The Steiner Problem in the Hypercube

Zevi Miller
Department of Mathematics and Statistics
Miami University
Oxford, Ohio 45056

Manley Perkel
Department of Mathematics and Statistics
Wright State University
Dayton, Ohio 45435

Let $Q(n)$ be the n-dimensional hypercube, and X, a set of points in $Q(n)$. The Steiner problem for the hypercube is to find the smallest possible number $L(n,X)$ of edges in any subtree of $Q(n)$ that spans X. We obtain the following results:

(1) An exact formula for $L(n,X)$, when $|X| \leq 5$.
(2) The bound $L(n,X) \leq \binom{n}{k+1} + (2 + o(1))\,([\log(k)]/k)\binom{n}{k}$ as $k \to \infty$, when X is the set of all points in $Q(n)$ of a given weight $k + 1$, provided $(k^2/[\log(k)])^{1+1/k} \leq n$.
(3) NP-completeness of deciding $L(n,X)$ even when every point of X has weight at most 2.

1. INTRODUCTION

Let $G = (V, E)$ be an undirected graph with vertex set V and edge set E, and let $X \subseteq V$ be a subset of the vertices of G. The Steiner problem for G is to find the minimum number of edges in any subtree of G that contains X among its vertices. Any subtree of G achieving this minimum will be called a *Steiner tree for* X. Apart from its intrinsic graph theoretic interest, the Steiner problem is motivated by various layout problems in VLSI design and in communication networks and by the construction of phylogenetic trees.

This problem has been considered for various classes of graphs G (see the survey [10]). We mention, in particular, the grid, where NP-completeness was proved [6] and polynomial algorithms for special sets X were found [1].

The related rectilinear problem is as follows: Given a set X of points in the plane, find the shortest rectilinear Steiner tree for X, that is, the shortest tree T in the plane such that $V(T) \supset X$ and each edge of T is either vertical or horizontal. It was shown in [7] that $t_s/t_m \geq 2/3$, where t_s is the shortest rectilin-

NETWORKS, Vol. 22 (1992) 1–19

CCC 0028-3045/92/01001-19$04.00

ear Steiner tree and t_m is the shortest rectilinear spanning tree (i.e., one whose vertex set is identical to X).

In this work, we study the graph Steiner problem in the case when G is the n-dimensional cube $Q(n)$, motivated by the importance of this graph and its derivatives (the butterfly and shuffle exchange networks) as parallel architectures and by a connection with phylogeny. The graph $Q(n)$ is defined as follows: The vertex set of $Q(n)$ is the set of all 2^n n-tuples of zeros and ones (viewed as strings of 0's and 1's), i.e., the vector space $\Omega = \mathbf{Z}_2^n$ (where $\mathbf{Z}_2$ denotes the field of two elements), and where two vertices are adjacent if and only if the corresponding n-tuples differ in exactly one coordinate position. If x is a vertex of $Q(n)$, we will denote by $x(i) \in \{0,1\}$ the ith coordinate of x.

The relation between the Steiner problem and phylogenetic trees is made clear in the case when $Q(n)$ is the underlying graph. Here, we consider a vector of 0's and 1's [i.e., a vertex of $Q(n)$] as a description of some individual—perhaps a genetic string in which each entry may take on one of two possible values. Then, a set of individuals may be viewed as a subset X of $Q(n)$. A rooted Steiner tree for X is then a possible explanation of how these individuals are related and how they evolved from a common ancestor (the root). Here, each edge of the tree represents an evolutionary change in exactly one of the the n entries. An overview of the literature on this subject is given in the forthcoming survey [8].

Throughout this article, if x and y are in Ω, then we will denote by $d(x,y)$ the distance from x to y in $Q(n)$, i.e., $d(x,y)$ is the Hamming distance function, being the number of coordinates in which x and y differ. For convenience, and when unlikely to cause confusion, we will denote by 0 the vertex 00 . . . 0 and by 1 the vertex 11 . . . 1. Define the weight of x to be $wt(x) = d(x,0) =$ the number of nonzero coordinates of x. So $d(x,y) = wt(x - y)$ and x is adjacent to y in $Q(n)$ if and only if $d(x,y) = 1$, that is, if and only if $wt(x - y) = 1$. It is easily checked that $Q(n)$ is regular of degree n. Further, the diameter of $Q(n)$ is n, and given any vertex x, there is a unique vertex y with $d(x,y) = n$ (we call x and y *antipodal* vertices).

