UC Irvine
ICS Technical Reports

Title
Maximization problems on graphs with edge weights chosen from a normal distribution

Permalink
https://escholarship.org/uc/item/4m41306k

Author
Lueker, George S.

Publication Date
1978

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

MAXIMIZATION PROBLEMS ON GRAPHS WITH EDGE WEIGHTS CHOSEN FROM A NORMAL DISTRIBUTION

by

George S. Lueker[+]

Notice: This Material may be protected by Copyright Law (Title 17 U.S.C.)

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

Technical Report #115

An extended abstract of this report is to appear in the Proceedings of the Tenth Annual ACM Symposium on Theory of Computing (1978).

Keywords and phrases:

Random graphs
Optimization problems
Normal distribution
Weighted graphs
Probabilistic analysis
Traveling salesman problem
Cliques

+Supported by NSF Grant MCS77-04410.

Notice: This Material
may be protected
by Copyright Law
(Title 17 U.S.C.)

Z
699
C3
no.115

Maximization Problems on Graphs with Edge Weights Chosen from a Normal Distribution

Abstract

We consider optimization problems on complete graphs with edge weights chosen from identical but independent normal distributions. We show some very general techniques for obtaining upper and lower bounds on the asymptotic behavior of these problems. Often, but not always, these bounds are equal, enabling us to state the asymptotic behavior of the maximum. Problems in which the bounds are tight include finding the optimum traveling salesman tour, finding a minimum cost spanning tree, and finding a heaviest clique on k vertices. We then discuss some greedy heuristic algorithms for these problems.

1. Introduction

Many results have been proven about the properties of random graphs. Some of these [AV77, BE75, ER59, ER60, ER66, GM75, Ma70, Po76, Wa77a] deal with graphs constructed by letting edges be present or absent according to some distribution; one then tries to estimate the probability that a subgraph of a given type will be present. 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 random distribution; one then tries to find the heaviest (or lightest) subgraph of a given type. We will call such problems subgraph optimization problems. 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 maximum solution is proportional to $\sqrt{n}$ [BHH59]; a very efficient algorithm has been designed whose expected behavior is asymptotically optimal [Ka76]. The assignment problem for the case in which weights are chosen from a uniform distribution has been studied by several people [Do69, Ku62,

wa77b]; Donath [Do69] has also considered the case in which the edge weights have a value x in [0,1] with probability proportional to x^k for some k. In this paper we investigate the behavior of a number of optimization problems on complete graphs with edge weights chosen independently from a normal distribution.

Throughout this paper, $\mathbf{G}_n$ will be a random variable which is a complete, undirected, weighted, labeled graph on n vertices; we will assume the vertices are labeled 1,2,...,n. Weights are chosen, independently, from a normal distribution with mean 0 and variance 1. (All of the results proved in this paper can immediately be generalized to the case in which some other mean and variance are specified, provided these quantities are the same for all edges; we assume zero mean and unit variance to minimize notation.) G will denote some particular weighted complete graph. The weight of the edge connecting vertex v and w will be denoted d(v,w).

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 $\mathbf{G}_n$. All elements of S_n are assumed to have the same number of edges; call this number m_n. 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 maximize W(G,H); this maximum will be called $W_{max}(G)$. Note that, for example, if S_n is the set of the (n-1)!/2 cycles on n vertices, $W_{max}(G)$ gives the solution to the traveling salesman problem. We wish to investigate the expected behavior of $W_{max}(\mathbf{G}_n)$. (Often in an optimization problem, we wish to minimize some quantity; for uniformity, however, we will always assume that we are maximizing quantities. The results obtained here will, by symmetry, immediately carry over to the corresponding minimization problem.)

In section 2, we will discuss some simple but useful facts about normal distributions. In section 3, we will present a very general technique for obtaining upper bounds on the expected values of maximum solutions to such problems. The method used is to examine the tail of the distribution of total weights of a subgraph from S_n in $\mathbf{G}_n$. A

similar idea was used by Donath in obtaining a lower bound on the solution to the assignment problem over n by n matrices whose columns are random permutations of the integers 1 to n [Do69]. There it was viewed as an enumeration argument; here the random elements are drawn from a continuous distribution so the argument has a somewhat different flavor. Section 4 discusses a very general technique for obtaining lower bounds on these expected values; the method is to relate a subgraph optimization problem to the corresponding subgraph existence problem. (Walkup [Wa77b] independently expoited a similar relationship, in estimating the optimum solution to random assignment problems with edge weights chosen from a uniform distribution.) It turns out that combining the bounds of sections 3 and 4 often enables us to make rather precise statements about the asymptotic behavior of the expected maximum, as will be shown in section 5. In section 6 we will investigate the behavior of some simple algorithms for some of these problems.

