The Hardest Constraint Problems: A Double Phase Transition

Tad Hogg Colin Williams

Artificial Intelligence **69**, 359–377 (1994)

Abstract

The distribution of hard graph coloring problems as a function of graph connectivity is shown to have two distinct transition behaviors. The first, previously recognized, is a peak in the median search cost near the connectivity at which half the graphs have solutions. This region contains a high proportion of relatively hard problem instances. However, the hardest instances are in fact concentrated at a second, lower, transition point. Near this point, most problems are quite easy, but there are also a few very hard cases. This region of exceptionally hard problems corresponds to the transition between polynomial and exponential scaling of the average search cost, whose location we also estimate theoretically. These behaviors also appear to arise in other constraint problems. This work also shows the limitations of simple measures of the cost distribution, such as mean or median, for identifying outlying cases.

1 Introduction

A number of recent studies have revealed a relation between the structure of constraint satisfaction problems and the difficulty of solving them with search [3, 12, 16, 4, 9, 17]. Specifically, the median search cost of many sophisticated algorithms for a variety of constraint-satisfaction problems exhibits a sharp peak as a structural parameter is varied. This peak closely coincides with the transition from under- to overconstrained problems, often manifested as an abrupt change in the probability a problem instance has a solution. This knowledge can be applied to find hard cases to test search algorithms and suggest new algorithms particularly suited for hard instances near the transition region [3, 6, 15].

Motivated by observations of rare, but very hard instances substantially outside this transition region, in this paper we investigate the distribution of hard cases in the context of graph coloring. We find that problem instances with sufficiently high search cost are more readily found well below the peak in median cost. Moreover, these cases are so hard that they dominate the mean search cost. This gives rise to a more subtle second transition, from polynomial to exponential average search cost, that occurs below the peak in median cost. The unexpected presence of such hard problems in supposedly easy regions, points to the need to examine the high-cost tail of the distribution of search costs, rather than relying primarily on mean or median costs.

2 Graph Coloring

A graph coloring problem consists of a graph, a specified number of colors, and the requirement to find a color for each node in the graph such that no pair of adjacent nodes (i.e., nodes linked by an edge in the graph) have the same color. Many important A.I. problems, such as planning and scheduling, can be mapped onto the graph coloring problem. Moreover, as a well-known NP-complete problem, graph coloring has received considerable attention and a number of search methods have been developed [11, 8, 15], including specific A.I. techniques that rely heavily on the use of heuristics. Thus graph coloring serves both as an example of a class of hard problems and one that has traditionally received much attention in the A.I. community as a test case for heuristic methods.

For this problem, the average degree of the graph γ (i.e., the average number of edges coming from a node in the graph) is an order parameter that distinguishes relatively easy from harder problems, on average. This is related to the number of edges e and number of nodes n in the graph by $e = \frac{1}{2}\gamma n$. In this paper, we focus on the case of 3–coloring (i.e., when 3 different colors are available), for which the transition between under- and overconstrained problems occurs near [3] $\gamma = 5$.

For simplicity, we consider the ensemble of problems given by random graphs, i.e., taking each graph with the specified number of nodes and connectivity to be equally likely. Similar behavior is seen for more restricted classes, such as for those graphs that have a 3–coloring or where certain "trivial" cases are removed, although the quantitative details change slightly. For example, graphs that have solutions tend to be somewhat easier to search, on average, than general random graphs. Conversely, removing trivial cases (e.g., by restricting consideration to graphs in which all nodes have at least 3 edges or also satisfy additional requirements [3]) increases the search cost, and can also slightly shift the transition point. Our focus on the simplest ensemble of graphs connects with the extensive literature on the properties of random graphs [2], while retaining the key behaviors seen in the more complex ensembles.

In the experiments presented below, unless otherwise stated, we used a complete, depth-first backtracking search based on the Brelaz heuristic [8] which assigns the most constrained nodes first (i.e., those with the most distinctly colored neighbors), breaking ties by choosing nodes with the most uncolored neighbors (with any remaining ties broken randomly). For each node, the smallest color consistent with the previous assignments is chosen first, with successive choices made when the search is forced to backtrack. We measure the search cost by the number of states in the search tree that are expanded until the first solution is found or, when there are no solutions, until no further possibilities remain to be examined. As a simple optimization, we never change the colorings for the first two nodes selected by this heuristic. Any such changes, which could only occur when the backtrack search has failed to find a solution starting from the initial assignments for the first two nodes, would amount to unnecessarily repeating the search with a permutation of the colors.

3 Distribution of Cost to First Solution

Fig. 1 shows the previously reported [3] peak in the median search cost. The observed peak for random graphs[1] is at $\gamma = 4.6$. This peak becomes sharper as larger graphs are considered and occurs at the connectivity at which the fraction of graphs with a solution drops from near one to near zero. This observation associates a region with a high density of relatively hard problems for particular search algorithms with an abrupt transition in the nature of the problems themselves.

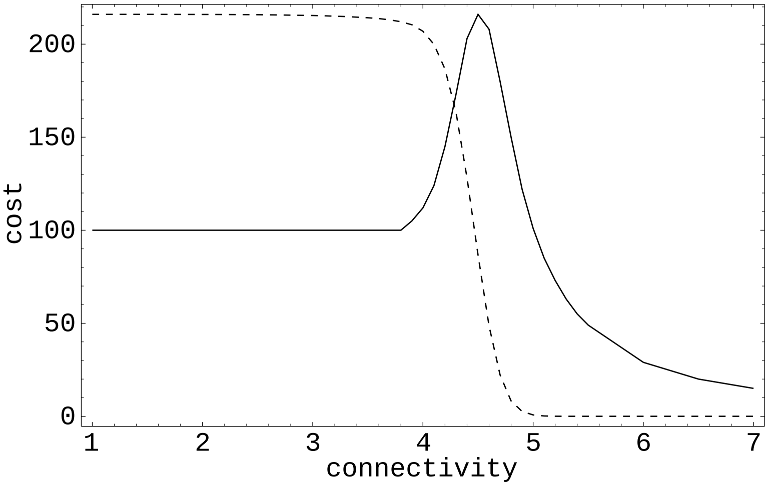

Figure 1: Behavior for 3–coloring of random graphs with 100 nodes as a function of connectivity γ in steps of 0.1. The solid curve shows the median search cost, and the dashed one is the fraction of graphs with a solution (ranging from one on the left to zero on the right).

