OPTIMIZATION PROBLEMS ON GRAPHS WITH INDEPENDENT RANDOM EDGE WEIGHTS[*]

by

George S. Lueker[+]

Department of Information and Computer Science
University of California, Irvine
Irvine, CA 92717

Technical Report #131[‡]
March, 1979

Keywords: Random graphs, optimization problems, normal distribution, weighted graphs, probabilistic analysis, traveling salesman problem, cliques, Boole's inequality, greedy algorithms, convergence in probability, almost sure convergence.

----

[*] This paper presents and extends the results in "Maximization Problems on Graphs with Edge Weights Chosen from a Normal Distribution," presented at the _Tenth ACM Symposium on Theory of Computing_, held in San Diego in May 1978.

[+] Supported by the National Science Foundation under grant MCS77-04410.

[‡] This TR is a substantially revised and extended version of TR #115.

OPTIMIZATION PROBLEMS ON GRAPHS WITH INDEPENDENT RANDOM EDGE WEIGHTS

George S. Lueker

Abstract

We consider optimization problems on complete graphs with edge weights drawn independently from a fixed distribution. We discuss several methods for analyzing these problems, including greedy methods, applications of Boole's inequality, and exploitation of relationships with results about random unweighted graphs. We illustrate these techniques in the case in which the edge weights are drawn from a normal distribution; in particular, we investigate the expected behavior of the minimum weight clique on k vertices. We describe the asymptotic behavior (in probability and/or almost surely) of the random variable which describes the optimum; we also discuss the asymptotic behavior of its mean. Next we demonstrate techniques by which we may determine an asymptotic description of the behavior of a greedy algorithm for this problem.

1. Introduction

Many results have been proven about the properties of random graphs. Some of these [1, 3, 9, 10, 12, 17, 23, 24, 26] deal with graphs constructed by letting each possible edge be present with a specified probability; one then tries to estimate the probability that a subgraph of a given type will be present. ([11] may be considered a paper about random directed graphs.) We will call such a problem a subgraph existence problem. Another area of interest is algorithms on graphs in which all edges are present, but weights are assigned to the edges according to some distribution; one then tries to find the minimum weight subgraph of a given type. We will call such problems subgraph optimization problems; they are the subject of this paper. For example, if a traveling salesman problem is constructed using the Euclidean distance between n points chosen from a uniform distribution in the unit square, then asymptotically the optimum solution is proportional to $n^{1/2}$ [2]; very efficient algorithms have been designed whose asymptotic behavior tends to be optimal [19, 21]. The assignment problem for the case in which edge weights are chosen from various distributions has been analyzed by Borovkov [4]; for this problem, the case in which the edge weights are chosen from a uniform distribution appears to be particularly difficult, and has been further pursued by Walkup [27].

In section 2, we present some basic definitions and facts. In section 3, we will discuss a very general technique for obtaining lower bounds on the values of solutions to subgraph optimization problems, based on Boole's inequality. Section 4 discusses a very general technique for obtaining upper bounds on these values, using theorems about subgraph existence problems. Combining the bounds of sections 3 and 4 often enables us to make rather precise statements about the asymptotic behavior of the minimum, as will be shown in section 5. Since many optimization problems are NP-complete [16, 20], it is useful to investigate the behavior of heuristics. In section 6 we will investigate the behavior of some greedy algorithms.

2. Definitions

We will frequently discuss probabilities and expected values. If X is a random variable and A and B are events, let P{A} be the probability of A, P{A|B} be the probability of A given B, E[X] be the expected value of X, and E[X|A] be the expected value of X given A.

Throughout this paper, $\mathbb{G}_n$ will be a random structure which is a complete, weighted, labeled graph on n vertices; we will assume the vertices are labeled 1,2,...,n. Weights are chosen, independently, from a distribution whose probability density function (pdf) is f, and whose (cumulative) probability distribution function (PDF) is F. X will denote the random variable whose PDF is F. G will denote some particular weighted complete graph. The weight of the edge joining vertex v and w will be denoted d(v,w). Depending on the application, $\mathbb{G}_n$ may be undirected or directed; in the former case, d(v,w) is of course symmetric. When we make asymptotic statements about the behavior of some random variable which is a function of $\mathbb{G}_n$, we will assume that an infinite sequence $\mathbb{G}_n$, n=1,2,...., is considered, with each graph drawn independently.

Let S_n be a set of labeled graphs on n vertices; again, the vertices are labeled 1,2,...,n, so there is a natural one-to-one correspondence between the vertices of an element of S_n and the vertices of $\mathbb{G}_n$. All elements of S_n are assumed to have the same number of edges. For any H in

S_n, and any weighted graph G, let W(G,H) be the number found by summing, over all edges in H, the weight of the corresponding edge in G. For a given G, we wish to choose H in S_n so as to minimize W(G,H); this minimum will be called $W_{min}(G)$. Note that, for example, if S_n is the set of the (n-1)!/2 cycles on n vertices in an undirected graph, $W_{min}(G)$ gives the solution to the traveling salesman problem. We wish to investigate the expected behavior of $W_{min}(\mathbb{G}_n)$. (Often in an optimization problem, we wish to maximize some quantity; for uniformity, however, we will always assume that we are minimizing quantities. The methods used here could also be applied to maximization problems.)