2. Some facts about normal distributions

We shall often use random variables chosen from a normal distribution with mean 0 and variance 1. Such a variable will be called a unit normal variable. For convenience, we shall let f (respectively F) be the corresponding probability density function (respectively probability distribution function). Thus

$$f(x) = (2\pi)^{-1/2}\, e^{-x^2/2}$$

$$F(x) = \int_{-\infty}^{x} f(t)\, dt$$

We will often be making statements about the asymptotic behavior of functions. The following definitions will be assumed.

$$g(x) \sim h(x) \iff g(x)-h(x) = o(h(x))$$

$$g(x) \lesssim h(x) \iff \max(g(x)-h(x),0) = o(h(x))$$

$$g(x) \gtrsim h(x) \iff \max(h(x)-g(x),0) = o(h(x))$$

Note that

$g(x) \lesssim h(x)$ and $g(x) \gtrsim h(x) \iff g(x) \sim h(x)$.

Also 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.

The following few observations, which are well-known or easily established, are useful.

Fact 1. As $r \to 0$,

a) $f^{-1}(r) \sim \pm\sqrt{2 \log r^{-1}}$

b) $F^{-1}(r) \sim -\sqrt{2 \log r^{-1}}$

c) $F^{-1}(1-r) \sim \sqrt{2 \log r^{-1}}$

Proof sketch. Part (a) is easily established; parts (b) and (c) follow from (a) and well-known facts about the relationship between f and F. (See [Fe68, p. 175].) ∎

Fact 2. Let X be a unit normal variable and let A be an event which happens with probability p. Then as $p \to 0$,

a) $|E[X \mid A]| \lesssim \sqrt{2 \log p^{-1}}$.

b) $E[|X| \mid A] \lesssim \sqrt{2 \log p^{-1}}$.

Proof sketch. For part (a), clearly the quantity in question will be maximized if the event A has the form "$X \geq a$" for some a; it is not hard to show that even in this case the bound holds. A similar argument holds for part (b). ∎

Fact 3. Let X_{max} be a random variable which is formed by taking the maximum of n unit normal variables. Then $E[X_{max}] \sim \sqrt{2 \log n}$. Moreover, for any $\epsilon > 0$,

$$P\{X_{max} \leq \sqrt{2(1-\epsilon) \log n}\} \lesssim e^{-n^{\epsilon/2}}. \qquad (1)$$

Proof sketch. First we prove that (1) holds. Let $M(x) = P\{X_{max} \leq x\}$. Then $M(x) = [F(x)]^n$. Let $a = \sqrt{2(1-e)\log n}$. As $n \to \infty$ (and hence $a \to \infty$), we have

$$F(a) = 1 - a^{-1} f(a) \ (1 + O(a^{-2}))$$

(See [Fe68, p. 175].) Thus,

$$\begin{aligned} M(a) &\sim [1 - a^{-1} f(a)(1 + O(a^{-2}))]^n \\ &\leq e^{-n\, a^{-1} f(a) \ (1 + O(a^{-2}))} \\ &= e^{-n\, a^{-1} (2\pi)^{-1/2} e^{-(1-e) \log n} \ (1 + O(a^{-2}))} \\ &= e^{-a^{-1} (2\pi)^{-1/2} n^{e} \ (1 + O(a^{-2}))} \\ &\leq e^{-n^{e/2}} \end{aligned}$$

Next we establish the asymptotic behavior of $E[X_{max}]$. First we show $E[X_{max}] \gtrsim \sqrt{2 \log n}$. Choose any $e>0$. Let OK be the event that $X_{max} \geq \sqrt{2(1-e) \log n}$. Then we may write

$$E[X_{max}] = P\{OK\}\ E[X_{max} | OK] + P\{\text{not OK}\}\ E[X_{max} | \text{not OK}]. \qquad (2)$$

By the results of the preceeding paragraph, $P\{OK\} \to 1$; also clearly

$$E[X_{max} | OK] \geq \sqrt{2(1-e) \log n}.$$

Thus, the first term on the right of (2) is $\gtrsim \sqrt{2(1-e) \log n}$. Further, $E[X_{max} | \text{not OK}]$ is surely less than the sum of the expectations of the magnitudes of the n random variables. Using Fact 2, we see that the second term on the right of (2) is

$$O(P\{\text{not OK}\}\ n \sqrt{2 \log P^{-1}\{\text{not OK}\}}),$$

which goes to zero, since

$$P\{\text{not OK}\} \lesssim e^{-n^{e/2}}.$$

Thus (2) becomes

$$E[X_{max}] \gtrsim \sqrt{2\ (1-e) \log n}.$$

Since e was chosen arbitrarily, the desired result follows.