We may also define $Q(n)$ inductively by letting $Q(0)$ be a single vertex, and then $Q(n)$ is obtained by taking two copies of $Q(n - 1)$ and joining corresponding vertices.

The rest of the article is divided into three parts: The first gives some general upper bounds for $L(n,X)$ and exact formulas for the same when $|X| \leq 5$. The second uses a beautiful result of Frankl and Rodl [5] on the generalized Turan problem to give an upper bound for $L(n,X)$, when X is the set of all points of some fixed weight in $Q(n)$. Finally, in the third part, we investigate the complexity of the problem by showing that even the restricted subproblem of determining $L(n,X)$ when each point of X has weight at most 2 is already NP-complete.

2. GENERAL UPPER BOUNDS AND EXACT RESULTS

Given a set X of vertices in $Q(n)$, denote by $T(X)$ any Steiner tree for X, and let $L(n,X)$ be the number of edges in $T(X)$ [naturally $L(n,X)$ is independent of which Steiner tree for X we pick].

The following basic facts concerning $T(X)$ are clear, so their proofs will be omitted.

Lemma 2.1.

(i) The leaves of $T(X)$ are all elements of X.
(ii) $L(n,X) \leq 2^n - 1$.
(iii) Let x be in X, and let S be a union of (nonempty) connected components of $T(X) - x$. Let Y be the set of vertices of X contained in S. Then, $S + x$ is a Steiner tree of $Y \cup x$.

Let $L(n,k) = \max\{L(n,X) : X \subseteq Q(n), |X| = k\}$. Clearly, $L(n,2) = n$, realized by letting $X = \{x_1x_2\}$ be a pair of antipodal vertices. The case $k = 3$ is covered in the following theorem:

Theorem 2.1. Let $X \subseteq Q(n)$, with $|X| = 3$, and denote by $r = r(X)$, the number of coordinate positions in which all three elements of X are equal (either all equal 0 or all equal 1). Then, $L(n,X) = n - r$ and $L(n,3) = n$.

Proof. Let $X = \{x_1,x_2,x_3\}$. We first define the *centroid* c of the set X. For each $j = 1,2, \ldots, n$, let $c(j) = 0$ if and only if two or more of $x_1(j), x_2(j), x_3(j)$ are 0, and let $c(j) = 1$ otherwise. Now define the point c by $c = c(1)c(2) \ldots c(n)$. It is then clear that in any coordinate position, c differs from none of the x_i, $i = 1$, 2, or 3, in those coordinate positions where they are all equal and differs from exactly one of the x_i otherwise. Thus, $\sum_{i=1}^{3} d(c,x_i) = n - r$, and so $L(n,X) \leq n - r$.

Consider the opposite inequality. Since each leaf of $T(X)$ must be an element of X, it is clear that $T(X)$ can have no vertex of degree [in $T(X)$] greater than 3 and can have at most one vertex of degree exactly 3. If $T(X)$ has precisely one vertex d of degree 3, then it is clear that all other interior vertices of $T(X)$ have degree 2 and each vertex of X must be a leaf of $T(X)$. Also, d must differ from at least one of the x_i in any coordinate position in which the x_i are not all equal. Hence, $L(n,X) \geq n - r$. If the tree has no vertex of degree 3, then it is a path, and again an endpoint of this path must differ from at least one of the other two x_i's in any coordinate position in which the x_i are not all equal. Thus, again, we get $L(n,X) \geq n - r$, and, hence, $L(n,X) = n - r$.

Now from above, $L(n,3) \leq n$. Let $X = \{0,1,x\}$, where $x \neq 0$ or 1. Then, $L(n,X) = n$, so that, in fact, $L(n,3) = n$, completing the proof. ■

For arbitrary k and $X = \{x_1,x_2, \ldots, x_k\}$, the definition of the centroid in the proof of Theorem 2.1 can, of course, be extended by taking $c(j)$ (for $j = 1, \ldots, n$) to be 0 if and only if 0 occurs as the jth coordinate of a majority (at least half) of the x_i (for $i = 1, \ldots, k$) and taking $c(j)$ to be 1 otherwise. This leads to the following:

The Centroid Upper Bound (CUB): $L(n,k) \leq \min\{\lfloor k/2 \rfloor n, 2^n - 1\}$.

