edges $e\in E$ are placed independently at random, with probability p, between any pair of vertices:*

$$\forall u\neq v\in V,\quad \{u,v\}\in E \text{ iff } X_{\{u,v\}}=1,$$

where $X_{\{u,v\}}\sim \text{Bernoulli}(p)$, i.i.d.

Note that since edges are placed independently, these graphs have no geometry. They are completely nonmetric, so for instance,

$$\forall u,v,w\in V,\quad \Pr[\{u,v\}\in E\,|\,\{u,w\}\in E,\{v,w\}\in E]=\Pr[\{u,v\}\in E]=p,$$

i.e., the existence of edge $\{u,v\}$ is completely independent of the existence of edges $\{u,w\}$ and $\{v,w\}$. So an edge cannot possibly represent any distance property satisfying the triangle inequality.

What do these graphs look like when n is large, and $p=\alpha/n$ for some fixed constant α? First consider the degree of a vertex:

$$\deg(v)=\big|\{u:\{u,v\}\in E\}\big|=\sum_{u\neq v\in V} X_{\{u,v\}}.$$

It follows from linearity of expectation that

$$\mathbb{E}[\deg(v)]=\sum_{u\neq v\in V}\mathbb{E}[X_{\{u,v\}}]=\sum_{u\neq v\in V} p=(n-1)p=\frac{n-1}{n}\,\alpha,$$

so for large n, the mean degree of the graph is approximately the fixed constant α. Moreover, the sum of Bernoulli random variables is a binomial random variable, so

$$\deg(v)\sim \text{Bin}(n-1,p)\sim \text{Bin}(n-1,\alpha/n).$$

In the large n limit, this becomes a Poisson random variable:

$$\deg(v)\sim \text{Poisson}(\alpha),$$

i.e., $\Pr[\deg(v)=k]=e^{-\alpha}\alpha^k/k!$

We now consider two cases for α:

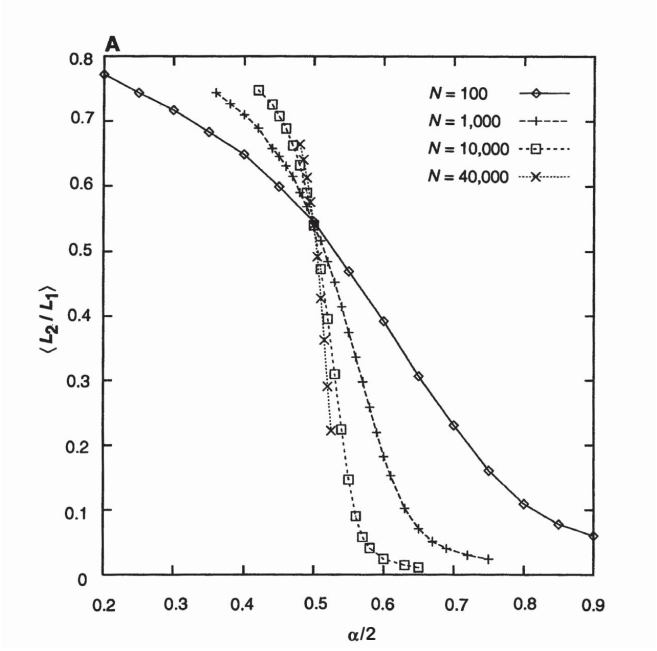
- $\alpha<1$. In this case, all components in the graph are very close to being trees. If we consider a vertex v to be the root of a tree, the probability of the tree having depth $\geq r$, for some fixed r, scales roughly as e^{-r} (think of it as r independent events taking place). Consequently, the tree dies off rapidly. A more careful analysis of the underlying "branching process" shows that all components in the graph are of size $O(\log n)$, and that the graph is very likely to be a *pseudoforest*: each component contains at most one single cycle.

- $\alpha > 1$. In this case, paths from v proliferate exponentially. The branching process no longer dies off rapidly. One can show that there is now a unique *giant component*, of size $O(n)$, and so a finite fraction of vertices are members of it. The rest of the components turn out to be small ones, of size $O(\log n)$.