Next we show that

$$E[X_{max}] \leq \sqrt{2 \log n}. \tag{3}$$

Note that, by symmetry,

$$E[X_{max}] = E[X_{max} \mid X_1 = \max(X_1, X_2, \ldots, X_n)]$$
$$= E[X_1 \mid X_1 = \max(X_1, X_2, \ldots, X_n)].$$

But clearly X_1 is the maximum with probability n^{-1}, so (3) follows by Fact 2. ∎

3. An upper bound.

Let $M_n = |S_n|$; throughout this paper, we assume that $\lim_{n\to\infty} M_n = \infty$. For example, if we are dealing with the traveling salesman problem, then $M_n = (n-1)!/2$. (Note that since our graphs are labeled, we can distinguish the elements of S_n even though they are all isomorphic.) The result of this section will establish that $E[W_{max}]$ cannot be much less than $\sqrt{2 m_n \log M_n}$.

Let $\mathbf{H}_n$ be a new random variable which is formed by choosing an element of S_n; each element is chosen with equal probability. When both $\mathbf{H}_n$ and $\mathbf{G}_n$ appear in an expression, we assume they are chosen independently.

Theorem 1. $E[W_{max}(\mathbf{G}_n)] \geq \sqrt{2 m_n \log M_n}$.

Proof. We begin by showing that

$$E[W_{max}(\mathbf{G}_n)] =$$
$$E[W(\mathbf{G}_n, \mathbf{H}_n) \mid W(\mathbf{G}_n, \mathbf{H}_n) = W_{max}(\mathbf{G}_n)]. \tag{4}$$

To see this, note that

$$E[W(\mathbf{G}_n, \mathbf{H}_n) \mid W(\mathbf{G}_n, \mathbf{H}_n) = W_{max}(\mathbf{G}_n)]$$
$$= E[W_{max}(\mathbf{G}_n) \mid W(\mathbf{G}_n, \mathbf{H}_n) = W_{max}(\mathbf{G}_n)] \tag{5}$$

Now clearly for any G, at least one $H \in S_n$ satisfies $W(\mathbf{G}_n,H)=W_{max}(\mathbf{G}_n)$; on the other hand, the set of G for which more than one $H \in S_n$ is maximal forms a space of measure zero. Thus for almost any G, the probability of $W(G,\mathbf{H}_n)=W_{max}(G)$ is precisely M_n^{-1}. Since this is independent of G, it follows that the right side of (5) is simply $E[W_{max}(\mathbf{G}_n)]$.

For any H in S_n, $W(\mathbf{G}_n,H)$ is simply the sum of m_n unit normal variables; hence $W(\mathbf{G}_n,H)$ has a normal distribution with mean 0 and variance m_n. Since this is true for any H, $W(\mathbf{G}_n,\mathbf{H}_n)$ must have this same distribution. Now $W(\mathbf{G}_n,\mathbf{H}_n) = W_{max}(\mathbf{G}_n)$ is an event which, as noticed before, has probability M_n^{-1}; thus by (4) and the obvious generalization of Fact 2 to normal variables which have nonunit variance, the theorem is established. ∎

4. A lower bound.

In this section we obtain an upper bound on $E[W_{max}]$. Many results have been obtained which demonstrate that for sufficiently dense graphs, certain properties are very likely to occur. More formally, define a random variable $\mathbf{G}_{n,p_n}$ to be a graph on n vertices, where each edge is present with probability p_n, independently of the others. Given a real sequence p_n and a sequence S_n of classes of subgraphs, let Q_n be the probability that the graph $\mathbf{G}_{n,p_n}$ fails to contain an element of S_n. Then, for example, it is known [AV77, Po76] that for any i, we can choose a c large enough so

$$Q_n=O(n^{-i})$$

if S_n is the set of hamiltonian cycles on n vertices, and $p_n = (c \log n)/n$.

In this section, we establish a theorem which relates results of this form to the optimization problems we are considering.