In this paper we will often wish to discuss inequalities which hold approximately, most of the time, for large enough n. In order to make such statements precisely, we need to introduce some notation. Let Y_n and Z_n be sequences of reals, and choose $\epsilon>0$. For any n and $\epsilon>0$, consider the following two propositions:

$$Y_n \leq Z_n + \epsilon |Z_n| \tag{1}$$

$$Y_n \geq Z_n - \epsilon |Z_n| \tag{2}$$

If for every $\epsilon>0$, (1) (respectively (2)) holds except for finitely many n, we will write $Y_n \lesssim Z_n$ (respectively $Y_n \gtrsim Z_n$). If for every $\epsilon>0$, both (1) and (2) hold except for finitely many n, we write $Y_n \sim Z_n$.

Now let Y_n and Z_n be sequences of random variables; we will not assume that Y_n and Z_n are independent, but we will assume that variables with different indices are independent. (In our applications, each Z_n will often be a constant.) Note that now (1) and (2) are events rather than simple predicates. Let $P_1(n,\epsilon)$ be the probability that (1) fails and $P_2(n,\epsilon)$ be the probability that (2) fails. If for each $\epsilon>0$, $P_1(n,\epsilon)$ (respectively $P_2(n,\epsilon)$) goes to 0 as n approaches infinity, we will say $Y_n \lesssim Z_n$ (respectively $Y_n \gtrsim Z_n$) in probability. If both P_1 and P_2 approach zero for all $\epsilon>0$, we write $Y_n \sim Z_n$ in probability. (The phrase "in probability" will be abbreviated "(pr.)".)

A much stronger notion is that of almost sure behavior. An asymptotic statement holds almost surely if the set of sequences Y and Z which do not obey the statement has probability measure 0. Suppose that, for each $\epsilon>0$, except with probability 0, the sequences Y_n and Z_n fail to satisfy (1) for only finitely many n. Then by an argument like that of Theorems 4.1.1 and 4.2.2 in [5], we may write

$$Y_n \lesssim Z_n \text{ almost surely.} \qquad (3)$$

("Almost surely" will be abbreviated "(a.s.)".) By the Borel-Cantelli lemmas (see, for example, [5]), an equivalent definition of (3) is

$$\forall\ \epsilon > 0,\ \sum_{n=0}^{\infty} P_1(n,\epsilon) < \infty .$$

We may similarly define statements that $Y_n \gtrsim Z_n$ or $Y_n \sim Z_n$

almost surely. Sometimes we will show that an asymptotic statement which is true in probability is not true almost surely; it then follows that the statement is almost surely false (see Corollary [5, p. 78]), even though it is true in probability!

Note that statements about probabilistic convergence and convergence of expected values are somewhat independent. In particular, either, both, or neither of the following two statements may be true:

$$E[Y_n] \sim E[Z_n]$$

$$Y_n \sim Z_n \quad \text{(a.s.)}$$

For more information and examples, see [5] and [28].

We will illustrate the methods discussed in this paper in the case in which f is the unit normal distribution. The following few observations, which are well-known or easily established, are useful. If X is some random variable, let $X_{i:n}$ denote the random variable obtained by selecting the i^{th} smallest of n independent observations of X.

<u>Fact 1</u>. Let X be a unit normal variable and let A be any event with probability p. Then as p→0,

a) $|E[X \mid A]| \underset{\sim}{<} (2 \log p^{-1})^{1/2}$.

b) $E[|X| \mid A] \underset{\sim}{<} (2 \log p^{-1})^{1/2}$.

The proof is easy and omitted.

Fact 2. Let X be a unit normal variable. Then

a) $E[X_{1:n}] \sim -(2 \log n)^{1/2}$.

Moreover, for any ϵ with $0<\epsilon<1$,

b) $P\{X_{1:n} \geq -(1-\epsilon)(\log n)^{1/2}\} \lesssim \exp(-n^{\epsilon})$.

c) $P\{X_{1:n} \leq -(1+\epsilon)(2 \log n)^{1/2}\}$

$\sim n\, F(-(1+\epsilon)(2 \log n)^{1/2})$

$= \Theta(n^{-2\epsilon-\epsilon^2} (\log n)^{-1/2})$.

See [7] for a thorough discussion of distributions of minima and other order statistics; parts (b) and (c) are simple calculations.

Note that for any k, and any $\epsilon>0$,

$$\lim_{n \to \infty} n^k \exp(-n^{\epsilon}) = 0;$$

we will describe this by saying that $\exp(-n^{\epsilon})$ *swallows polynomials*. Note also that we may conclude that

$$X_{1:n} \lesssim -(2 \log n)^{1/2} \quad \text{(a.s.)}$$

while on the other hand

$$X_{1:n} \gtrsim -(2 \log n)^{1/2} \quad \text{(pr. but not a.s.)}.$$

Intuitively, this is because the minimum cannot be greater (algebraicly) than some bound unless all n observations are greater than the bound; for it to be less than the bound, only a single observation needs to be low. (Sen probably discusses related phenomena in [25], but I have not yet gotten hold of a copy of this paper.) This sort of behavior will arise again when we consider the problem of finding minimum weight cliques.