The CUB turns out, however, not to be best possible, except for the cases $k = 2$ and 3. However, we can use the CUB with the centroid idea to get a

better upper bound as follows: Let $X = \{x_1, x_2, \ldots, x_k\}$ with k even. Let $Y = \{x_1, x_2, \ldots, x_{k-1}\}$, let c be the centroid of Y, and let T be the tree on Y obtained by using the centroid construction, so that $|T| \leq [(k-2)/2]n$. Then, it must be the case that for some subset Z of Y of cardinality $k-2$ we have $\Sigma_{z \in Z} d(c,z) \leq [(k-2)/(k-1)][(k-2)/2]n$. Without loss of generality, let $Z = \{x_1, x_2, \ldots, x_{k-2}\}$. Now, if $T(x_{k-2}, x_{k-1}, x_k)$ denotes the Steiner tree on $\{x_{k-2}, x_{k-1}, x_k\}$ (of size $\leq n$, by Theorem 2.1), then attaching T to $T(x_{k-2}, x_{k-1}, x_k)$ and deleting all the edges from c to x_{k-1} gives the following:

The Modified Centroid Upper Bound (MCUB):

$$L(n,k) \leq \min\left\{\left\lfloor \frac{k}{2} \right\rfloor n,\ 2^n - 1\right\}, \text{ for } k \text{ odd (the CUB), and}$$

$$L(n,k) \leq \min\left\{\left\lceil k - 1 + \frac{1}{k-1} \right\rceil \cdot \frac{n}{2},\ 2^n - 1\right\}, \text{ for } k \text{ even.}$$

[*Note:* When k is odd, breaking the set X up into a sequence of $\lfloor k/2 \rfloor$ subsets of size three, each intersecting the next in a single vertex, and then using $L(n,3) \leq n$, leads again to the upper bound given in the CUB for k odd.]

For $k = 4$, the MCUB gives $L(n,4) \leq \lfloor (5/3)n \rfloor$. For example, with $n = 3$, the set $X = \{0, a = 011, b = 101, c = 110]$ satisfies $L(3,X) = 5$, whence $L(3,4) = 5$. In fact, the MCUB upper bound is exact for $k = 4$, i.e., $L(n,4) = \lfloor (5/3)n \rfloor$, the extremal set being obtained as follows:

Let A, B, and C be subsets of the coordinate set $N = \{1,2, \ldots, n\}$ that partition N, and such that $|A| = |B| = \lceil n/3 \rceil$, and $|C| = n - 2\lceil n/3 \rceil$. Let $X = \{0, a = 0^A 1^B 1^C, b = 1^A 0^B 1^C, c = 1^A 1^B 0^C\}$, where, if $\varepsilon = 0$ or 1 and M is a subset of N, then ε^M means place the symbol ε in all the coordinate positions determined by M. We will show later that $L(n,X) = \lfloor (5/3)n \rfloor$.

Before considering $k > 3$ in more detail, we have the following definitions and results. Let T be a connected subtree of $Q(n)$. Any vertex of T whose degree (in T) is at least 3 will be called a *branch point* of T. Let x be a leaf of T, and y, the closest branch point of T (if any) to x. If there is such a branch point, then the smallest connected subgraph of T containing y and all leaves of T whose closest branch point is also y will be called the *cluster determined by* y. Thus, the cluster determined by y can be viewed as a subdivision of a star with central point y, that is, subgraph that can be obtained from a star (with center y) by, if necessary, inserting points of degree 2 along edges of the star. The endpoints of the cluster are precisely those endpoints of T having y as their closest branch point in T.

Lemma 2.2. Let $X \subseteq Q(n)$, and $T(X)$, a Steiner tree on X. Let $Y = X \cup \{\text{branch points of } T\}$. Let $x, y \in Y$ and such that the path P in $T(X)$ from x to y contains no vertex in Y other than x and y.

(i) If for some $1 \leq i \leq n$, $x(i) = y(i)$, then $z(i) = x(i) = y(i)$ for all vertices z in P.

(ii) If for some $1 \leq i \leq n$, $x(i) \neq y(i)$, then we may either assume that $z(i) \leftarrow x(i)$ for all vertices of P [other than $y(i)$] or $z(i) = y(i)$ for all vertices z of P [other than $x(i)$].

Proof.

(i) Without loss of generality, suppose $i = n$, and $x(n) = y(n) = 0$. Now suppose that some vertex in P has its nth coordinate equal to 1. Since $x(n) = 0$, there is some pair of adjacent vertices z_1 and z_2, such that $z_1(n) = 0$ and $z_2(n) = 1$. Since $d(z_1,z_2) = 1$, we can write $z_1 = w0$ and $z_2 = w1$, where w is some string of $n - 1$ zeros and ones. Thus, the path P' obtained from P by changing the nth coordinate of every vertex in P to a 0 (and deleting duplications) is a path in $Q(n)$ from x to y [since $y(n) = 0$] and $|P'| < |P|$. Replacing the path P in $T(X)$ by P' then leads to a tree T' on X containing fewer vertices than $T(X)$, a contradiction.