Theorem 2. Suppose Q_n goes to zero rapidly enough so that

$$Q_n \sqrt{m_n \log Q_n^{-1}} = o(m_n \sqrt{\log p_n^{-1}})$$

then

$$E[W_{max}] \gtrsim m_n \sqrt{2 \log p_n^{-1}}$$

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

1. Let $a = F^{-1}(1-p_n)$, and let $\bar{H}$ be some fixed element of S_n.
2. Let E be the set of edges in G whose weight is greater than a.
3. Let H be any element of S_n all of whose edges are in E, and stop. If no such H can be found, go on to step 4.
4. Let $H = \bar{H}$.

Note that if this algorithm stops at step 3, we surely have

$$W(G,H) \geq a\, m_n,$$

so by Fact 1,

$$W(G,H) \gtrsim m_n \sqrt{2 \log p_n^{-1}}.$$

We must also consider the possibility that H is set to $\bar{H}$ in step 4, that is, that no element of S_n can be constructed from the edges in E; call this event FAIL. Now note that distribution of graphs obtained by choosing all edges of weight greater than $F^{-1}(1 - p_n)$ is identical to $\mathbf{G}_{n,p_n}$. Thus the probability of FAIL is just Q_n. By Fact 2, and the fact that $W(\mathbf{G}_n,\bar{H})$ is normally distributed with variance m_n, we may conclude that the expected weight of $\bar{H}$ in the event FAIL is $O(\sqrt{m_n \log Q_n^{-1}})$. Then by the hypotheses of the theorem, the error we commit by ignoring the possibility of event FAIL is negligible. ∎

5. Some examples.

We now consider several examples of the applications of Theorems 1 and 2. We shall often observe a happy occurence--the upper and lower bounds coincide, enabling us to determine the asymptotic behavior of the maximum solution.

We begin by considering the traveling salesman problem; $W_{max}^{TSP}(G)$ will denote the maximum weight traveling salesman tour in a weighted graph G. Theorem 1 easily gives an asymptotic upper bound of $n\sqrt{2 \log n}$. On the other hand, it is known [AV77, Po76] that the probability that $\mathbf{G}_{n,p_n}$ fails to contain a hamiltonian cycle can be made to be $O(n^{-\alpha})$, for any α, by letting $p_n = (c \log n)/n$, where c is large enough. Thus, by Theorem 2,

$$E[W_{max}^{TSP}(\mathbf{G}_n)] \gtrsim n\sqrt{2 \log ((c \log n)/n)^{-1}}$$
$$\sim n\sqrt{2 \log n}.$$

Thus we easily obtain

Corollary 1. The expected maximum for the traveling salesman problem is given by $E[W_{max}^{TSP}(\mathbf{G}_n)] \sim n\sqrt{2 \log n}$.

A very similar result is very easily established for the expected weight of the maximum cost spanning tree in G, denoted $W_{max}^{ST}(G)$.

Corollary 2. $E[W_{max}^{ST}(\mathbf{G}_n)] \sim n\sqrt{2 \log n}$.

Actually, the examples considered so far have not fully tested the power of Theorems 1 and 2. For both the TSP and the maximum cost spanning tree problem, the upper bound could have been obtained by simply calculating the expected total weight of the heaviest n or n-1 edges. Also, as we shall see in Section 6, the lower bound can be achieved by a very simple algorithm. The bounds achieved in the next example do not appear to be obtainable by such simple arguments. Consider the problem of finding the weight of the heaviest k-clique in a graph G, denoted $W_{max}^{CLIQ(k)}(G)$. (In the asymptotic statements which follow, we assume that both k and n tend to infinity, but n goes to infinity much faster than k.) First note that the number of edges in a k-clique is $\sim k^2/2$. Further, the number of distinct k-cliques is C(n,k). Thus an upper bound is

$$E[W_{max}^{CLIQ(k)}(\mathbf{G}_n)] \lesssim \sqrt{2(k^2/2) \log C(n,k)}$$
$$\sim k^{3/2}\sqrt{\log n}. \qquad (6)$$

Now the proof of Theorem 1 in [GM75] can easily be adapted to show that if we let $p_n = n^{-2/(k+2)}$, then the probability that $\mathbf{G}_{n,p_n}$ fails to contain a k-clique is $O(n^{-3/2})$. Thus Theorem 2 gives

$$E[W_{max}^{CLIQ(k)}(\mathbf{G}_n)] \gtrsim (k^2/2)\sqrt{2 \log n^{2/(k+2)}}$$
$$\sim (k^2/2)\sqrt{2\,(2/(k+2))\log n}$$
$$\sim k^{3/2}\sqrt{\log n} \qquad (7)$$

Combining (6) with (7), we get the following.