Fact 3. Let X be a unit normal variable and let F be its PDF. Then as $p \rightarrow 0$,

$$P\{X \leq (1+\epsilon)\ F^{-1}(p)\ |\ X \leq F^{-1}(p)\} = \Theta(p^{2\epsilon+\epsilon^2}).$$

Now let F again be an arbitrary PDF, and f the corresponding pdf. Often we will need to consider the sum of several of these variables; we will let X^{*k} be the random variable corresponding to the sum of k random variables chosen independently according to F, and let F^{*k} be the corresponding PDF. Note that if F is unit normal, then

$$F^{*k}(x) = F(k^{-1/2}\ x). \tag{4}$$

In order to discuss minimization problems, we will need to discuss the expected value of a sum given that a certain event is true; the following notation will be helpful. If m is a positive integer and p is a real in [0,1], let

$$B(m,p,F) = E[X^{*m}\ |\ F^{*m}(X^{*m}) \leq p].$$

Note that if A is any event with probability p, then

$$E[X^{*m} \mid A] \geq B(m,p,F);$$

note also that if F is unit normal,

$$B(m,p,F) \sim -(2\ m \log p^{-1})^{1/2} \tag{5}$$

3. A lower bound

In this section we derive a simple bound on the expected behavior of $W_{min}(\mathbb{G}_n)$ and on the PDF of $W_{min}(\mathbb{G}_n)$. The method is a straightforward application of Boole's inequality; see [6, 7, 14]. See [15] for another application of this inequality to an optimization problem; there a problem involving points distributed uniformly over the Euclidean plane was investigated. (Donath [8] used a combinatorial argument, for a version of the assignment problem with integer weights, which is in some ways similar to that given here.) Let M_n be the cardinality of S_n; recalling that each element of S_n has the same number of edges, let m_n be this common number.

Lemma 1. $P\{W_{min}(\mathbb{G}_n) \leq x\} \leq M_n\ F^{*m_n}(x)$.

Proof. We have

$$P\{W_{min}(\mathbb{G}_n) \leq x\}$$

$$= P\{\exists H \in S_n \text{ such that } W(\mathbb{G}_n,H) \leq x\}$$

(by the definition of W_{min})

$$\leq \sum_{H \in S_n} P\{W(\mathbb{G}_n, H) \leq x\}$$

(by Boole's inequality [14, p. 23])

$$= \sum_{H \in S_n} F^{*m_n}(x)$$

(since each H has m_n edges)

$$= M_n \, F^{*m_n}(x).$$ []

Corollary 1. $E[W_{min}(\mathbb{G}_n)] \geq B(m_n, M_n^{-1}, F)$.

Proof. Note that if a random variable had a PDF of $\min(1, M_n F^{*m_n})$, its expectation would be precisely $B(m_n, M_n^{-1}, F)$. []

Corollary 2. If F is unit normal, and $M_n \to \infty$, then

$$E[W_{min}(\mathbb{G}_n)] \gtrsim -(2\, m_n \log M_n)^{1/2}$$

Proof. This follows immediately from the previous corollary and (5). []

Corollary 3. If F is unit normal, and M_n approaches infinity, then

$$W_{min}(\mathbb{G}_n) \gtrsim -(2\, m_n \log M_n)^{1/2} \quad \text{(pr.)}.$$

Moreover, if M_n^{-1} swallows polynomials, then this bound holds almost surely.

Proof. Using Lemma 1, we know that the probability of

$$W_{min}(\mathbb{G}_n) \leq -(1+\epsilon)\ (2\ m_n \log M_n)^{1/2}$$

is bounded by

$$M_n\ F^{*m_n}(-(1+\epsilon)\ (2\ m_n \log M_n)^{1/2})$$

$$= M_n\ F(-(1+\epsilon)\ (2 \log M_n)^{1/2})$$

(by (4))

$$= O(M_n^{-2\epsilon}).$$

(by Fact 2c)

Clearly this goes to zero if M_n goes to infinity; moreover, if M_n^{-1} swallows polynomials, the sum of this probability must converge, so by the Borel-Cantelli lemma the almost sure convergence is established. []

4. An upper bound

In this section we obtain an upper bound on the expected behavior of $W_{min}(\mathbb{G}_n)$. We will use some results about subgraph existence problems on random graphs. Define a random structure $\mathbf{G}_{n,p_n}$ to be a graph on n vertices, where each edge is present with probability p_n, independently of the others. Then, for example, it is known [1, 24] that for any i, we can choose a c large enough so that $\mathbf{G}_{n,p_n}$ has a hamiltonian cycle except with probability $O(n^{-i})$, if

$p_n=(c \log n)/n$.