However, a more complex story is seen from the full distribution of the cost to first solution or failure. Surprisingly, as shown in Fig. 2, this distribution develops a long tail for intermediate values of connectivity,

[1] Note this is slightly lower than the value of 5.1 reported for graphs not reducible with respect to a variety of local simplification operators [3].

so that exceptionally hard instances are concentrated not around the peak in the median, but rather at lower connectivities. Thus, for 100–node graphs, $\gamma = 4.5$, near the median peak, gives many more cases with cost above 1000 than $\gamma = 3$, but the reverse is true for costs above 100,000. Furthermore, the nearly straight-line behavior of the tail for $\gamma = 3$ on the log-log plot of Fig. 2 shows that the fraction of searches, f, whose cost is greater than C, decays as a power of the cost, i.e., $f \propto C^{-x}$, over at least a few orders of magnitude. This compares with an exponentially decaying tail for higher connectivity cases.

These fundamentally different behaviors for the tail of the cost distribution suggest, in contrast to previous studies, that there are actually two qualitatively distinct regions of hard problems: 1) a region containing a high density of relatively hard problems giving the peak in the median, and 2) a region, with somewhat lower connectivity, in which most problems are very easy but which also contains a few exceptionally hard instances.

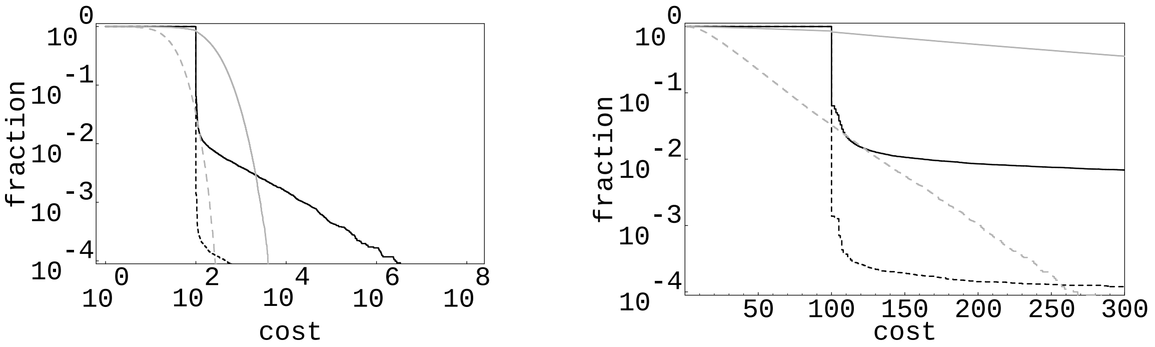

Figure 2: Distribution of search costs for 100–node graphs: fraction of searches with greater than the specified cost. The curves correspond to different connectivities: 2 (dashed), 3 (solid black), 4.5 (gray) and 6.0 (gray, dashed). On the left is a log-log plot, where linear behavior corresponds to a power-law function, and on the right is a log plot, over a limited range of the same data, where linear behavior corresponds to an exponential function. Each curve represents 10000 samples. The long tail behavior for intermediate connectivities is also seen when the samples are restricted to cases with solutions, although it is less extreme.

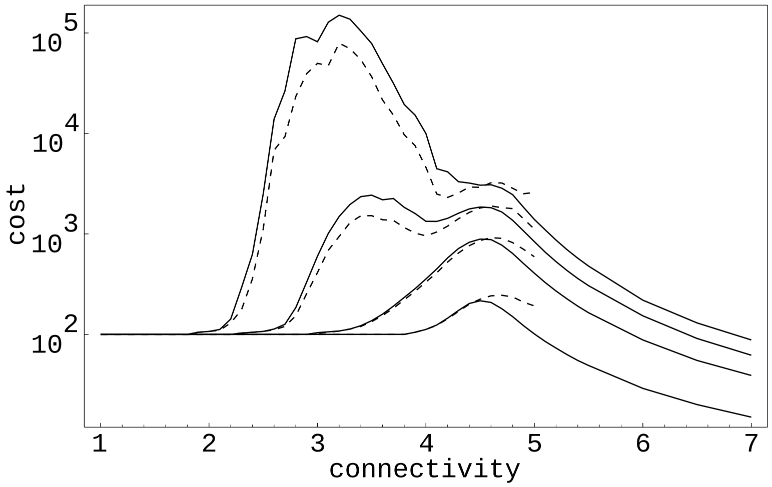

Figure 3: Cost percentiles vs. connectivity for 100–node graphs, based on 50000 samples at each value of γ, given in increments of 0.1. Two sets of curves are shown: solid, for the behavior of all samples, and dashed, for those samples with solutions. The dashed curves extend only up to $\gamma = 5$ since beyond that point few instances have solutions. For each set of curves, the lowest shows the 50% cost (i.e., the median). Successively higher curves show the costs for the top 0.05, 0.005 and 0.0005 of the problems.

The changing nature of the tail of the cost distribution as a function of connectivity can be seen more

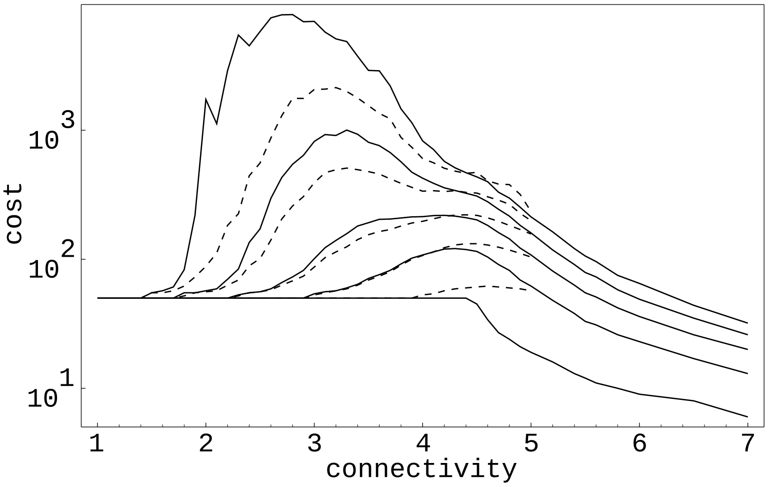

Figure 4: Cost percentiles vs. connectivity for 50–node graphs. The set of solid curves include all instances, and dashed curves are for cases with solutions. For each set, the curves, from bottom to top, show the costs for the top 0.5 (median), 0.05, 0.005, 0.0005 and 0.00005 of the problems, respectively, based on 10^6 samples, and shown in increments of 0.1 in γ.

clearly in Fig. 3 and 4. They show the percentiles of the distribution, i.e., the cost above which are a specified fraction of the problems. In these figures, closely spaced percentile curves correspond to a rapidly decaying tail, while the widely separated ones indicate an extended tail.