Corollary 3:

$$E[W_{max}^{CLIQ(k)}(\mathbf{G}_n)] \sim k^{3/2}\sqrt{\log n}.$$

Finally, we give an example in which the upper and lower bounds guaranteed by Theorems 1 and 2 do not coincide. We will say a graph H has property X(k) if

a) H has 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). Also, let C(a,b) denote the number of combinations of a things taken b at a time.

First consider the upper bound of Theorem 1. To obtain a simple lower bound on M_n, imagine we first select a fixed set of k vertices to be the clique; now choose the remaining k^2-C(k,2) edges in any way from the remaining n-k vertices. This gives a lower bound of

$$\log M_n \gtrsim \log C(C(n-k,2),k^2-C(k,2))$$
$$\gtrsim k^2 \log n$$

On the other hand, M_n is surely less than the number of ways of choosing k vertices times the number of ways of choosing any k^2-C(k,2) edges joining the n vertices. Thus an upper bound is

$$\log M_n \lesssim \log C(n,k) + \log C(C(n,2),k^2-C(k,2))$$
$$\sim k \log n + k^2 \log n$$
$$\sim k^2 \log n.$$

Combining these upper and lower bounds, we obtain

$$\log M_n \sim k^2 \log n$$

The upper bound of Theorem 1 becomes

$$E[W_{max}^{X(k)}(\mathbf{G}_n)] \lesssim \sqrt{k^2\, 2\, k^2 \log n}$$
$$= k^2 \sqrt{2 \log n}. \qquad (8)$$

Now consider the lower bound provided by Theorem 2. In order to get as strong a result from this theorem as possible, we wish to let p_n go to $\emptyset$ as fast as possible. We will show that even if we let it go to $\emptyset$ too fast, the bound is not tight. In particular, if we let p_n be as small as $n^{-2(1+e)/k}$, for any $e>\emptyset$, then the probability that $\mathbf{G}_{n,p_n}$ has a subgraph with property X(k) goes to zero, since the probability of having a clique on k vertices goes to zero. (Again, this is an easy consequence of the proof in [GM75].) Thus we are letting p_n go to zero faster than the conditions of the theorem allow. Even with this choice of p_n, however, Theorem 2 gives a lower bound of

$$k^2 \sqrt{2 \log n^{2(1+e)/k}}$$
$$= 2\, k^{3/2} \sqrt{\log n\ (1+e)}$$

Letting e go to zero would give a lower bound of

$$E[W_{max}^{X(k)}(\mathbf{G}_n)] \gtrsim 2\, k^{3/2} \sqrt{\log n} \qquad (9)$$

Note that the upper and lower bounds do not coincide. In fact, we can show that neither is tight, by determining $E[W_{max}^{X(k)}(\mathbf{G}_n)]$. Let $W_{max}^{HEAV(k)}(G)$ be the weight of the heaviest k^2-C(k,2) edges in G. It is not hard to see that

$$E[W_{max}^{HEAV(k)}(\mathbf{G}_n)] \sim (k^2/2) \sqrt{2 \log n} \qquad (1\emptyset)$$

and

$$E[W_{max}^{X(k)}(\mathbf{G}_n)] \gtrsim E[W_{max}^{HEAV(k)}(\mathbf{G}_n)]. \qquad (11)$$

On the other hand, an upper bound is

$$E[W_{max}^{X(k)}(\mathbf{G}_n)] \lesssim E[W_{max}^{CLIQ(k)}(\mathbf{G}_n)] + E[W_{max}^{HEAV(k)}(\mathbf{G}_n)] \qquad (12)$$

since by choosing the clique and extra edges independently we can certainly do as well as when we must avoid duplication of edges. Combining (12), (10), and Corollary 3 we obtain

$$E[W_{max}^{X(k)}] \leq (k^2/2)\sqrt{2 \log n} + k^{3/2}\sqrt{\log n}$$
$$\sim (k^2/2)\sqrt{2 \log n} \qquad (13)$$

Then (10), (11), and (13) give

$$E[W_{max}^{X(k)}] \sim (k^2/2)\sqrt{2 \log n}.$$

Comparing this with the bounds in (8) and (9), we conclude that neither bound is tight.

Thus there are sets S_n such that the bounds of Theorem 1 and Theorem 2 are not asymptotically tight.