In this section, we discuss a simple lemma which relates results of this form to the optimization problems we are considering. (This lemma was used in [22] for the case of normal distributions; it can be stated much more generally, as was suggested by the referees and also observed by Weide [28], who attributes the idea of the lemma to T. Nishizeki. Walkup [27] has also exploited the relationship between existence and optimization problems.)

Lemma 2 [22, 28]. Let p_n be a sequence of reals in $[0,1]$ and let q_n be the probability that $\mathbb{G}_{n,p_n}$ fails to contain an element of S_n. Then

$$W_{\min}(\mathbb{G}_n) \leq m_n F^{-1}(p_n) \qquad (6)$$

except with a probability of at most q_n.

Proof. Consider the following algorithm for choosing an element H of S_n.

1. Let $a = F^{-1}(p_n)$, and let H_0 be some fixed element of S_n.

2. Let E be the set of edges in G whose weight is less than a; call these light edges.

3. Let H be any element of S_n all of whose edges are light, and stop. If no such H can be found, go on to step 4.

4. Let $H = H_0$.

Note that, except with probability of at most q_n, this algorithm returns a subgraph whose weight satisfies the desired inequality. □

<u>Corollary 1</u> [22]. Suppose that F is unit normal, and that q_n goes to zero rapidly enough that

$$q_n \, (m_n \log q_n^{-1})^{1/2} = o(m_n \, (\log p_n^{-1})^{1/2}).$$

Then

$$E[W_{min}(\mathbb{G}_n)] \lesssim - m_n \, (2 \log p_n^{-1})^{1/2}.$$

<u>Proof</u>. Consider again the algorithm in the proof of the lemma. Note that with probability approaching 1, the algorithm will return a subgraph H whose weight is at most

$$m_n \, F^{-1}(p_n) \sim - m_n \, (2 \log p_n^{-1})^{1/2}.$$

Let FAIL be the event that we fail to find an element of S_n among the light edges, and must therefore set H to H_0; the probability of FAIL is just q_n. By Fact 1, and the fact that $W(\mathbb{G}_n, H_0)$ is normally distributed with variance m_n, we may conclude that the expected weight of H_0 in the event FAIL is $O((m_n \log q_n^{-1})^{1/2})$. Then by the hypotheses of the corollary, the error we commit by ignoring the possibility of event FAIL is negligible. □

Corollary 2 [28]. If the sum of the q_n in the lemma converges, then

$$W_{min}(\mathbb{G}_n) \lesssim m_n \, F^{-1}(p_n) \quad \text{(a.s.)}$$

5. Some examples

In this section we show some applications of the methods discussed so far. As mentioned earlier, we will assume that edge weights are unit normal variables. The assignment problem for the normal distribution (and others) was analyzed by Borovkov [4]; he observed that a lower bound for this problem may be obtained by taking the sum of the minimum element in each row of the input matrix; similarly, he observed that a simple greedy algorithm yields a fairly good upper bound. Using these results he showed that

$$W_{min}(\mathbb{G}_n) \sim - n \, (2 \log n)^{1/2} \quad \text{(pr.)}.$$

His method also easily establishes a similar result for the traveling salesman problem.

Weide has used results about the probability of finding a hamiltonian circuit in $\mathbb{G}_{n,p_n}$ [1, 24], to show that for the traveling salesman problem with unit normal edge weights,

$$W_{min}(\mathbb{G}_n) \lesssim - n \, (2 \log n)^{1/2} \quad \text{(a.s.)}.$$

Using Corollary 3 to Lemma 1 we can easily extend this to also be an upper bound. A very similar analysis holds for the assignment problem, so we obtain the following.

Theorem 1. For the traveling salesman problem or the assignment problem, with unit normal edge weights,

$$W_{\min}(\mathbb{G}_n) \sim - n\ (2 \log n)^{1/2} \quad \text{(a.s.)}.$$

Of course, these examples do not provide much evidence for the power of the methods discussed here, since we have only slightly extended a long-known result. The bounds achieved in the next example, however, do not appear to be obtainable by simple greedy arguments. Consider the problem of finding the weight of the heaviest k-clique in a graph G. (By a k-clique we mean a subgraph on k vertices, all of which are adjacent. In the asymptotic statements which follow, we assume that k is fixed and n goes to infinity.) It is not at all easy to devise a greedy algorithm which gives good bounds for this problem; in fact, in the next section we will see that the natural greedy algorithm, in probability, fails to produce a good bound. The results of the previous sections, however, easily lead to a tight description of the behavior of this problem.

Theorem 2. For the problem of finding the lightest k-clique in an n-vertex graph with unit normal edge weights,

a) $W_{\min}(\mathbb{G}_n) \lesssim - k\ ((k-1) \log n)^{1/2}$ (a.s.)

b) $W_{\min}(\mathbb{G}_n) \sim -k\,((k-1)\log n)^{1/2}$ (pr. but not a.s.)

c) $E[W_{\min}(\mathbb{G}_n)] \sim -k\,((k-1)\log n)^{1/2}$.

