JOURNAL OF ALGORITHMS 6, 275–282 (1985)

A Theoretical Analysis of Backtracking in the Graph Coloring Problem

EDWARD A. BENDER

University of California at San Diego, La Jolla, California 92093

AND

HERBERT S. WILF

University of Pennsylvania, Philadelphia, Pennsylvania 19104

Received February 13, 1984

The graph coloring problem is: Given a positive integer K and a graph G. Can the vertices of G be properly colored in K colors? The problem is NP-complete. The average behavior of the simplest backtrack algorithm for this problem is studied. Average run time over all graphs is known to be bounded. Average run time over all graphs with n vertices and q edges behaves like $\exp(Cn^2/q)$. It is shown that similar results hold for all higher moments of the run time distribution. For all graphs and for graphs where $\lim n^2/q$ exists, the run time has a limiting distribution as $n \to \infty$.

1. INTRODUCTION

The graph coloring problem is: Given a positive integer K and a graph G. Can the vertices of G be properly colored in K colors? The problem is *NP*-complete.

The backtrack algorithm that we have in mind for this problem begins by coloring vertex 1 in color 1. In general, if vertices $1, 2, \ldots, L-1$ have been colored, then vertex L receives the color of lowest number that is consistent with the colors already assigned. If no such color exists, then backtrack by increasing the color of vertex $L-1$, etc. The algorithm will not halt until *all* possible assignments have been tried in this fashion, even if a proper coloring is found.

A graph on n labelled vertices is specified by indicating which of the possible $\binom{n}{2}$ edges is present. Let each edge be present with independent

0196-6774/85 $3.00

probability p. Define a random variable $X_n(p)$ equal to running time on a randomly generated graph as measured by the number of nodes in the search tree that the algorithm generates, described below. In [1] it was shown that backtracking is an $O(1)$ average time algorithm for all graphs; that is, the expected value of $X_n(\frac{1}{2})$ is bounded.

In this paper we investigate the distribution of $X_n(p)$ both for constant p and for $p = p(n) \to 0$. For fixed $p \neq 0$, all moments of $X_n(p)$ are bounded and $\Pr\{X_n(p) = k\}$ is independent of n when $n \geq k$. If $p \geq C/\log n$, the moments of $X_n(p)$ grow no faster than polynomially with n; else if $p \geq C/n^{1-\epsilon}$, they grow subexponentially. With minor modifications, these results carry over to graphs chosen randomly from the set of those with n vertices and q edges.

2. Notation and Terminology

Our graphs are simple and undirected, with vertex set $V(G)$ labelled $1, 2, \ldots, |V(G)|$. The set of all labelled n-vertex graphs is $\mathcal{G}_n$ and the subset with exactly q edges is $\mathcal{G}_{n,q}$. If $0 \leq p \leq 1$, we make $\mathcal{G}_n$ into a probability space $\mathcal{G}_n(p)$ by assigning $G \in \mathcal{G}_{n,q}$ a probability $p^q(1-p)^{\binom{n}{2}-q}$. When $p = \frac{1}{2}$, this assigns each $G \in \mathcal{G}_n$ the same probability. $\Pr\{G\}$ is the probability assigned to the graph by $\mathcal{G}_n(p)$.

An *induced subgraph* H of a graph G is obtained by choosing some subset $S \subseteq V(G)$ for the vertices of H. The edges of H are precisely all edges of G both of whose endpoints lie in S. We then say that *H is the subgraph of G that is induced by the set S*. By $H_L(G)$ we mean the subgraph of G that is induced by the set $\{1, 2, \ldots, L\}$.

The *chromatic polynomial* $P(\lambda, G)$ of a graph G is the polynomial in λ of degree $|V(G)|$ such that for each $K = 1, 2, \ldots,$ the value of $P(K, G)$ is the number of ways to color the vertices of G properly in the colors $1, 2, \ldots, K$.

The action of the backtrack algorithm, as described in the introduction, on a graph G generates a *search tree* $T_K(G)$ whose nodes are arranged in levels $0, 1, \ldots, n = |V(G)|$. The node at level $L = 0$ is the *root*. At level $L > 0$ there is a node corresponding to each proper coloring of vertices $1, 2, \ldots, L$ of G in K colors. Thus there are precisely $P(K, H_L(G))$ nodes at level L. There is an edge between nodes v at level L and v' at level $L + 1$ if and only if they agree in the colors that they assign to vertices $1, 2, \ldots, L$. Figure 2 shows the backtrack tree $T_3(G)$ for the graph in Fig. 1. That tree has 46 nodes, each labelled by the colors it assigns to the vertices of G. For example, the node labelled "12132" in Fig. 2 corresponds to coloring vertices 1, 2, 3, 4, 5 of G in colors 1, 2, 1, 3, 2, respectively.