6. Some algorithms.

In this section we investigate the expected behavior of some simple heuristic algorithms for the traveling salesman problem and the heaviest clique problem. Since both of the corresponding subgraph existence problems are NP-complete [Ka72], it is likely that there is no efficient algorithm which produces exact answers all of the time. Nonetheless, we shall see that some simple fast algorithms have average behavior which is close to the average behavior of the optimum. (Since a fast exact algorithm for the spanning tree problem is well-known [Kr56], we will not discuss heuristic approaches for it.)

It is not hard to show that a very simple greedy algorithm for the TSP, which constructs a tour by starting at an arbitrary vertex and iteratively walking to the closest unvisited vertex, achieves the expected asymptotic behavior described in Corollary 1.

The problem of finding the heaviest clique on k vertices is considerably more interesting. Consider the following greedy approach.

```
procedure CLIQ_GREEDY(G);
   begin
      C <- {an arbitrary vertex of G};
      while |C| < k do
         begin
            let v be the vertex not in C
              which maximizes the sum, over
              all w in C, of d(v,w);
            add v to C;
         end
      return the total weight of all edges
        joining vertices of C;
   end;
```

Lemma 1:

$$E[\text{CLIQ\_GREEDY}(\mathbf{G}_n)] \geq \sqrt{8/9}\, k^{3/2} \sqrt{\log n}.$$

Proof. Note that if we consider, for some fixed i, the i^{th} pass through the loop, we are choosing the maximum of n-i sums of i unit normal variables. Unfortunately, the i^{th} pass through the loop is affected by the previous passes, which complicates the analysis somewhat. However, an idea similar to that used in [ES74,GS76] is useful here--we can simply eliminate all cases in which things don't work out as we like. More formally, let us choose an $\epsilon > 0$ and consider the probability that for any set C of vertices, $|C|<k$,

$$\max_{v \notin C} \sum_{w \in C} d(v,w) \leq \sqrt{2\,|C|\,(1-\epsilon)\,\log\,(n-|C|)}$$

Using Fact 3, 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; call this probability P(n,k). (Recall that we are letting n go to infinity much faster than k.) Accordingly, we conclude that the algorithm produces a clique of weight at least

$$\sum_{i=1}^{k-1} \sqrt{2\,i\,(1-\epsilon)\,\log\,(n-i)}$$

$$\sim \sqrt{8/9}\, k^{3/2} \sqrt{(1-e) \log n}$$

except with probability $P(n,k)$. But since $P(n,k)$ vanishes so rapidly an argument like that in the proof of Theorem 2 tells us that the result is

$$E[\text{CLIQ\_GREEDY}(\mathbf{G}_n)]$$
$$\gtrsim \sqrt{8/9}\, k^{3/2} \sqrt{(1 - e) \log n}.$$

Since e may be arbitrarily small, we conclude that

$$E[\text{CLIQ\_GREEDY}(\mathbf{G}_n)] \gtrsim \sqrt{8/9}\, k^{3/2} \sqrt{\log n}.$$ ∎

Next we will show that this is in fact a tight description of the behavior of the algorithm. The following lemma will be useful.

Lemma 2. Let $\mathbf{V}_m$ 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

$$V_1 \leq V_2 \Longrightarrow g(V_1) \leq g(V_2).$$

(V_1 is said here to be less than or equal to V_2 if the inequality holds componentwise.) Finally, let B be an r by m matrix of nonegative reals, and b be a column vector of r reals. Then

$$E[g(\mathbf{V}_m) | B\, \mathbf{V}_m \leq b] \leq E[g(\mathbf{V}_m)].$$

Proof. We prove the lemma by induction on m. For m=∅ it is trivial. Suppose it holds for m=k-1. We may decompose $\mathbf{V}_k$ as

$$\mathbf{V}_k = \mathbf{V}_{k-1} \,||\, \mathbf{V},$$

where || denotes concatenation, $\mathbf{V}_{k-1}$ is the first k-1 components of $\mathbf{V}_k$, and $\mathbf{V}$ is the last component of $\mathbf{V}_k$. Then

$$E[g(\mathbf{V}_k) \mid B\, \mathbf{V}_k \leq b]$$
$$= E[g(\mathbf{V}_{k-1} \,||\, \mathbf{V}) \mid B_1\, \mathbf{V}_{k-1} + B_2\, \mathbf{V} \leq b] \qquad (14)$$

where B_1 and B_2 are appropriate submatrices of B. Let G' be the distribution function of $\mathbf{V}$. Then we may write the right hand side of