But we skipped the case of $\alpha = 1$. What happens there?

At $\alpha = 1$, the structure of the graph changes very suddenly. By analogy with physical models, we call this structural change a *phase transition*, and $\alpha = 1$ is the *critical point*. The $\alpha < 1$ case is called *subcritical*, and the $\alpha > 1$ case is called *supercritical*.

We can understand the phase transition quantitatively, as follows. Define L_k to be the size of the kth-largest component in the graph. If we generate hundreds of random graphs at different values of α, take the quantity L_2/L_1 averaged over all these graphs, and plot the results as a function of α, we get the following curves for increasing n (from a 1994 paper by Kirkpatrick and Selman):



As n increases, these curves become increasingly steep at $\alpha = 1$, approaching a step function in the limit $n \to \infty$. That represents the separation between the subcritical phase, where no giant component exists and so $L_2/L_1 = O(1)$, and the supercritical phase, where a giant component exists and so $L_2/L_1 = O(\log n/n)$.

One way of characterizing the sharpness of the transition is as follows. For a given graph size n, let $\alpha_1(\epsilon)$ be the value at which $L_2/L_1 = 1 - \epsilon$, and let $\alpha_2(\epsilon)$ be the value at which $L_2/L_1 = \epsilon$. If we take some small $\epsilon > 0$, then we may define the width of the *critical window* to be

$$\lambda(\epsilon) = \alpha_2(\epsilon) - \alpha_1(\epsilon),$$

i.e., the width of the window in which L_2/L_1 drops from very near 1 to very near 0. As n increases, $\lambda(\epsilon)$ shrinks. Precisely *how* $\lambda(\epsilon)$ scales in n is one of the key factors in describing the nature of the phase transition.

Note that a sharp transition of this kind can arise as a simple consequence of the central limit theorem. In such cases, which are manifestations of *Kolmogorov's zero-one law*, one can often calculate the critical window width exactly. This is shown in the following example.

Example 6.3 (Biased coin tossing)**.** A biased coin has probability p of landing on heads. If the coin is tossed n times, what is the probability that it will land on tails more often than heads?

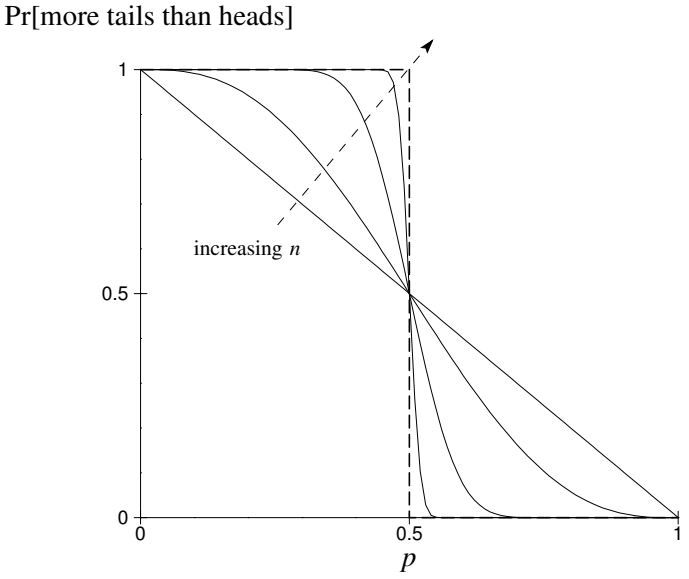
Let X be the number of times the coin lands on heads, out of n trials. In that case,

$$\mathbb{P}\mathrm{r}[\text{more tails than heads}] = \mathbb{P}\mathrm{r}[X < n/2].$$