We will write $\beta_K(G)$ for $|V(T_K(G))|$. It measures the "run time" of the algorithm. If G is chosen by $\mathcal{G}_n(p)$, then the value of $\beta_K(G)$ is a random

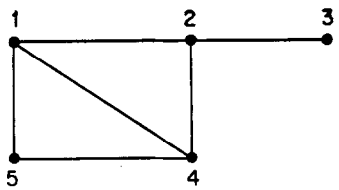

FIGURE 1

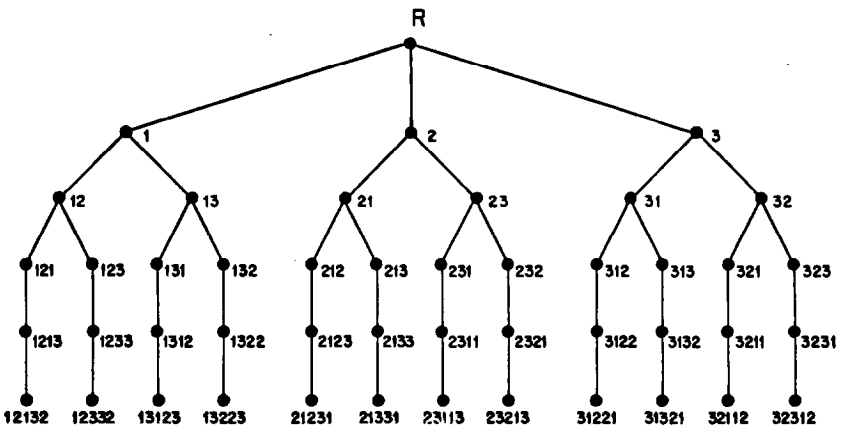

FIGURE 2

variable denoted by $X_n(p)$. Our analysis can be regarded as combinatorial, but the language of probability is convenient in this context. For the *moments* of $X_n(p)$, $E(X_n(p)^m)$, we shall write $\beta_K^{(m)}(n, p)$. This is the average of $(\beta_K(G))^m$ over all $G \in \mathcal{G}_n(p)$, weighted according to probability.

3. STATEMENT OF RESULTS

The following theorems will be proved in the next section. First, as regards the growth of the moment sequence, we have

THEOREM 1. *Let $p = p(n)$ be such that $pn \to \infty$ as $n \to \infty$. Then for some $C_i(K) > 0$,*

$$C_1(K)m^2/p \le \log\bigl(\beta_K^{(m)}(n, p)\bigr) \le C_2(K)m^2/p \tag{3.1}$$

as $n \to \infty$. *If, in addition* $p \to 0$, *then*

$$\log\left(\beta_K^{(1)}(n, p)\right) \sim \frac{K(\log K)^2}{2p}. \tag{3.2}$$

Next, in case of constant edge density, we will prove a good deal more, namely

THEOREM 2. *Let* $p \in (0, 1)$ *be fixed. Then*

(A) (*Existence of limiting moments*). *For each fixed* m,

$$\lim_{n \to \infty} \beta_K^{(m)}(n, p) =: \beta_K^{(m)}$$

exists and is finite.

(B) (*Existence of probability density function*). *There is a function* $f_K(j, p) \geq 0$ *such that* $\sum_j f_K(j, p) = 1$ *and* $\Pr\{X_n(p) = j\} = f_K(j, p)$ *for* $n \geq j$.

(C) (*Estimate of tail of distribution*). *There is a* $C = C(K, p) > 0$ *such that, for all* $t \geq 0$, $\Pr\{X_n(p) \geq t\} \leq e^{-C(\log t)^2}$.

Instead of using $\mathcal{G}_n(p)$, we could assign equal probabilities to all graphs in $\mathcal{G}_{n,q}$. We would use the random variable $X_{n,q}$ in place of $X_n(p)$. As is usual in the theory of random graphs, our theorems would remain nearly valid if p were replaced by $q/\binom{n}{2}$. The reason for "nearly" valid is that Theorem 2(B) must then be replaced by

$$\lim_{n \to \infty} \Pr\{X_{n,q} = j\} = f_K(j, p).$$

The proofs would be somewhat more complicated since powers of probabilities would be replaced by ratios of binomial coefficients.