Proof. Observe that Corollary 4 to Theorem 1 in [10], combined with the method of proof of Theorem 2(ii) in [3], can easily be used to show that if we let $p_n = n^{-2/(k-1)+\epsilon}$, then the probability that $\mathbb{G}_{n,p_n}$ fails to contain a k-clique is $O(\exp(-n^{\epsilon}))$. Thus Corollary 1 to Lemma 2 gives

$$E[W_{\min}(\mathbb{G}_n)] \lesssim -C(k,2)\ (2\log n^{2/(k-1)-\epsilon})^{1/2}$$

$$\sim -k\,((k-1)\log n)^{1/2}\ O(1+\epsilon),$$

so

$$E[W_{\min}(\mathbb{G}_n)] \lesssim -k\,((k-1)\log n)^{1/2},$$

and by Corollary 2 part (a) holds almost surely.

Next note that the number of edges in a k-clique is C(k,2), where C(i,j) denotes the number of combinations of i things taken j at a time. Further, the number of distinct k-cliques is C(n,k). Thus, by Lemma 1 and its corollaries, we see that

$$E[W_{\min}(\mathbb{G}_n)] \gtrsim -(2\ C(k,2)\log C(n,k))^{1/2}$$

$$\sim -k\,((k-1)\log n)^{1/2},$$

and

$$W_{min}(\mathbb{G}_n) \underset{\sim}{>} -k\ ((k-1)\ \log\ n)^{1/2}. \qquad \text{(pr.)} \tag{7}$$

Note that M_n does not become infinite fast enough to guarantee almost sure behavior by Corollary 3. We now sketch a proof that the bound in (7) does <u>not</u> hold almost surely. Choose δ small enough so that

$$(2\,\delta + \delta^2)\ (2/(k-1)) < 1. \tag{8}$$

Now let

$$p_n = n^{-2/(k-1)+\epsilon}, \tag{9}$$

as above. Then if we pick out the light edges of $\mathbf{G}_n$ as in the proof of Lemma 2, we can almost surely construct a k-clique using only light edges. But by (8), (9), and Fact 3, an arbitrarily chosen one of the light edges used in the clique will be less than $F^{-1}(p_n)$ by a factor of $(1+\delta)$ with a probability whose sum does not converge as n goes to infinity. Since this remains true even if we make ϵ arbitrarily close to zero, and since the number of edges in a k-clique is independent of n, this likelihood of a single excessively light edge must prevent (7) from holding almost surely. []

The results obtained so far demonstrate that the bounds discussed in the previous section can give tight descriptions for some interesting problems. However, these bounds are not always tight, even when the edge weights have unit normal distributions. Say a graph H has property X(k) if

a) H contains a clique of size k, and

b) H has k^2 edges.

Let S_n be the set of all n vertex graphs with property X(k). It is not difficult to show that for this choice of S_n, neither the lower nor the upper bound on the asymptotic expected behavior is tight; see [22] for details.

6. Regular greedy algorithms

It is easy to devise greedy algorithms for subgraph optimization problems. For example, to find the lightest hamiltonian path in a graph, one can start at an arbitrary vertex and iteratively walk to the nearest unused vertex. A greedy algorithm was used by Borovkov [4] in his analysis of the assignment problem. For the lightest clique problem, one can start with the cheapest edge and iteratively add the vertex which increases the weight of the clique by the smallest amount.

Such algorithms can be viewed in the following way. The desired output is a list L of the edges in the subgraph found; by a slight abuse of notation, we will say that a vertex is in L if it is an endpoint of an edge in L. Initially L is the null list. Each partial list L will somehow determine a family CHOICE(L) of sets of edges in G;

each set in CHOICE(L) must be disjoint from L. At each iteration, we choose the set of edges in CHOICE(L) of smallest total weight, and append it to L.

For example, in the k-clique algorithm discussed above, after r>0 iterations L would be a clique on r+1 vertices. If L is the empty list, CHOICE(L) would contain a set of singleton sets whose elements were the edges of the graph; for nonempty L, CHOICE(L) would contain one set for each vertex v not in L, namely, the set of r+1 edges which join v to vertices in L.

A Pidgin-Algol specification of the algorithm appears below; here cost(E), where E is a set of edges, denotes the total weight of the edges in E.

```
begin
  L ← the empty list;
  for r ← 1 until t do
    begin
      let E be the set in CHOICE(L) which minimizes
        cost(E);
      append the elements of E to the end of L;
    end;
end;
```

Let **A** be such an algorithm. If the length of L determines the cardinality of CHOICE(L) and of each element of CHOICE(L), we will say the algorithm is *regular*; henceforth we assume the algorithm is regular. This means that we know, for each r, how many choices are possible at iteration r (call this number c(r)) and how many edges will be added during iteration r (call this number e(r)). It is tempting at this point to use the following argument, which we shall call the naive analysis. At the r^{th} iteration, we choose the minimum of c(r) variables each of which is the sum of e(r) random variables chosen according to CHOICE; thus the amount we add to the cost of L is $X^{*e(r)}_{1:c(r)}$. (In notation like $X^{*a}_{b:c}$, the superscript is considered to have higher precedence; thus this would mean to choose the b^{th} smallest of c independent observations, each of which was the sum of a observations of X.) Let $\hat{X}_{\mathbb{A}}$ be the random variable corresponding to the sum of these variables over all iterations; i.e.,

$$\hat{X}_{\mathbb{A}} = \sum_{r=1}^{t} X^{*e(r)}_{1:c(r)}.$$

Let $\hat{F}_{\mathbb{A}}$ be the corresponding PDF. Let $F_{\mathbb{A}}$ be the PDF which describes the true distribution of outputs of $\mathbb{A}$, and let $X_{\mathbb{A}}$ be the corresponding random variable.