(ii) Again without loss of generality, suppose $i = n$, $x(n) = 0$, and $y(n) = 1$. In a manner similar to the proof above, we may assume that if $P = (z_1, z_2, \ldots, z_p)$, where $x = z_1$ and $y = z_p$, then for some $m < p$, $z_j(n) = 0$ for all $j \leq m$, and $z_j(n) = 1$ for all $j > m$. Then, $z_m(n) = w0$ and $z_{m+1}(n) = w1$, where w is a string of $n - 1$ zeros and ones. We may assume $2 \leq m \leq p - 2$ or else we are done.

Let $P' = (z_1,z_2, \ldots, z_m, z'_{m+2}, \ldots, z'_p, y)$, where for $j \geq m + 2$, z'_j denotes z_j, with its nth coordinate changed to 0. Let $P'' = (x,z'_1,z'_2, \ldots, z'_{m-1}, z_{m+1}, \ldots, z_p)$, where, for $j \leq m - 1$, z'_j denotes z_j, with its nth coordinate changed to 1. Then $|P'| = |P''| = |P|$, so replacing P by either P' or P'' in $T(X)$ leads to the result. ■

Corollary 2.1. Let $X \subseteq Q(n)$, and let $T(X)$ be a Steiner tree on X. Let y be a branch point of T such that the cluster C determined by y is defined. Let $Z \subseteq X$ denote the set of leaves of C. Choose i, $1 \leq i \leq n$. Then, we have $z(i) = y(i)$ for all but possibly one $z \in Z$.

Proof. Suppose that u and v are elements of Z such that $u(i) = v(i) \neq y(i)$, for some i, $1 \leq i \leq n$. Without loss of generality, suppose $i = n$, $u(n) = v(n) = 1$, and $y(n) = 0$. By Lemma 2(ii), we may assume that the nth coordinates of all the vertices (except y) on the subpaths of $T(X)$ from u to y and from v to y are equal to 1. But then the vertices adjacent to y on these subpaths are one and the same. This common vertex is then a branch point of $T(X)$, contradicting the definition of the cluster C. ■

Lemma 2.3. Let X be a set of vertices of $Q(n - 1)$, $X = \{x_1,x_2, \ldots, x_k\}$. For $i = 1, \ldots, k$, choose $\varepsilon_i \in \{0,1\}$ and denote by $x_i\varepsilon_i$ the vertex of $Q(n)$ whose first $n - 1$ coordinates agree with x_i and whose nth coordinate is ε_i (i.e., $x_i\varepsilon_i$ is x_i "concatenated" with ε_i). Define $X\varepsilon$ to be the set $\{x_i\varepsilon_i : 1 \leq i \leq k\}$. Then, we have the following:

(i) If all the ε_i are the same, then $L(n,X\varepsilon) = L(n - 1,X)$.
(ii) If all but one of the ε_i are the same, $L(n,X\varepsilon) = L(n - 1,X) + 1$.

Proof. First we have the following notation. Let T be a tree in $Q(n-1)$. We will denote by $T0$ the tree in $Q(n)$ obtained by concatenating every vertex of T with a 0. Clearly, $T0$ is also connected, and $|T0| = |T|$. Let $\pi: Q(n) \to Q(n-1)$ be the canonical projection obtained by "truncating" the nth coordinate of elements of $Q(n)$, i.e., if $e_1e_2 \ldots e_n$ is a vertex in $Q(n)$, then $\pi(e_1e_2 \ldots e_n) = e_1e_2 \ldots e_{n-1}$ in $Q(n-1)$. Then, it is clear that if T is a tree in $Q(n)$, then $\pi(T)$, obtained by projecting every vertex of T into $Q(n-1)$, is a tree in $Q(n-1)$ and that $|T| \geq |\pi(T)|$.

(i) If $T(X)$ is a Steiner tree on X in $Q(n-1)$, and if all the ε_i are the same, say all equal 0, then $T(X)0$ is a tree containing $X\varepsilon$ and $|T(X)| = |T(X)0|$. Also, it is clear that $\pi(T(X)0) = T(X)$, so that $T(X)0$ must, in fact, be a Steiner tree on $X\varepsilon$. The result follows.