In addition to showing the contrast in the nature of the tails of the distributions, these figures show that, like the median cost (which is the 50% percentile), the percentiles exhibit peaks as connectivity is varied. Because of the extended tail at intermediate connectivities, the peak shifts downward as successively higher percentiles of solution cost are considered. Thus problems with costs substantially above that achieved by the median peak are more readily found at a lower value of connectivity than the peak in the median cost. Examining this behavior for graphs of different sizes we see that the peaks appear to sharpen as larger graphs are considered, although the large fluctuations make this difficult to determine precisely. Moreover, the shift is also seen for cases with solutions, so the hard cases for intermediate connectivities and the extended tail are not entirely due to those rare instances with no solutions.

4 The Long Tail of the Cost Distribution

The backtrack-based search algorithms we examined give the same qualitative form for the cumulative cost distribution for intermediate connectivities. As illustrated in Fig. 2, this distribution consists of an abrupt step, indicating that most instances are rapidly solved, followed by a long tail. This tail is approximately described by a power-law with three key parameters:

- its gradient, $-x$
- its intercept with the step, $f_{\rm step}$ and
- its length, ℓ

In practice, the tail of cost distribution for intermediate connectivities cannot follow a power-law out to arbitrarily large costs since there is a finite maximum search cost for any given n. That is, $f(C)$, the fraction of searches, f, whose cost is greater than C, is strictly zero above this maximum. The mathematical interpretation of these observations is that the tail eventually decays much faster than a power-law. However, to a first approximation, the tail consists of a long power-law decay described by[2] $f = f_{\rm step} n^x C^{-x}$ for values of C ranging from slightly above n to some large value ℓ beyond which the distribution drops off more rapidly.

[2] derived from $\ln f = -x \ln C + \beta$ with the special case $\ln f_{\rm step} = -x \ln n + \beta$ and β is a constant.

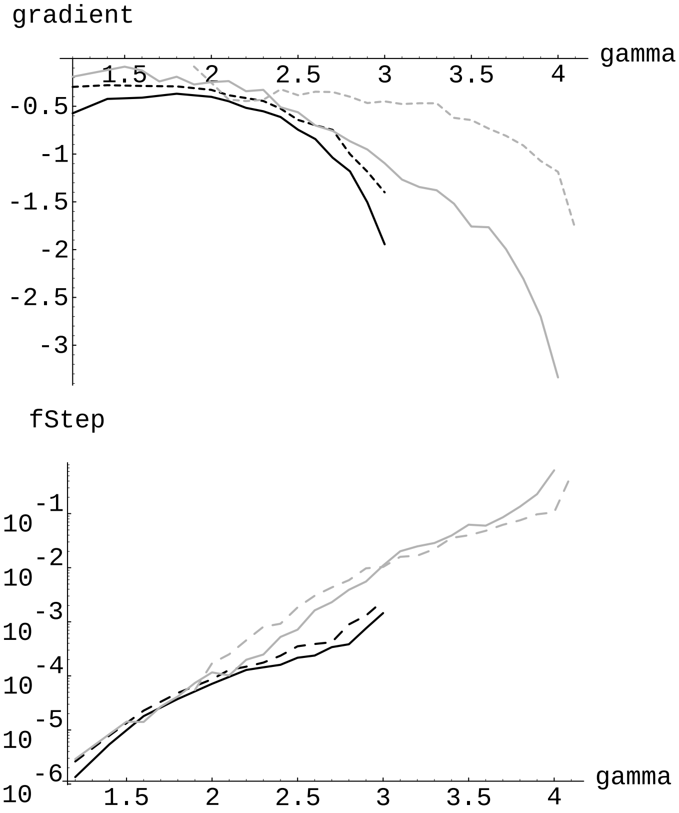

Figure 5: Experimental results at $n = 20$ (dark, solid), $n = 30$ (dark, dashed), $n = 50$ (light, solid) and $n = 100$ (light, dashed) showing how $-x$ and f_{step} vary with γ and n.

The quantitative behavior of the parameters describing the tail change as n is increased as shown in Fig. 5 and 6. From Fig. 5 we can infer some features of the gradient and the intercept with the step, namely:

- *Gradient*: For intermediate connectivities (i.e., $1.2 < \gamma < 2.7$) the power-law tail is very flat (i.e., $0 < x < 1$) but eventually (when $\gamma \geq 2.7$ for $n = 20$ and $n = 30$) the gradient of the power-law tail again becomes much steeper ($x > 1$). The connectivity out to which this flat region extends becomes progressively greater for larger values of n.

- *Intercept with Step*: f_{step} grows exponentially with γ until a connectivity is reached beyond which the power-law tail disappears. The growth rate is insensitive to n.

Notice that $-x$ and f_{step} are fairly insensitive to n.

Next we explored what happened to the length of the tail, at a given connectivity, as larger graphs were considered. Fig. 2 shows that such tails certainly appear at connectivities in the range $2 \leq \gamma \leq 3$. Consequently, we focussed on what happened to the length of the tail at $\gamma = 2.5$ as the number of nodes, n, was increased. The results of numerical experiments are shown in Fig. 6.

The data suggests that:

- *Tail Length*: increases rapidly with n.

In practice it is extremely difficult to sample the complete tail at the higher n values because billions of data points, each a complete search, are needed.