Now $\hat{F}_{\mathbb{A}}$ and $F_{\mathbb{A}}$ may be different; the flaw in the above analysis is twofold:

a) after several iterations the edge weights have been conditioned by previous choices made during the algorithm, and

b) the sets in CHOICE(L), for some L, may overlap, and we are thus not choosing the minimum of independent variables.

If we rule out such problems, the analysis becomes much easier. Define a simple greedy algorithm to be a regular greedy algorithm for which, at each iteration, none of the edges in CHOICE(L) can have appeared in a set in CHOICE(L) at some previous iteration, and the sets in CHOICE(L) are disjoint from each other. Then if $\mathbb{A}$ is a simple greedy algorithm, we have $F_{\mathbb{A}}=\hat{F}_{\mathbb{A}}$ (see [28]).

Note that the natural greedy algorithm for the assignment problem is simple, so the analysis of this algorithm is easily carried out. The k-clique algorithm described above, however, is not simple; an edge may be considered at many different iterations. We shall, in the remainder of this paper, undertake the analysis of regular greedy algorithms which are not simple.

<u>Theorem</u> <u>3</u>. If $\mathbb{A}$ is a regular greedy algorithm, then for all x

$$F_{\mathbb{A}}(x) \leq \hat{F}_{\mathbb{A}}(x).$$

Hence $E[X_{\mathbb{A}}]\geq E[\hat{X}_{\mathbb{A}}]$.

For the proof we will need a lemma, whose proof is uninteresting and deferred to the appendix.

<u>Lemma</u> <u>3</u>. Let w be a column vector of m independent real random variables chosen with a distribution function G. Let g be a real-valued function of m-vectors which is monotonic nondecreasing, in the sense that

$$w \leq w' \implies g(w) \leq g(w').$$

(Here w is said to be less than or equal to w' if the inequality holds in each component.) Finally, let B be an r by m matrix of nonnegative reals, and b be a column vector of r reals. Then

$$P\{g(w) \leq x \mid B\ w \geq b\} \leq P\{g(w) \leq x\}.$$

Proof of theorem. Suppose we are at the beginning of iteration r. Let L_0 be some possible value for L at this point, and let A_0 be the event that $L=L_0$. Consider the function

$$\min_{E \in CHOICE(L_0)} cost(E). \qquad (10)$$

Were it not for the conditioning on the probabilities of the edge weights due to previous iterations, the PDF for this minimum cost would be less than or equal that for $X^{*e(r)}_{1:c(r)}$. (The inequality is necessary because the sets in CHOICE(L_0) may not be disjoint.) We now show that this statement remains true even when we bear in mind that the edge weights are conditioned. Note that (10) depends only on edges which have not yet been chosen, and is monotonic increasing in these edges. Now since the choice of edges to add to L is determined by comparisons of sums of edge weights, the event A_0 can be phrased as a set of inequalities on the edge weights; each inequality expresses that fact that the selected set of edges was less than or equal to some other set allowed by CHOICE. Note that each edge not yet chosen must appear only on the greater side of these inequalities. Thus by the lemma, the true PDF for the variable which describes the total weight of edges added to L during this iteration can only be decreased by this conditioning. Summing over all possible L_0, and integrating over all values of the variables chosen so far, we obtain the theorem. []

To illustrate the application of this theorem, consider the greedy k-clique algorithm mentioned above, with unit normal edge weights. The naive analysis says that the algorithm returns a clique of weight

$$\hat{X}_{\mathbb{A}} = X_{1:C(n,2)} + \sum_{i=2}^{k-1} X^{*i}_{1:n-i}$$

$$\sim - s_k (\log n)^{1/2}, \quad \text{(pr.) where}$$

$$s_k = 2 + \sum_{i=2}^{k-1} (2i)^{1/2}.$$

Lemma 4.

a) $X_{\mathbb{A}} \gtrsim - s_k (\log n)^{1/2}$ (pr.), and

b) $E[X_{\mathbb{A}}] \gtrsim - s_k (\log n)^{1/2}$.

Proof. $\hat{X}_{\mathbb{A}}$ satisfies the indicated bounds, and hence so does $X_{\mathbb{A}}$ by the previous theorem. []

In order to complete our analysis of the behavior of the greedy algorithm for k-cliques, it would be desirable to have an upper bound on the behavior of the solution it obtains. The past theorem gives us little help in this direction, but we may nonetheless establish the desired bound.