But X is simply a binomial random variable $X \sim \mathrm{Bin}(n, p)$, so

$$\mathbb{P}\mathrm{r}[\text{more tails than heads}] = \sum_{k=0}^{\lceil n/2-1 \rceil} \binom{n}{k} p^k (1-p)^{n-k},$$

which looks like



It is not difficult to show that in this case, the width of the critical window is $\lambda = \Theta(1/\sqrt{n})$. One way of showing it is to consider the large n limit, where $\mathrm{Bin}(n, p)$ approaches a normal random variable $\mathrm{N}(np, np(1-p))$. Then

$$\begin{aligned}
\mathbb{P}\mathrm{r}[\text{more tails than heads}] &\approx \int_{-\infty}^{n/2} \frac{1}{\sqrt{2\pi np(1-p)}}\, e^{-(x-np)^2/2np(1-p)}\, dx \\
&= \int_{-\infty}^{n\left(\frac{1}{2}-p\right)} \frac{1}{\sqrt{2\pi np(1-p)}}\, e^{-x^2/2np(1-p)}\, dx \\
&= \int_{-\infty}^{\sqrt{\frac{n}{p(1-p)}}\left(\frac{1}{2}-p\right)} \frac{1}{\sqrt{2\pi}}\, e^{-u^2/2}\, du.
\end{aligned}$$

In the limit $n \to \infty$, if p is bounded away from $1/2$, the upper limit of integration is either $-\infty$ or $+\infty$, and so $\mathbb{P}\mathrm{r}[\text{more tails than heads}]$ is either 0 or 1. But what happens if p is very close to $1/2$? Rewriting p as

$$p = \frac{1}{2} + \frac{y}{\sqrt{n}},$$

where y is a constant, it follows that

$$\begin{aligned}\mathbb{P}\text{r}[\text{more tails than heads}] &\approx \int_{-\infty}^{-y\Big/\sqrt{\left(\frac{1}{2}+\frac{y}{\sqrt{n}}\right)\left(\frac{1}{2}-\frac{y}{\sqrt{n}}\right)}} \frac{1}{\sqrt{2\pi}}\, e^{-u^2/2}\, du \\ &\approx \int_{-\infty}^{-2y} \sqrt{\frac{1}{2\pi}}\, e^{-u^2/2}\, du \quad \text{for } n \to \infty.\end{aligned}$$

This limiting function is exactly $\Phi(-2y)$, where Φ is the cumulative distribution function for the standard normal distribution. So if y_1 is the value of y at which $\mathbb{P}\text{r}[\text{more tails than heads}] = 1 - \epsilon$ and y_2 is the value of y at which $\mathbb{P}\text{r}[\text{more tails than heads}] = \epsilon$, then for large n,

$$\Phi(-2y_1) = 1 - \epsilon \quad \text{or} \quad y_1 = -\frac{\Phi^{-1}(1-\epsilon)}{2}$$

and

$$\Phi(-2y_2) = \epsilon \quad \text{or} \quad y_2 = -\frac{\Phi^{-1}(\epsilon)}{2} = \frac{\Phi^{-1}(1-\epsilon)}{2}.$$

The width of the critical window is then given by

$$\lambda = p_2 - p_1 = \left(\frac{1}{2} + \frac{y_2}{\sqrt{n}}\right) - \left(\frac{1}{2} + \frac{y_1}{\sqrt{n}}\right) = \frac{y_2 - y_1}{\sqrt{n}} = \frac{\Phi^{-1}(1-\epsilon)}{\sqrt{n}}.$$

Consequently, for increasing n, λ shrinks as $\Theta(n^{-1/2})$.

Calculating the width of the critical window for the phase transition in random graphs is far less straightforward, but it is known that there, $\lambda = \Theta(n^{-1/3})$.

6.2 Graph coloring

We now return to the problem of graph coloring, on random graphs. Suppose that a graph G is drawn from the $\mathcal{G}(n,p)$ ensemble, for some fixed values of n and p. Is $\chi(G) \leq 3$?

6.2.1 Sparse case

We saw earlier that if $|E| \ll |V|$, the graph is so sparse that the answer is almost certainly yes. But now we can make this statement both more precise and stronger.