(14) as

$$\frac{\int dG'(x)\ E[g(\mathbf{V}_{k-1}||x) \mid B_1\ \mathbf{V}_{k-1} \le b - B_2\ x]\ h(x)}{\int dG'(x)\ h(x)}$$

where

$$h(x) = P\{B_1\ \mathbf{V}_{k-1} \le b - B_2\ x\}.$$

Now by the inductive hypothesis, for any x,

$$E[g(\mathbf{V}_{k-1}\ ||\ x) \mid B_1\ \mathbf{V}_{k-1} \le b - B_2\ x] \le E[g(\mathbf{V}_{k-1}||x)]$$

Thus an upper bound is

$$\frac{\int dG'(x)\ h(x)\ E[g(\mathbf{V}_{k-1}||x)]}{\int dG'(x)\ h(x)}$$

But since h(x) is easily seen to be monotonic decreasing, while $E[g(\mathbf{V}_{k-1}||x)]$ is monotonic increasing, this ratio is bounded above by

$$\int dG'(x)\ E[g(\mathbf{V}_{k-1}||x)],$$

which is precisely $E[g(\mathbf{V}_k)]$. This completes the induction. ∎

Lemma 3:

$$E[\text{CLIQ\_GREEDY}(\mathbf{G}_n)] \le \sqrt{8/9}\, k^{3/2} \sqrt{\log n}. \qquad (15)$$

Proof. Note that this lemma asserts that the lower bound of Lemma 1 is also an upper bound. If, on any given pass through the main loop of the algorithm, the edge weight probabilities were not conditioned by previous passes, it would be a simple matter to analyze the expected weight of the new edges added to the clique. We begin the proof by showing that the conditioning of edge weights which has taken place can only hurt the average behavior of the algorithm.

Assume that the vertices of G are labeled $v_1, v_2, \ldots, v_n$. Suppose we have completed r-1 iterations. Then |C| = r. If L is a list of r vertices, let A(L) be the event that L contains the vertices of C, in the order in which they were added to C, and that no ties were present

during the selection of maxima. For the time being, consider the case

$$L = v_1, v_2, \ldots, v_r.$$

Then to determine the contribution of the next vertex to the total clique weight, we must evaluate

$$E[\max_{r<i\leq n} \sum_{j=1}^{r} d(v_i, v_j) \mid A(L)]. \tag{16}$$

Now since the choice of vertices to add to C is determined by comparisons of sums of edge weights, the event A(L) can be phrased as a set of inequalities on the edge weights. In particular, the inequalities which must be satisfied are

$\forall m$ with $1 \leq m \leq r-1$,

$\forall i$ with $m+2 \leq i \leq n$,

$$\sum_{j=1}^{m} d(v_{m+1}, v_j) > \sum_{j=1}^{m} d(v_i, v_j) \tag{17}$$

If G is the weighted graph on n vertices, let V1(G) be a vector which contains the weights of edges joining vertices numbered r or lower, in some arbitrary order; let V2(G) be a vector which contains the remaining edge weights. We will use $\mathbf{V}1$ (respectively $\mathbf{V}2$) as an abbreviation for $V1(\mathbf{G}_n)$ (respectively $V2(\mathbf{G}_n)$).

Note that we can find matrices B_1 and B_2, with all elements of B_2 positive, such that (17) can be written as

$$B_2 \, V2(G) < B_1 \, V1(G)$$

Thus if we let

$$g(V2(G)) = \max_{r<i\leq n} \sum_{j=1}^{r} d(v_i, v_j),$$

we may rewrite (16) as

$$E[g(\mathbf{V}2) \mid B_2 \, \mathbf{V}2 < B_1 \, \mathbf{V}1] \tag{18}$$

By an application of Lemma 2, this can be seen to be bounded above by $E[g(\mathbf{V}2)]$. But this is just the expected value of the maximum of n-r independent sums of r unit normal variables, so

$$E[g(\mathbf{V}2)] \sim \sqrt{2 \, r \log n}.$$

Thus far we have shown that

$$E[\max_{r \le i \le n} \sum_{j=1}^{r} d(v_i,v_j) \mid A(L)] \lesssim \sqrt{2\, r \log n}$$

Now by symmetry, the choice of L does not affect the analysis; moreover, the events A(L), over all possible L, together with the space of measure $\emptyset$ in which ties are present during the selection of maxima, exactly cover the entire probability space. Thus, the expected weight of the set of edges added at the r^{th} iteration is at most asymptotic to $\sqrt{2\, r \log n}$. Summing from r equalling 1 to k-1, we obtain the lemma. ∎

Theorem 3:

$$E[\text{CLIQ\_GREEDY}] \sim \sqrt{8/9}\, k^{3/2} \sqrt{\log n}.$$

Proof. This follows immediately from Lemmas 1 and 3. ∎

Thus the algorithm has an average result which is less than 6 percent away from the average optimum.