Lemma 5.

a) $X_{\mathbb{A}} \lesssim - s_k (\log n)^{1/2}$ (a.s.), and

b) $E[X_{\mathbb{A}}] \lesssim - s_k (\log n)^{1/2}$.

Proof. An idea similar to that used in [13, 18] is useful here--we can simply eliminate all cases in which things do not work out as we like. Choose any $\epsilon>0$. Note that the probability that the first edge selected is above $-2(1-\epsilon)(\log n)^{1/2}$ goes to zero fast enough to swallow polynomials. Next consider the probability that for some set C of vertices, $|C|<k$,

$$\min_{v \notin C} \sum_{w \in C} d(v,w) \geq -(1-\epsilon)\,(2\,|C|\,\log\,(n-|C|))^{1/2}$$

Using Fact 2b, we see that for any fixed choice of C, this probability goes to zero fast enough to swallow polynomials. But, for fixed k, there are only polynomially many choices for C, so the sum of this probability, over all possible C with $|C|<k$, must go to zero fast enough to swallow polynomials. We may conclude that the algorithm produces a clique of weight less than $(1-\epsilon)$ times the expected value predicted by the naive analysis except with a probability which swallows polynomials. Thus the sum of this probability over all n must converge, so we have part (a). Part (b) is then easily obtained using Fact 1. []

Theorem 4. For the greedy k-clique algorithm with unit normal edge weights,

a) $X_{\mathbb{A}} \lesssim - s_k (\log n)^{1/2}$ (a.s.),

b) $X_{\mathbb{A}} \sim - s_k (\log n)^{1/2}$ (pr. but not a.s.), and

c) $E[X_{\mathbb{A}}] \sim - s_k (\log n)^{1/2}$.

Proof. Most of the theorem follows directly from Lemmas 4 and 5. To show that the asymptotic behavior does not hold almost surely, we may use an argument similar to that used in the proof of Theorem 2. []

Combining Theorems 2 and 4, we see that for $k \geq 3$,

$$X_{\mathbb{A}}/W_{min} \sim s_k/(k\ (k-1)^{1/2}) \quad \text{(pr. but not a.s.).} \qquad (11)$$

For k=2, the algorithm is of course exact, since it merely chooses the cheapest edge; as k approaches infinity the ratio on the right of (11) approaches $(8/9)^{1/2}$.

7. Conclusions

We have demonstrated the use of some basic methods for analysis of the expected behavior of subgraph optimization problems. These methods have enabled us not only to determine the expected behavior of the optimum, but also to demonstrate that the asymptotic behavior held in probability, and to determine whether or not it held almost surely. In addition, we have demonstrated how to analyze the behavior of a greedy algorithm, even when edge weights were conditioned as the algorithm proceeded, and when the algorithm was provably suboptimal.

Although many of the techniques discussed here are of fairly general applicability, we have demonstrated them only in the case where edge weights are chosen from a unit normal distribution. It would be easy to state the results in the case where an arbitrary mean and variance were stated for the normal distribution. In a later paper, we plan to investigate these same problems under distributions other than normal.

Appendix: Proof of Lemma 3.

Lemma 3. Let w be a column vector of m independent real random variables chosen with a continuous distribution function G. Let g be a real-valued function of m-vectors which is monotonic nondecreasing, in the sense that

$$w \leq w' \implies g(w) \leq g(w').$$

Finally, let B be an r by m matrix of nonnegative reals, and b be a column vector of r reals. Then

$$P\{g(w) \leq x \mid B\ w \geq b\} \leq P\{g(w) \leq x\}.$$

Proof. We prove the lemma by induction on m. For m=1 it is easy. Suppose it holds for m=k-1. We may decompose w as

$$w = (w^*, w_n),$$

where w^* is the first k-1 components of w, and w_n is the last component of w. Then

$$P\{g(w) \leq x \mid B\ w \geq b\}$$

$$= P\{g(w^*, w_n) \leq x \mid B_1\ w^* + B_2\ w_n \geq b\} \qquad (12)$$

where B_1 and B_2 are appropriate submatrices of B. We may write the right hand side of (12) as

$$\frac{\int dG(\xi)\ h(\xi)\ P\{g(w^*,\xi)\leq x \mid B_1\ w^* \geq b - B_2\,\xi\}}{\int dG(\xi)\ h(\xi)}$$

where

$$h(\xi) = P\{B_1\ w^* \geq b - B_2\,\xi\}.$$

Now by the inductive hypothesis, for any ξ,

$$P\{g(w^*,\xi)\leq x \mid B_1\ w^* \geq b - B_2\,\xi\} \leq P\{g(w^*,\xi)\leq x\}.$$

Thus an upper bound is

$$\frac{\int dG(\xi)\ h(\xi)\ P\{g(w^*,\xi)\leq x\}}{\int dG(\xi)\ h(\xi)}$$

But since $h(\xi)$ is easily seen to be monotonic increasing, while $P\{g(w^*,\xi)\leq x\}$ is monotonic decreasing in ξ, this ratio is bounded above by

$$\int dG(\xi)\ P\{g(w^*,\xi)\leq x\},$$

which is precisely $P\{g(w)\leq x\}$. This completes the induction.