We generated the complete backtrack search trees for 1000 random graphs with $n = 20$, $p = \frac{1}{2}$, and $K = 3$. The observed mean was $\hat{\beta}_3^{(1)}(20, \frac{1}{2}) = 204.7$ and the observed standard deviation was (coincidentally) also 204.7. We then repeated this with $n = 40$, observing a mean and standard deviation of 199.6 and 252.0, respectively. In Table 1 we show the observed

TABLE 1

Observed Tree Size Distribution $F_3(j, \frac{1}{2})$ for $n = 20, 40$

$n \backslash j$	50	100	150	200	300	400	500	600	700	800	900	1000
20	.144	.367	.573	.687	.823	.893	.936	.956	.968	.980	.985	.990
40	.149	.358	.557	.680	.832	.904	.935	.963	.975	.982	.988	.992

cumulative distributions $F_K(j, p) = \sum_{i \le j} f_K(i, p)$. Thus, the numerical results are already in good agreement with Theorem 2(B), i.e., that the cumulative distribution function of search tree size approaches a limit.

4. Proofs

Before proving the theorems, we establish some lemmas.

Lemma 1. *For some* $C(K) > 0$,

$$K^n(1-p)^{(n^2/2K)-n/2} \ge \sum_{H \in \mathcal{G}_n} P(K, H)\Pr\{H\}$$

$$\ge C(K)\frac{K^n(1-p)^{(n^2/2K)+n/2}}{n^{K/2}}.$$

Proof. Let π be a partition of $\{1, 2, \ldots, n\}$ into an ordered collection of K (possibly empty) blocks, and let $S(\pi)$ be the sum of the probabilities of those graphs in $\mathcal{G}_n$ which are properly colored when the vertices in the ith block of π are assigned color i. Then

$$\sum_{H \in \mathcal{G}_n} P(K, H)\Pr\{H\} = \sum_{\pi} S(\pi). \tag{4.1}$$

A graph G will contribute to $S(\pi)$ if and only if for each i none of the vertices in the ith block of π are joined by edges of G. If the cardinality of the ith block is s_i, we have $S(\pi) = (1-p)^{\sigma}$, where $\sigma = \sum\binom{s_i}{2}$. Thus we can collect terms in (4.1) according to the cardinality of the blocks of π,

$$\sum_{H \in \mathcal{G}_n} P(K, H)\Pr\{H\} = \sum_{s_1 \cdots s_K} \frac{n!}{s_1! \cdots s_K!}(1-p)^{\sigma}, \tag{4.2}$$

where the sum ranges over all non-negative integers s_i adding to n. The factor $(1-p)^{\sigma}$ never exceeds $(1-p)^{n(n/K-1)/2}$. Since the first factor on the right side of (4.2) sums to K^n, the upper bound is proved. We can certainly choose integers s_i so that they sum to n and $|s_i - n/K| \le 1$. We bound (4.2) from below by this term. In this case, $\binom{s_i}{2} \le (n/K + 1)n/2K$ and by Stirling's formula we find

$$\frac{n!}{s_1! \cdots s_K!} \ge \frac{C(K)K^n}{n^{K/2}}. \quad \square$$

LEMMA 2.

$$\beta_K^{(m)}(n,p) = \sum_{L=0}^{n} \sum_{H \in \mathscr{G}_L} \Pr\{H\} \sum_{\substack{\max L_i = L \\ 0 \le L_i \le L}} \prod_{i=1}^{m} P\big(K, H_{L_i}(H)\big).$$

Proof. We have

$$\begin{aligned}
\beta_K^{(m)}(n,p) &= \sum_{G \in \mathscr{G}_n} \Pr\{G\} |V(T_K(G))|^m \\
&= \sum_{G \in \mathscr{G}_n} \Pr\{G\} \left(\sum_{L=0}^{n} P(K, H_L(G)) \right)^m \\
&= \sum_{L_1, \ldots, L_m} \sum_{G \in \mathscr{G}_n} \Pr\{G\} \prod_{i=1}^{m} P\big(K, H_{L_i}(G)\big) \\
&= \sum_{L=0}^{n} \sum_{H \in \mathscr{G}_L} \Pr\{H\} \sum_{\max L_i = L} \prod_{i=1}^{m} P\big(K, H_{L_i}(H)\big),
\end{aligned}$$

where the last line is based on the following two observations. If $L = \max L_i$, then $H_{L_i}(G) = H_{L_i}(H)$, where $H = H_L(G)$. Second, the sum of $\Pr\{G\}$ over all $G \in \mathscr{G}_n$ with $H = H_L(G)$ is simply $\Pr\{H\}$ because only the edges connecting $1, \ldots, L$ are restricted. □