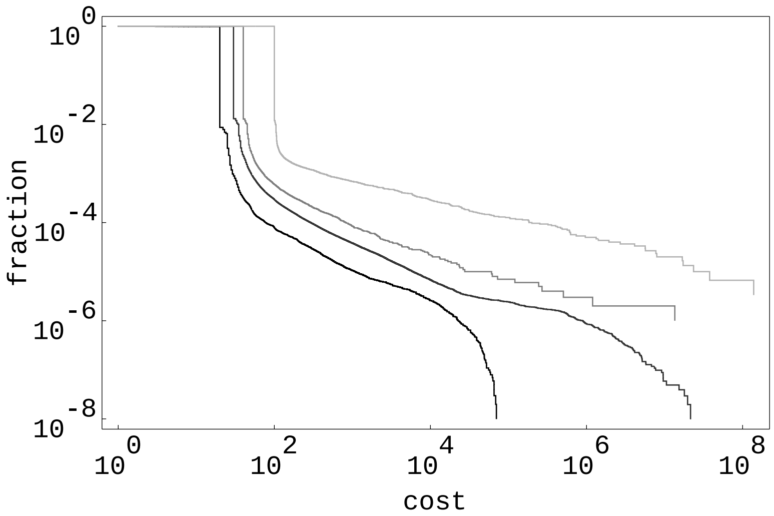

Figure 6: Full cost distribution curves showing explicitly how the tail of the distribution becomes flatter and longer as n is increased at constant γ. The data shown is for $n = 20, 30, 40, 100$ (dark to light) at $\gamma = 2.5$ but similar behavior is seen elsewhere.

5 Consequence of Long Tails for Mean Cost

When describing the typical behavior of searches, often the mean or median costs are used. For tightly clustered distributions these are similar and convey the overall behavior. However, when there is an extended power-law tail, such as those shown Fig. 2 and 6, the mean cost and median cost can be very different. In particular, the long tails cause rare, high-cost cases to contribute significantly to the mean.

To see this consequence of the long tails more precisely, recall that the mean cost is given by $\langle C \rangle = -\int_0^\infty C \frac{df}{dC} dC$. As a specific example, suppose the tail is *exactly* given by a power-law as described above. In other words, we approximate the exact distribution by assuming that $f(C) = 0$ for all costs, C such that $C > n + \ell$. In this case we have:

$$\begin{aligned} \langle C \rangle &= n(1 - f_{\text{step}}) + f_{\text{step}} x n^x \int_n^{n+\ell} C^{-x} dC \\ &\approx n + f_{\text{step}} \frac{x n^x}{1-x} \left((n+\ell)^{1-x} - n^{1-x} \right) \end{aligned} \qquad (1)$$

where we used $1 - f_{\text{step}} \approx 1$ based on the observations that $f_{\text{step}} \ll 1$ but is fairly insensitive to changes in n. Moreover, we observe that $\ell \gg n$ and so $(n + \ell)^{1-x} \approx \ell^{1-x}$. Hence the behavior of $\langle C \rangle$ is particularly sensitive to whether x is greater than or less than 1. Thus, when $x > 1$ (a rapidly decaying power-law tail), $\ell^{1-x} \ll n^{1-x} \to 0$ giving $\langle C \rangle \approx n$. In this case, the rare cases in the tail do not contribute significantly to the mean. By contrast, when $x < 1$ (a slowly decaying tail), $\ell^{1-x} \gg n^{1-x}$ which dominates the mean for large n giving

$$\langle C \rangle \approx f_{\text{step}} \frac{x n^x \ell^{1-x}}{1-x} \gg n$$

In this case the mean cost is dominated by the rare cases in the tail of the distribution. Thus we see that the value of x, the gradient of the power-law tail, determines whether the anomalously high costs contribute significantly to the mean.

With this insight, we now turn to the behavior of the mean cost for the observed distributions. As shown in Fig. 5, for intermediate connectivities, x is indeed less than one, and both x and f_{step} are approximately independent of n. Thus the behavior of the mean is dominated by the rare cases in the tail, and is quite large, as shown in Fig. 7. Moreover, accurate empirical estimates of the mean in this region require an enormous number of samples to cover the significant part of the tail (i.e., enough of the tail to include the full region where its gradient satisfies $x < 1$ and the truncated end is becoming apparent). More limited sampling is likely to substantially underestimate the true mean, a common difficulty with such extended distributions [14]. Moreover, from Fig. 7 we also see that the mean appears to develop two peaks. In light of the behavior of x and f_{step} this means that, as γ increases beyond 3, the length of the power-law tail decreases, consistent with the observations in Fig. 2.

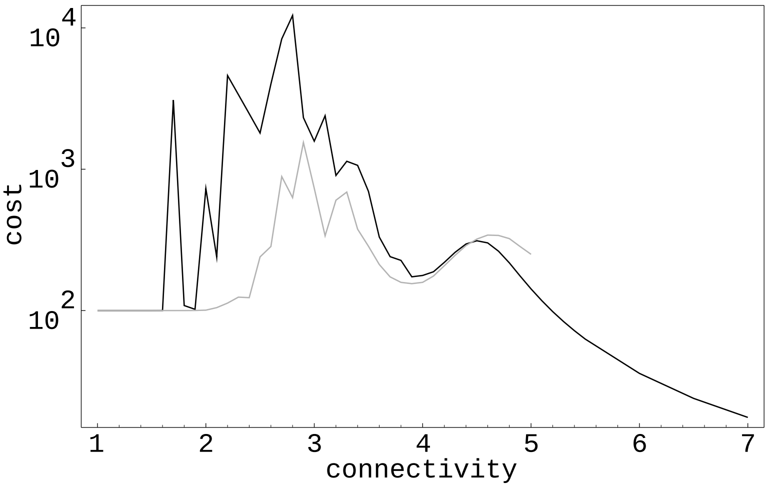

Figure 7: Behavior for 3–coloring of random graphs with 100 nodes vs. γ in increments of 0.1. The black curve shows the average search cost for 50000 samples. The gray curve is the average for those graphs that have a solution. There is extremely large variance for the intermediate connectivities giving a relatively large error to the estimate of the mean in that region.

By contrast, the behavior of the median is determined by the size of the step before the power-law tail. When $f_{\text{step}} < \frac{1}{2}$, the large cost instances will not contribute to the median cost, which will be equal to n. This is in fact the case at intermediate connectivities since we observe $f_{\text{step}} \ll 1$, even as high as $\gamma = 3$. Hence we find substantial difference between mean and median in this regime as seen by comparing Fig. 1 and 7.