[]

References

[1] D. Angluin and L. G. Valiant, Fast Probabilistic Algorithms for Hamiltonian Circuits and Matchings, Proc. 9th ACM Symp. on Theory of Computing, 1977, pp. 30-41.

[2] J. Beardwood, J. H. Halton, and J. M. Hammersley, The Shortest Path Through Many Points, Proc. Camb. Phil. Soc. 55 (1959), pp. 299-327.

[3] B. Bollabȧs and P. Erdős, Cliques in Random Graphs, Math. Proc. Camb. Phil. Soc. 80 (1976), pp. 419-427.

[4] A. A. Borovkov, A Probabilistic Formulation of Two Economic Problems, Soviet Mathematics 3:5 (1962), pp. 1403-1406.

[5] K. L. Chung, A Course in Probability Theory, 2nd ed., Academic Press, New York, 1976.

[6] F. N. David and D. E. Barton, Combinatorial Chance, Charles Griffin and Company Limited, London, 1962.

[7] H. A. David, Order Statistics, John Wiley and Sons, New York, 1970.

[8] W. E. Donath, Algorithm and Average-value Bounds for Assignment Problems, IBM J. Res. Dev. 13 (1969), pp. 380-386.

[9] P. Erdős and A. Rėnyi, On Random Graphs I, Publicationes Mathematicae 6 (1959), pp. 290-297.

[10] P. Erdős and A. Rėnyi, On the Evolution of Random Graphs, Publ. Math. Inst. Hung. Acad. Sci. 5A (1960), pp. 17-61.

[11] P. Erdős and A. Rėnyi, On Random Matrices, Publ. Math. Inst. Hung. Acad. Sci., 8A (1963), pp. 455-461.

[12] P. Erdős and A. Rėnyi, On the Existence of a Factor of Degree One of a Connected Random Graph, Acta Math. Acad. Sci. Hung. 17 (1966), pp. 359-368.

[13] P. Erdős and J. Spencer, Probabilistic Methods in Combinatorics, Academic Press, New York, 1974.

[14] W. Feller, An Introduction to Probability Theory and Its Applications, Vol. I, Third Edition, John Wiley and Sons, New York, 1968.

[15] M. L. Fisher and D. S. Hochbaum, Probabilistic Analysis of the Euclidean K-Median Problem, Technical report 78-06-03, Wharton Department of Decision Sciences, University of Pennsylvania, May 1978.

[16] M. R. Garey and D. S. Johnson, Computers and Intractibility: A Guide to the Theory of NP-Completeness, W. H. Freeman and Company, San Francisco, 1979.

[17] G. R. Grimmett and C. J. H. McDiarmid, On Coloring Random Graphs, Math. Proc. Camb. Phil. Soc. 77 (1975), pp. 313-324.

[18] L. J. Guibas and E. Szemeredi, The Analysis of Double Hashing, Proceedings of the Eighth Annual ACM Symposium on Theory of Computing, 1976, pp. 187-191.

[19] J. H. Halton and R. Terada, An Almost Surely Optimal Algorithm for the Euclidean Traveling Salesman Problem, Computer Sciences Technical Report #335, University of Wisconsin-Madison, October 1978.

[20] R. M. Karp, Reducibility among Combinatorial Problems, in Complexity of Computer Computations, R. E. Miller and J. W. Thatcher, eds., Plenum Press, N. Y., 1972, pp. 85-104.

[21] R. M. Karp, Probabilistic Analysis of Partitioning Algorithms for the Travelling-Salesman Problem in the Plane, Math. Op. Res. 2:3 (1977), pp. 209-224.

[22] G. S. Lueker, "Maximization Problems on Graphs with Edge Weights Chosen from a Normal Distribution," Technical Report 115, University of California at Irvine, March 1978. A condensed version appears in Proc. 10th Annual ACM Symp. on Theory of Computing, May 1978, pp. 13-18.

[23] D. W. Matula, On the Complete Subgraphs of a Random Graph, Proc. 2nd Chapel Hill Conference on Combinatorial Math. and its Applications, University of North Carolina, Chapel Hill, May, 1970, pp. 356-369.

[24] L. Pósa, Hamiltonian Circuits in Random Graphs, Disc. Mathematics 14 (1976), pp. 359-364.

[25] P. K. Sen, On Stochastic Convergence of the Sample Extreme Values from Distributions with Infinite Extremities, J. Ind. Soc. Agric. Statist. 16 (1964), 189-201.

[26] D. W. Walkup, Matchings in Random Regular Bipartite Graphs, draft, December, 1977.

[27] D. W. Walkup, On the Expected Value of a Random Assignment Problem, draft, December, 1977.

[28] B. W. Weide, Statistical Methods in Algorithm Design and Analysis, Ph.D. thesis, Department of Computer Science, Carnegie-Mellon University; available as technical report CMU-CS-78-142, August 1978.