LEMMA 3. *Suppose that $a > 0$ and $b \ge 0$ are functions of n with $b/a = o(n)$ and $B \ge a \ge B/n^2$. Then there are $C_i(B) > 0$ such that for all large n*

$$C_1(B) \le \sqrt{a}\, e^{-b^2/4a} \sum_{L=0}^{n} e^{-aL^2 + bL} \le C_2(B).$$

Proof. We have $-aL^2 + bL = -ax^2 + b^2/4a$, where $x = L - b/2a$. Thus the expression in the lemma can be written as $\sqrt{a} \sum e^{-ax^2}$, the summation ranging over all x for which $x + b/2a$ is an integer between 0 and n. This sum is an approximation to $\int e^{-u^2}\, du$ with step size $\sqrt{a}$. Since x ranges from $-b/2a \le 0$ to $n - b/2a > 0$, the interval of integration contains 0 and has length $n\sqrt{a} \ge \sqrt{B}$. □

We now prove Theorem 1, beginning with the lower bound in (3.1). Consider those terms in Lemma 2 with every $L_i = L$ and H a graph with no edges. Thus

$$\beta_K^{(m)}(n,p) \ge \sum_{L=0}^{n} (1-p)^{\binom{L}{2}} K^{mL}.$$

Apply Lemma 3 with $a = -\frac{1}{2}\log(1-p)$ and $b = m\log K + a$. Use $-\log(1-p) \sim p$ as $p \to 0$.

The upper bound in (3.1) is somewhat more complicated. Remove from the product in Lemma 2 some factor with $L_i = L$ and re-index the remainder. Since there are m choices for i, we have

$$\beta_K^{(m)}(n,p) \le m \sum_{L=0}^{n} \sum_{H \in \mathscr{G}_L} P(K,H)\Pr\{H\} \sum_{L_i \le L} \prod_{i=2}^{m} P\big(K, H_{L_i}(H)\big)$$

$$= m \sum_{L=0}^{n} \sum_{H \in \mathscr{G}_L} P(K,H)\Pr\{H\} \left(\sum_{j=0}^{L} P\big(K, H_j(H)\big) \right)^{m-1}.$$

Since the chromatic polynomial of a j vertex graph is at most K^j, we can bound the inner sum by K^{L+1}. Using the upper bound in Lemma 1, we get

$$\beta_K^{(m)}(n,p) \le m \sum_{L=0}^{n} K^L (1-p)^{L^2/2K - L/2} K^{(L+1)(m-1)}$$

$$= mK^{m-1} \sum_{L=0}^{n} K^{mL} (1-p)^{L^2/2K - L/2}.$$

Apply Lemma 3.

We now prove (3.2). With $m = 1$ in Lemma 2, we obtain as the inner sum precisely the sum estimated in Lemma 1. Thus

$$\sum_{L=0}^{n} K^L (1-p)^{L^2/2K - L/2} \ge \beta_K^{(1)}(n,p)$$

$$\ge C(K) \left(\sum_{L=1}^{n} \frac{K^L (1-p)^{L^2/2K + L/2}}{L^{K/2}} + 1 \right). \tag{4.3}$$

In the sum on the right, the single term with

$$L = \left\lfloor \frac{-K\log K}{\log(1-p)} \right\rfloor$$

already contributes an amount

$$\exp\left(\frac{K(\log K)^2}{2p} (1 + o(1)) \right)$$

as required in (3.2). That the sum on the left in (4.3) is also of that size follows from Lemma 3. This completes the proof of Theorem 1.

The existence of the limit in Theorem 2 follows from the observation that $\beta_K^{(m)}(n, p)$ is monotonic increasing in n by Lemma 2 and bounded above by (3.1).

Suppose that $n \geq j$ and $G \in \mathscr{G}_n$. If $T_K(G)$ has precisely j nodes, then it has no nodes at level j and so $|V(T_K(G))| = j$ if and only if $|V(T_K(H_j(G)))| = j$. Hence $\Pr\{X_n(p) = j\}$ equals the sum of $\Pr\{H\}$ over all $H \in \mathscr{G}_j$ with $|V(T_K(H)| = j$. Call this sum $f_K(j, p)$. We must show that $\sum_j f_K(j, p) = 1$. Since $\beta_K^{(1)}(n, p)$ is bounded, for every $\delta > 0$ there is an i such that $\Pr\{X_n(p) \leq i\} > 1 - \delta$ for all n. Set $n = i$ to get $\sum_{j \leq i} f_K(j, p) \geq 1 - \delta$.

For the last claim in Theorem 2, use (3.1) and $\Pr\{X_n(p) \geq t\} t^m \leq \beta_K^{(m)}(n, p)$ with $m = (p \log t)/2C_2(K)$.

Reference

1. H. S. Wilf, Backtrack: An $O(1)$ expected time algorithm for the graph coloring problem, *Inform. Processing Lett.* 18 (1984), 119–122.