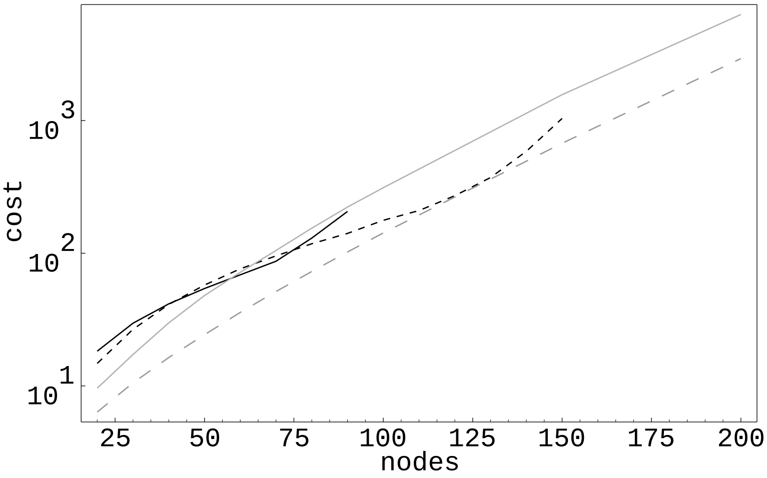

Figure 8: Behavior of average cost vs. number of nodes in the graph, with a logarithmic scale. Shown is the scaling for connectivities 3.5, 4.0, 4.5 and 5.0 (respectively, solid black, dashed, gray and gray dashed curves). Each point is the mean cost of 10^6 samples (up to 50 nodes) or 10^5 (above 50 nodes).

A final question concerns how the mean scales with increasing n. From the above discussion, the mean grows rapidly when dominated by the hard cases in the tail. In fact, as shown in Fig. 8 this growth is approximately exponential, at least for the higher connectivities. Fluctuations preclude a definitive conclusion from a limited number of samples in the range $1 < \gamma < 3.5$. For smaller values of the connectivity, we obtain data consistent with a linear scaling, but there remains the possibility of seeing more rapid growth as more samples are considered.

6 A Phase Transition Producing the Long Tail

Our observations show that at intermediate connectivities rare, hard instances dominate the mean cost and make it grow exponentially. This raises the issue of whether this behavior persists to arbitrarily small values of γ or it disappears below a specific, nonzero value γ_{crit}. This question is difficult to address empirically due to the large number of samples required to adequately examine the tail of the distribution for smaller values of γ. We consider instead the topological structure of random graphs to show that there is indeed a regime in which the average cost grows linearly.

A simple argument is based on the fact that, for connectivities below one, as the number of nodes increases, almost all random graphs consist of either trees or components with a single cycle [2] (which we refer to as *trivial* components). These types of components, characterized by having no more edges than nodes, can be colored without backtracking by the Brelaz heuristic (provided there are at least three available colors, i.e., $b \geq 3$), so the search cost for almost all graphs will be just the number of nodes n. Naively, this suggests that the average search cost will scale linearly at least up to $\gamma = 1$, at which point more complex components become common. However, this is not definitive since the harder components, although rare, could still contribute significantly to the mean. For example, suppose there is a difficult component of size n that requires an exponentially large search cost, e.g., 2^n, but whose probability of existing decreases only as a power of n, e.g., $1/n$, (which would satisfy the requirement that the probability of such components goes to zero for $\gamma < 1$). Such a component would contribute significantly to the mean, e.g., $\frac{1}{n}2^n$ in this case. Thus it is important to recognize not only that such components are rare, but that they are sufficiently rare to avoid any significant contribution to the mean.

6.1 A Polynomial to Exponential Transition

To make this argument more precise we introduce a simple bound on the search cost and show that, for sufficiently small γ the mean cost indeed grows linearly. Specifically, the overall search cost is determined by the cost of searching each connected component of the graph. For trivial components this can be done without backtracking. For others, we use the very crude upper bound of $(b-1)^k$, where k is the number of nodes in the component. This bound arises by noting that, within a connected graph, the Brelaz heuristic always selects the next node to color from among those that already have colored neighbors. Hence for each such choice there will be at most $b-1$ available color choices. The cost to search the whole graph can be bounded by the sum of the costs to search each component[3]. For graphs with n nodes and $m = \frac{1}{2}\gamma n$ edges, we obtain a bound on the mean cost:

$$\langle C \rangle \leq n + \sum_{k=1}^{m-1} (b-1)^k \sum_{e=k+1}^{m} E_{ke} \tag{2}$$

where E_{ke} is the expected number of components with k nodes and e edges.

The expected number of components with k nodes and e edges is given by

$$E_{ke} = \frac{\binom{n}{k} c_{ke} \binom{\binom{n-k}{2}}{m-e}}{\binom{\binom{n}{2}}{m}} \tag{3}$$

where c_{ke} is the number of connected graphs with k nodes and e edges. In this expression, the first two factors in the numerator together count the number of ways to form a connected subgraph with the specified number of nodes and edges. Such a subgraph will be a component of the overall graph provided no edges link it to any of the remaining $n-k$ nodes in the graph. The third factor in the numerator counts the number of ways this can be done. Finally, the denominator is just the total number of graphs with n nodes and m

[3] When there are no solutions, the search could terminate as soon as an unsolvable component is found, giving a smaller cost.

edges. This can be evaluated explicitly using a recursion relation for the number of connected graphs [13], but for our case the following simple bound is sufficient. The connected graphs with k nodes and $k+i$ edges can be generated by adding a single extra edge, in all possible ways, to the connected graphs with k nodes and $k+i-1$ edges, although this procedure can generate the graphs multiple times and so produces an upper bound on the number of such graphs. Thus $c_{k,k+i} \leq \frac{1}{2}k^2 c_{k,k+i-1}$ because $\frac{1}{2}k^2$ is an upper bound for the number of additional possible edges for a graph with k nodes. Finally since $c_{k,k-1}$ is just the number of trees with k nodes, i.e., k^{k-2} we have $c_{k,k+i} \leq k^{k+2i}/2^{i+1}$. When there are many edges, a better bound is just the total number of graphs:

$$c_{k,k+i} \leq \binom{\binom{k}{2}}{k+i}$$

In the remainder of this discussion we consider the behavior of $E_{k,k+i}$ as determined by the smallest of these two bounds.