Recall that a *subcritical* random graph ($\alpha < 1$) is very likely to be a pseudoforest. This means that *with high probability* (with probability $1 - o(1)$, often abbreviated "w.h.p."), no component contains more than one cycle. But we know that at most 3 colors are needed to color the cycle, and at most 2 colors are needed to color the tree that makes up the rest of the component. So if G is drawn from $\mathcal{G}(n,p)$ with $p = \alpha/n$, for $\alpha < 1$,

$$\chi(G) \leq 3 \quad \text{w.h.p.}$$

6.2.2 Dense case

At the other extreme, where $|E| \gg |V|$, the graph is so dense that the answer to whether $\chi(G) \leq 3$ is almost certainly no. But what more precise statement can we make about $\chi(G)$, and for what range of α? Here, we can use a simple probabilistic approach known as the *first moment method*.

Let the random variable X denote the number of ways of coloring G using exactly 3 colors, so that $\chi(G) \leq 3$ iff $X \geq 1$. Recall Markov's inequality for nonnegative random variables:

$$\mathbb{P}\mathrm{r}[X \geq t] \leq \frac{\mathbb{E}[X]}{t}.$$

Setting $t = 1$ gives

$$\mathbb{P}\mathrm{r}[X \geq 1] \leq \mathbb{E}[X].$$

What can we say about $\mathbb{E}[X]$? Given n vertices, there are 3^n possible color assignments $\mathbf{a}_1, \ldots, \mathbf{a}_{3^n}$. If assignment $\mathbf{a}_i$ has probability q_i of being a valid coloring, then from linearity of expectation,

$$\mathbb{E}[X] = \sum_{i=1}^{3^n} q_i.$$

If we assume that n is divisible by 3, then it is not hard to show that the probability q_i is highest in the case where there are $n/3$ vertices with each color: that is where the fewest pairs of vertices have the same color. So if q is that maximum probability, then

$$\mathbb{P}\mathrm{r}[X \geq 1] \leq \mathbb{E}[X] \leq 3^n q.$$

We approximate q by modifying the random model slightly: instead of placing edges independently with probability $p = \alpha/n$, assume that there are *exactly* $m = \alpha(n-1)/2$ edges, chosen independently and with replacement among all possible pairs of vertices. If a vertex is equally likely to take on any color, the probability that a randomly drawn edge connects vertices of different colors is $2/3$. In that case,

$$q = \left(\frac{2}{3}\right)^m = \left(\frac{2}{3}\right)^{\alpha(n-1)/2},$$

and so

$$\mathbb{P}\mathrm{r}[X \geq 1] \leq \mathbb{E}[X] \leq 3^n \left(\frac{2}{3}\right)^{\alpha(n-1)/2} = \left(\frac{2}{3}\right)^{-\alpha/2} \left[3\left(\frac{2}{3}\right)^{\alpha/2}\right]^n.$$

Consequently, $\mathbb{P}\mathrm{r}[X \geq 1]$ vanishes for large n as long as

$$3\left(\frac{2}{3}\right)^{\alpha/2} < 1,$$

or

$$\alpha > 2\,\frac{\ln 1/3}{\ln 2/3} \approx 5.419.$$

With a little bit of work, one can use probabilistic tail bounds to show that our approximation of q becomes exact for large n, meaning that this bound holds for the $\mathcal{G}(n,p)$ ensemble. So the conclusion is that if G is drawn from $\mathcal{G}(n,p)$ with $p = \alpha/n$, for $\alpha > 5.419$,