Conclusion

We have examined the problem of finding a heaviest instance of a subgraph, from a given set, in an n-vertex weighted complete graph. Assuming that the edge weights are chosen from a normal distribution, we have investigated the expected behavior of this optimization problem. Two very general theorems were discussed, one of which gave an upper bound and one of which gave a lower bound. In a number of interesting problems, these bounds turned out to be tight, enabling us to state the asymptotic behavior of the optimum; these problems included the traveling salesman problem, the maximum weight spanning tree problem, and the problem of finding the heaviest clique on k vertices, where $1 \ll k \ll n$. However, an example showed that in some cases neither bound was tight; it would be interesting to obtain good sufficient conditions under which either bound was tight.

Next some greedy approximation algorithms were discussed. For the travelling salesman problem, a very simple greedy algorithm gave results whose average was the same as the average optimum. For the clique

problem, a simple greedy algorithm was analyzed and found to produce an average result which was about 6 percent lower than the average optimum. It would be interesting to find an algorithm with even better average behavior.

References

[AV77] Dana Angluin and Leslie G. Valiant, "Fast Probabilistic Algorithms for Hamiltonian Circuits and Matchings," Proceedings of the Ninth Annual ACM Symposium on Theory of Computing, 1977, pp. 30-41.

[BHH59] Jillian Beardwood, J. H. Halton, and J. M. Hammersley, "The Shortest Path Through Many Points," Proceedings Camb. Phil. Society 55 (1959), pp. 299-327.

[BE75] B. Bollabás and P. Erdös, "Cliques in Random Graphs," Math. Proc. Camb. Phil. Soc. 80 (1976), pp. 419-427.

[Do69] W. E. Donath, "Algorithm and Average-value Bounds for Assignment Problems," IBM J. Res. Develop. 13 (1969), pp. 380-386.

[ER59] P. Erdös and A. Rényi, "On Random Graphs I," Publicationes Mathematicae 6 (1959), pp. 290-297.

[ER60] P. Erdös and A. Rényi, "On the Evolution of Random Graphs," Publ. Math. Inst. Hung. Acad. Sci. 5A (1960), pp. 17-61.

[ER66] P. Erdös and A. Rényi, "On the Existence of a Factor of Degree One of a Connected Random Graph," Acta Math. Acad. Sci. Hung. 17 (1966), pp. 359-368.

[ES74] P. Erdös and J. Spencer, Probabilistic Methods in Combinatorics, Academic Press, New York, 1974.

[Fe68] William Feller, An Introduction to Probability Theory and Its Applications, Vol. I, Third Edition, John Wiley and Sons, New York, 1968.

[GM75] G. R. Grimmett and C. J. H. McDiarmid, "On Coloring Random Graphs," Math. Proc. Camb. Phil. Soc. 77 (1975), pp. 313-324.

[GS76] Leo J. Guibas and Endre Szemeredi, "The Analysis of Double Hashing," Proceedings of the Eighth Annual ACM Symposium on Theory of Computing, 1976, pp. 187-191.

[Ka72] Richard 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.

[Ka76] Richard M. Karp, "The Probabilistic Analysis of Some Combinatorial Search Algorithms," Algorithms and Complexity: New Directions and Recent Results, J. F. Traub, ed., Academic Press, New York, 1976, pp. 1-19.

[Kr56] J. B. Kruskal, Jr., "On the Shortest Spanning Subtree of a Graph and the Traveling Salesman Problem," Proc. Amer. Math. Soc. 7 (1956), pp. 48-50.

[Ku62] Jerome M. Kurtzberg, "On Approximation Methods for the Assignment Problem," JACM 9 (1962), pp. 419-439.

[Ma70] David 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.

[Po76] L. Pósa, "Hamiltonian Circuits in Random Graphs," Discrete Mathematics 14 (1976), pp. 359-364.

[Wa77a] David W. Walkup, "Matchings in Random Regular Bipartite Graphs," draft, December, 1977.

[Wa77b] David W. Walkup, "On the Expected Value of a Random Assignment Problem," draft, December, 1977.