These bounds and Stirling's formula [1]

$$1 < \frac{x!}{x^x e^{-x}\sqrt{2\pi x}} < e^{1/12x}$$

determine the behavior of $E_{k,k+i}$ as a function of i. We find that, when $\gamma < 1$ and $n \to \infty$, there are two types of behavior. The first, for $k < \sqrt{2n/\gamma}$, has $E_{k,k+i}$ monotonically decreasing with i so the maximum is at $i_{\max} = 1$. For larger k, the maximum occurs at $i_{\max} \sim (1-\ln 2)k/\ln k$. We then use the simple bound $\sum_{e=k+1}^{m} E_{ke} < m E_{k,k+i_{\max}}$ in Eq. (2) to obtain

$$\begin{aligned}
\langle C \rangle &< n + \frac{1}{4}m \sum_{k=1}^{m-1} (b-1)^k k^{k+1} R_{nmk} I_{nmk} \\
R_{nmk} &\equiv \binom{n}{k}\binom{\binom{n-k}{2}}{m-(k+1)} \Big/ \binom{\binom{n}{2}}{m} \\
I_{nmk} &\equiv E_{k,k+i_{max}}/E_{k,k+1}
\end{aligned} \tag{4}$$

Note that $I_{nmk} = 1$ for $k < \sqrt{2n/\gamma}$. Stirling's formula gives $\ln R_{nmk} < r_{n\gamma k}$ where

$$r_{n\gamma k} \equiv k(-\gamma + \ln\gamma) + \ln\gamma - \ln n - \ln\left(\sqrt{2\pi k}k^k e^{-k}\right) \tag{5}$$

The increased contribution from I_{nmk} for $k > \sqrt{2n/\gamma}$ introduces a factor which is less than $e^{k(1-\ln 2)}$. Combining these results gives

$$\langle C \rangle < n + \frac{\frac{1}{8}\gamma^2}{\sqrt{2\pi}} \sum_{k=1}^{m} k^{3/2} e^{k(2-\ln 2-\gamma+\ln(b-1)+\ln\gamma)} \tag{6}$$

As n increases, this sum approaches a finite constant value provided $-\gamma + \ln(b-1) + \ln\gamma + 2 - \ln 2 < 0$ since the ratio test shows the sum converges in that case. Hence the root of this equation $\gamma_{\text{crit}} = 0.16$, marks the connectivity below which the cost of graph coloring is surely linear. Note that this is only a lower bound on the extent of the linear regime since many of the nontrivial components could be much easier to search than implied by our crude estimate of $(b-1)^k$. Nevertheless it is sufficient to conclude that long tails cannot persist down to arbitrarily small γ.

We should note that attempts to improve this bound on the mean cost must focus on the behavior of the exponent in the sum, since the result for γ_{crit} is not affected by improvements in the bound by polynomial factors in k. This can be approached in three ways. First, a tighter bound on the number of connected graphs may eliminate the contribution from components with many edges. In particular, numerical evaluation of Eq. (2) suggests it remains linear up to about $\gamma = 0.23$. Second, we can improve our general search bound of $(b-1)^k$ for nontrivial components by accounting to some extent for the fact that graphs should get easier to color when a large number of colors is available, and that when there are many edges in the graph, increased pruning will limit the depth of backtrack. Third, we can use a better understanding of the relation of components' topological structure to search cost. For instance, there are additional types of components that can be searched without backtrack, e.g., a component consisting of non-overlapping cycles. These need not be counted among the nontrivial components used in our cost bound. Furthermore, for those components that do generally require backtracking, only those nodes that are actually part of a cycle in the component could cause some backtracking, for dependency directed methods. For example, for a component of size k in which only $\frac{1}{2}k$ nodes were actually part of cycles, and the rest were in trees connected to these cycles, we could use the search cost bound of $k + (b-1)^{k/2}$ instead of $(b-1)^k$. These improvements in the component search bound can result in a smaller exponent in the sum, and hence a larger value for γ_{crit}.

6.2 Location of the Polynomial/Exponential Transition Point

Associating the connectivity value γ_{crit} above which the cost distribution develops a long tail with a phase transition provides a way to compute its location theoretically. Specifically, it can be approximated using a model that relates the mean search cost to the behavior of the number of partial solutions of various sizes [18]. In this context, a partial solution of size s is an assignments of colors to s nodes so that no constraint is violated. The search cost is dominated by a bulge, at $s = s_{\max}$, in the number of partial solutions as a function of their size. When this bulge occurs at a size smaller than that of complete solutions, i.e., $s_{\max} < n$, the theory predicts that cost grows exponentially. However, when there are few constraints (e.g., sparse graphs for graph coloring), this bulge will be at the solution level itself, i.e., $s_{\max} = n$ and the predicted cost grows only polynomially. Identifying the largest value of connectivity, γ, for which the bulge is still at the solution level, then allows a prediction of the transition point. For the best quantitative prediction, the model can be specialized to the particular constraints of graph coloring [18, 17], and this then predicts the transition will occur at $\gamma = (b-1)\ln b$ where b is the number of colors used in the problem.[4] Thus for the 3–coloring searches considered here this predicts $\gamma = 2.2$. By contrast, this theory predicts the peak in the median cost (the transition from under- to overconstrained problems) at [18, 17]

$$\gamma = -\frac{2\ln b}{\ln(1-1/b)}$$

or 5.4 for 3–coloring[5]. Thus this theory makes the qualitative prediction that the transition from polynomial to exponential behavior is separate from and below the transition in the probability for a solution.

7 Cost Distributions for Other Search Algorithms

An important result of previous studies is that the peak in the median search cost is seen using a variety of search methods, though at slightly different locations. This establishes the generality of the observation, and gives it much more significance than if it were limited to a particular heuristic search method. To address this issue for the behavior of the full cost distribution, in this section we present the behavior for additional A.I. search methods.

7.1 Alternative 1: Heuristic Repair

As one very different search method, we used heuristic repair [11] from randomly selected initial colorings of all the nodes. At each step, this method selects a node whose current color conflicts with that of a neighbor and changes that node's color to reduce as much as possible the number of violated constraints in the problem. If the total number of violations is not reduced within a prespecified number of steps, the search restarts from a new initial coloring. This never terminates for problems with no solutions, so we applied it only to cases with solutions. Our cost measure is just the number of states examined during the search.

Fig. 9 shows the behavior of heuristic repair on 50–node graphs with solutions. By comparison with Fig. 4 we see that heuristic repair typically requires substantially more search steps than the Brelaz heuristic, and exhibits a single, broad peak, which shifts slightly to lower values of connectivity as higher percentiles are

[4] The basic model [18, Eq. 8] gives the polynomial to exponential transition at $\gamma = (b - \frac{1}{b})\ln b$ or 2.9 for 3–coloring.