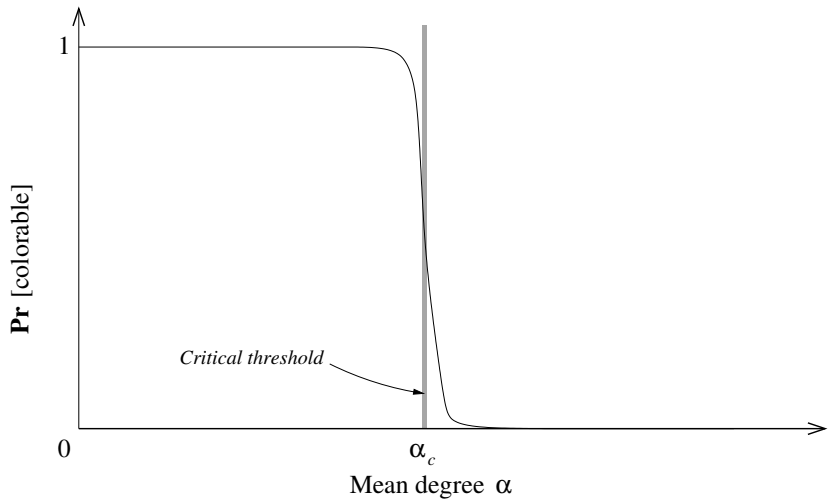
$$\chi(G) > 3 \quad \text{w.h.p.}$$

6.2.3 Phase transition

So for $\alpha < 1$, random graphs from $\mathcal{G}(n, \alpha/n)$ almost certainly are 3-colorable, and for $\alpha > 5.419$, they almost certainly are not. What happens for α in between these two constants? In 1991, Peter Cheeseman, Bob Kanefsky and William Taylor, three artifical intelligence researchers at NASA Ames, conjectured based on numerical simulations that there is a critical value α_c at which a phase transition takes place:

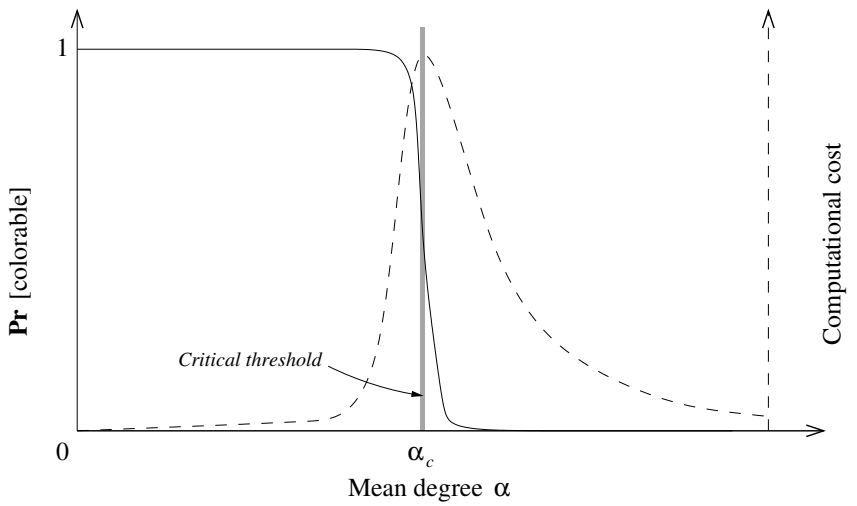
- $\alpha < \alpha_c$ *(subcritical)*. Graphs are 3-colorable with probability approaching 1 for $n \to \infty$.
- $\alpha > \alpha_c$ *(supercritical)*. Graphs are 3-colorable with probability approaching 0 for $n \to \infty$.

That gives rise to the following phase diagram:



This conjecture was *almost* proved in 1999 by Dimitris Achlioptas and Ehud Friedgut. What remains an open problem is showing that α_c is well-defined in the limit $n \to \infty$, although in view of the increasingly tight bounds that researchers have found since that time, it is difficult to imagine otherwise. A possibly more significant open problem is determining the value of α_c. Unlike for the phase transition of the giant component in random graphs, where the critical value is $\alpha = 1$, the exact value of α_c for random graph coloring is not known. Numerically, it is estimated to be around $\alpha_c \approx 4.69$.

In fact, the claim by Cheeseman et al. was not only a structural one, but an algorithmic one as well. Their paper, which had the title "Where the *really* hard problems are," speculated that the phase transition is closely connected with the average-case complexity of the problem. For graphs deep in the "uncolorable" or "colorable" phases, it is easy to answer whether or not $\chi(G) \leq 3$, even though in principle the problem is NP-complete. But near the sharp threshold separating the two phases, it is presumably hard for any algorithm to answer the question. Accompanying the critical value α_c is then a peak in complexity:

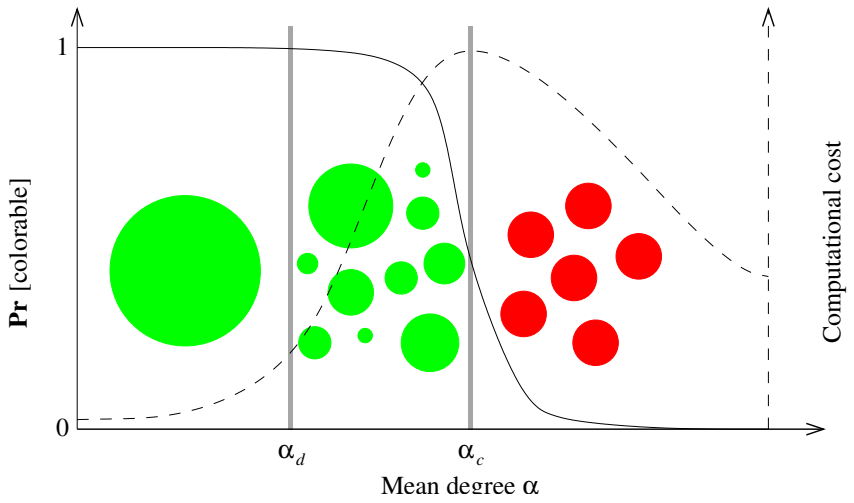


There has been some controversy over how general this effect really is, and whether it really is a statement about the *intrinsic* complexity of the problem rather than simply about the performance of a certain class of algorithms. However, similar effects have been noticed in many other NP-complete problems over random ensembles. The past decade has seen a flurry of research activity in this area. One of the most useful consequences has been a more detailed understanding of the phase transition, and how such an understanding can motivate better algorithms.

To see this, consider one of the optimization problems that we originally asked in the context of graph coloring. How can we color vertices of the graph, using only 3 colors, such that the number of edge violations is *minimized*? (This minimum will generally be zero in the colorable phase.)

Imagine the set of **all** such optimal color assignments for a given graph. This set undergoes an interesting qualitative change as we consider graphs of increasing density. When α is small, not only are there very many colorings possible, but any one coloring is "close" to many others. It is possible to convert one solution into another by changing the color assignment of only a very small number of vertices. Furthermore, there exists a chain of "close" solutions taking us from *any* one coloring to *any* other. We can think of all solutions as lying within a single cluster.

But somewhere within the subcritical phase, this picture changes. Above a certain threshold value α_d which is *strictly below* α_c, the single cluster shatters into many non-adjacent clusters. It is impossible to find a chain of small moves going from a solution in one cluster to a solution in another. Wide gaps develop in the solution space. This persists up to the critical value α_c, where optimal solutions no longer provide perfect colorings.



So in fact, there is not only one phase transition taking place, but two! The solution space structure changes suddenly at $\alpha = \alpha_d$, creating two subphases within the colorable phase, and then the entire nature of solutions changes at $\alpha = \alpha_c$, where the uncolorable phase starts.

But the cluster fragmentation at α_d turns out to have another consequence: a "backbone" starts forming within the graph, consisting of variables that always adopt the same color within a given solution cluster. This has the effect of trapping algorithms such as the depth-first search method that we studied. If all solutions lie within a single cluster, the algorithm is very likely to find one without needing to explore much of the search tree. But if there is a backbone that needs to be set correctly, and depth-first search does not do so at the outset, it will get stuck in an exponential dead end. The fragmentation transition at α_d marks the boundary between an "easy colorable" subphase and a "hard colorable" subphase, where the complexity starts escalating.

Fortunately, the structural picture also suggests a promising alternative to depth-first search: instead of sequentially fixing the colors of vertices, we need to identify the backbone variables and treat them separately. This is the motivation behind recent probabilistic inference methods based on *belief propagation*. Perhaps the most successful is *survey propagation*, proposed in 2002 by Marc Mézard, Giorgio Parisi and Riccardo Zecchina. The idea is to find a self-consistent set of probabilities that a given vertex will adopt a given color, over the set of all optimal colorings. In some cases, this can be done in such a way that the backbone vertices get fixed to their most probable values first, and the other vertices then follow. The result is that even within the "hard colorable" subphase, we can often find the few perfect colorings that exist for a given graph.