[5] The prediction of 5.4 is quite close to the observed transition for reduced graphs [3] but somewhat higher than the value for our ensemble of random graphs.

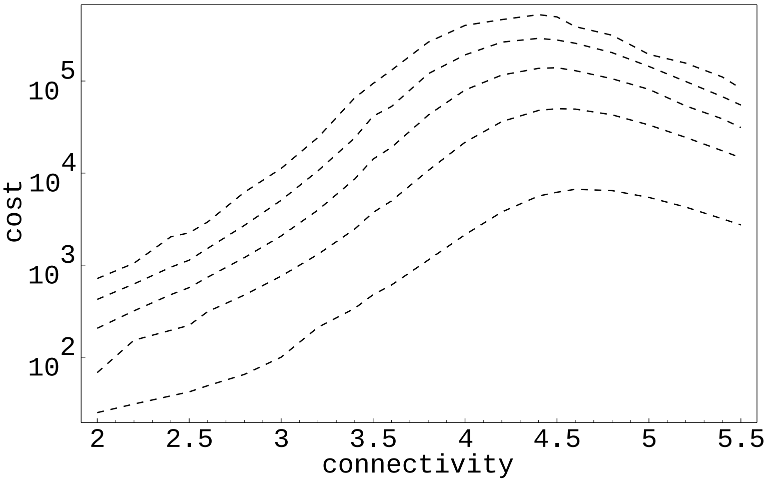

Figure 9: Cost percentiles vs. connectivity for 50–node graphs, with solutions, using heuristic repair search. The curves show the same percentiles as in Fig. 4 for γ in increments of 0.2. The heuristic repair search was restarted whenever 100 successive steps gave no reduction in the number of conflicts, and this was repeated until a solution was found. Qualitatively similar behavior is seen with a 1000–step limit.

considered. While it is possible that harder instances at low connectivity could exist at higher percentiles, or for larger sized graphs, this observation suggests that heuristic repair is not sensitive to the same structural properties of the graphs as the Brelaz algorithm.

One factor that contributes to this difference is that the heuristic repair search is restarted from a new initial state whenever it fails to reduce the number of constraint violations after a limited number of steps. This prevents the search from continuing with potentially lengthy searches due to a poor choice of initial state. Thus graphs will have relatively high cost for this method only to the extent that good initial states are rare. By contrast, a backtracking search method is forced to explore a large search space whenever its first few choices preclude any solution but that fact cannot be determined until many additional choices have been made. We found, however, that this only accounts for part of the difference in behavior between these methods: restarting the backtrack search after a prespecified number of unsuccessful steps eliminates some of the exceptionally hard cases but not all of them [7].

Finally, we should note that the exceptionally hard instances for Brelaz at low connectivities are not especially easy for heuristic repair. This was established by considering the behavior of the minimum cost of the two methods, and seeing that the second peak, at low connectivities, remained.

7.2 Alternative 2: Local Reduction

Many graphs, particularly with low connectivities, can be substantially reduced by eliminating from the search nodes that are trivial to color. For example, any node with two or fewer links can always be colored since at most two of the three available colors are eliminated by the neighbor colorings. This suggests another search method: first reduce the graph as much as possible using a simple set of three reduction operations [3], and then search the remaining reduced graph with Brelaz backtrack. The total search cost must now include some measure of the reduction operators as well as that of the Brelaz search. For simplicity, we counted the number of times the reduction operators were applied, plus the number of Brelaz search states, plus one step to count the search-free coloring of eliminated nodes. Each of these operations involves examining the local neighborhood of many nodes in the graph so are of roughly comparable magnitude[6]. The results, shown in Fig. 10, show the extended tails for intermediate connectivities and give the same qualitative shift in the peak for higher percentiles.

[6] Note that measurements based on actual execution time would give somewhat different weights to the different operations.

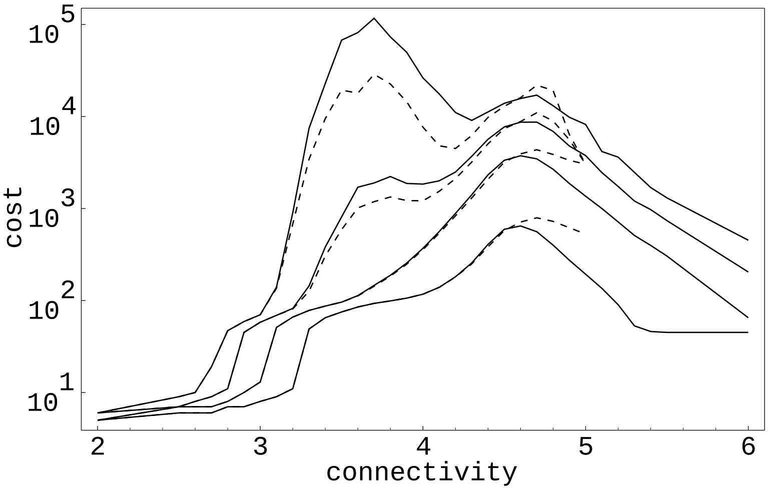

Figure 10: Cost percentiles vs. connectivity for 150–node graphs using graph reduction followed by Brelaz backtrack on the reduced graph. The curves, based on 10^5 samples, show the same percentiles as in Fig. 3 for γ in increments of 0.1.

8 Discussion

In summary, we have shown that the distribution of hard graph coloring problems is more subtle than previously reported. In particular, the hardest search examples are concentrated below the peak in the median cost, and appear to be associated with a transition between polynomial and exponential average search cost. We also derived an approximate analytic expression for the location of this transition. A full analysis of this behavior remains for future work.

A practical consequence of our observations on the complete cost distribution of solving graph coloring problems, is to provide information about where to look for intrinsically hard problem instances, i.e., problems which are hard for many dissimilar A.I. algorithms. Such problems might be sought after to serve as test cases for benchmarking new algorithms. To give a specific example, if one wanted to find a 3-coloring problem for a 100 node graph with a search cost in excess of 1 million steps, one would be better off looking at problems at $\gamma = 3$ (roughly where the mean peaks) than at problems at $\gamma = 4.5$ (roughly where the median peaks).

Moreover, the fact that such instances are rare in the ensemble of problems we have investigated (the random graph ensemble) should not be taken to imply that they will be rare in *all* ensembles. For example, graphs corresponding to class scheduling problems are quite unlike the distribution of random graphs due to implicit correlations between the course preferences of students [10]. Consequently, in practice, it is entirely possible that the kinds of troublesome problems we have identified might be encountered more frequently than the random graph ensemble would suggest.

An interesting open question is whether these hard, sparse graphs have qualitatively different structure than the more densely connected cases associated with the peak in the median. If so, this may suggest specialized heuristics for use in these cases. Alternatively, by examining the detailed topological structure of exceptionally hard low connectivity problems, which are stripped of extraneous topological details, we might be able to spot what topological features make such problems truly hard and use the insight gained to shed light on the higher connectivity cases. Exceptionally hard problems are therefore a natural topic of study for A.I. because they represent the most challenging cases that Nature supplies.

There is also the question of the generality of the behavior seen here. As one way to address this, we have also examined the behavior of other graph ensembles. These include restricting attention to those cases with solutions, graphs with a prespecified solution, and various “reduced” graphs in which some trivial cases are removed by simple local operations. All these cases show the same qualitative behavior, i.e., a concentration of hard problems below the peak in the median due to rare hard cases, but with different absolute cost values and somewhat shifted peak locations.

Finally we should note that indications of a second peak in the mean search cost, the existence of rare, exceptionally hard cases below the transition in the median cost, and the transition between polynomial and exponential search cost, have been observed in the satisfiability problem [4, 5]. The second peak is also evident in constraint satisfaction models based on random selection of minimized nogoods [18]. These observations, as well as the generality of the theoretical arguments given above, suggest that the behavior observed here for graph coloring occurs in a wide range of constraint problems. An interesting open question is posed by other kinds of problems such as optimization or satisficing searches (in which one desires high quality solutions within limited resources). Like constraint satisfaction, these problems are known to exhibit complexity transitions, such as the transition from polynomial to exponential cost [19] and a peak in median cost for some cases [3]. Thus it is of interest to see if they also exhibit a separate region of exceptionally hard problems.

9 Acknowledgments

We thank Mauricio Lomelin and Anant Sahai for helpful discussions and improvements to the search code used in these experiments.

References

[1] M. Abramowitz and I. Stegun, editors. *Handbook of Mathematical Functions.* Dover, New York, 1965.

[2] B. Bollobas. *Random Graphs.* Academic Press, NY, 1985.

[3] Peter Cheeseman, Bob Kanefsky, and William M. Taylor. Where the really hard problems are. In J. Mylopoulos and R. Reiter, editors, *Proceedings of IJCAI91*, pages 331–337, San Mateo, CA, 1991. Morgan Kaufmann.

[4] James M. Crawford and Larry D. Auton. Experimental results on the cross-over point in satisfiability problems. In *Proc. of the 11th Natl. Conf. on Artificial Intelligence (AAAI93)*, pages 21–27, Menlo Park, CA, 1993. AAAI Press.

[5] Ian P. Gent and Toby Walsh. Easy problems are sometimes hard. *Artificial Intelligence*, 70:335–345, 1994.

[6] Tad Hogg and Colin P. Williams. Solving the really hard problems with cooperative search. In *Proc. of the 11th Natl. Conf. on Artificial Intelligence (AAAI93)*, pages 231–236, Menlo Park, CA, 1993. AAAI Press.

[7] Tad Hogg and Colin P. Williams. Expected gains from parallelizing constraint solving for hard problems. In *Proc. of the 12th Natl. Conf. on Artificial Intelligence (AAAI94)*, pages 331–336, Menlo Park, CA, 1994. AAAI Press.

[8] David S. Johnson, Cecilia R. Aragon, Lyle A. McGeoch, and Catherine Schevon. Optimization by simulated annealing: An experimental evaluation; part ii, graph coloring and number partitioning. *Operations Research*, 39(3):378–406, May-June 1991.

[9] Tracy Larrabee and Yumi Tsuji. Evidence for a satisfiability threshold for random 3CNF formulas. In Haym Hirsh et al., editors, *AAAI Spring Symposium on AI and NP-Hard Problems*, pages 112–118. AAAI, 1993.

[10] Gary Lewandowski and Anne Condon. Experiments with parallel graph coloring heuristics. In *Proc. of 2nd DIMACS Challenge*, 1993.

[11] Steven Minton, Mark D. Johnston, Andrew B. Philips, and Philip Laird. Solving large-scale constraint satisfaction and scheduling problems using a heuristic repair method. In *Proceedings of AAAI-90*, pages 17–24, Menlo Park, CA, 1990. AAAI Press.

[12] David Mitchell, Bart Selman, and Hector Levesque. Hard and easy distributions of SAT problems. In *Proc. of the 10th Natl. Conf. on Artificial Intelligence (AAAI92)*, pages 459–465, Menlo Park, 1992. AAAI Press.

[13] E. M. Palmer. *Graphical Evolution: An Introduction to the Theory of Random Graphs.* Wiley Interscience, NY, 1985.

[14] S. Redner. Random multiplicative processes: An elementary tutorial. *Am. J. Phys.*, 58(3):267–273, March 1990.

[15] Bart Selman, Hector Levesque, and David Mitchell. A new method for solving hard satisfiability problems. In *Proc. of the 10th Natl. Conf. on Artificial Intelligence (AAAI92)*, pages 440–446, Menlo Park, CA, 1992. AAAI Press.

[16] Colin P. Williams and Tad Hogg. Using deep structure to locate hard problems. In *Proc. of the 10th Natl. Conf. on Artificial Intelligence (AAAI92)*, pages 472–477, Menlo Park, CA, 1992. AAAI Press.

[17] Colin P. Williams and Tad Hogg. Extending deep structure. In *Proc. of the 11th Natl. Conf. on Artificial Intelligence (AAAI93)*, pages 152–157, Menlo Park, CA, 1993. AAAI Press.

[18] Colin P. Williams and Tad Hogg. Exploiting the deep structure of constraint problems. *Artificial Intelligence*, 70:73–117, 1994.

[19] Weixiong Zhang and Richard E. Korf. An average-case analysis of branch-and-bound with applications: Summary of results. In *Proc. of the 10th Natl. Conf. on Artificial Intelligence (AAAI92)*, pages 545–550, Menlo Park, CA, 1992. AAAI Press.