(ii) Without loss of generality, we may assume that $\varepsilon_i = 0$ for $1 \leq i \leq k-1$ and $\varepsilon_k = 1$. Let $X0$ denote $\{x_i0 : 1 \leq i \leq k\}$. Then, if $T(X0)$ denotes a Steiner tree on $X0$, we have $|T(X0)| = L(n-1,X)$ by part (i) above. Adding to $T(X0)$ the edge from x_k0 to x_k1 shows that $L(n,X\varepsilon) \leq L(n-1,X) + 1$.

On the other hand, if $T(X\varepsilon)$ denotes a Steiner tree on $X\varepsilon$, then $\pi(T(X\varepsilon))0$ is a tree on $X\varepsilon$ not containing the vertex x_k1, so $|T(X\varepsilon)| \geq L(n-1,X) + 1$; hence, the result follows. ■

We now need some notation for coordinates on which vertices are identical, as well as notation for certain types of vertices, obtained by repeating a single string a number of times and then truncating the last copy.

Let $X \subseteq Q(n)$, and N, the set of coordinate positions as before. Let $M \subseteq N$. Denote by $M[X]$ the subset of M consisting of those coordinate positions in M on which all elements of X agree, i.e., $i \in M[X]$ if and only if $i \in M$ and either the ith coordinate of every element in X is a 0 or the ith coordinate of every element in X is a 1. So $M[X] = M \cap N[X]$.

If x is a string of 0's and 1's, of length c, say, and if d and e are integers with $0 \leq e < c$, then $(x; d, e)$ will stand for the string of 0's and 1's of length $cd + e$ obtained by taking $(d+1)$ copies of x (concatenated, one following the other) and truncating the first $cd + e$ entries (or, alternatively, deleting the last $c - e$ entries). For example, if $x = 01100$ (so $c = 5$), $d = 3$, and $e = 2$, then

$$(x; 3, 2) = 01100011000110001, \text{ of length } 17.$$

We are now in a position to consider the case $k = 5$.

Let $X \subseteq Q(n)$ with $|X| = 5$. Let $N_1 = N \backslash N[X]$, $N_2 = \cup_{A \subseteq X \text{ and } |A|=4} N_1[A]$, and $N_3 = N \backslash (N[X] \cup N_2)$. For $A, B \subseteq X$ with $|A| = |B| = 3$, and $X = A \cup B$, define $t(A,B) = |N_3[A]| + |N_3[B]|$. Let $t = \max\{t(A,B) : A, B \subseteq X, |A| = |B| = 3, \text{ and } X = A \cup B\}$. Thus, N_1 is the set of coordinate positions on which either 3 or 4 points of X agree; N_2, the set on which some 4 points of X agree; and the N_3, the set on which some 3 points agree but no set of four points agree. We have $N = N_1 \cup N_2 \cup N_3$.

Theorem 2.2. Let $X \subseteq Q(n)$ with $|X| = 5$. Let $r = |N[X]|$, $s = |N_2|$, and let t be as above. Then, $L(n,X) = 2n - 2r - s - t$.

Proof. First suppose that $r = s = 0$. Then, $N_3 = N$. Let $A, B \subseteq X$ with $|A| = |B| = 3$, and $X = A \cup B$, such that $t(A,B) = t$. By Theorem 2.1, $L(n,A) = n - |N[A]|$ and $L(n,B) = n - |N[B]|$. If $T(A)$ and $T(B)$ denote Steiner trees on A and B, respectively, then the subgraph of $Q(n)$ obtained by taking all vertices and edges in $T(A)$ together with all the vertices and edges of $T(B)$ is connected, since it contains the vertex in $A \cap B$ and contains X. Taking a spanning subtree of this subgraph gives $L(n,X) \leq L(n,A) + L(n,B) = 2n - t$.

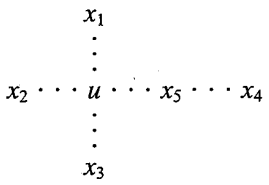
Note that in every coordinate position there are exactly three 0's or exactly three 1's, so that $t \geq 1$. Thus, certainly, $L(n,X) \leq 2n - 1$.

Now let $T(X)$ be a Steiner tree on X. Suppose there is an x in X and a (nonempty) connected component S of $T(X) - x$ such that $|Y| = 2$, where Y is the set of vertices of X contained in S. Then, by Lemma 2.1(iii), $|T(X)| = L(n,A) + L(n,B)$, where $A = Y \cup x$, $B = X - Y$, are such that $|A| = |B| = 3$, and $X = A \cup B$. Thus, by Theorem 2.1, $|T(X)| = n - |N[A]| + n - |N[B]| \geq 2n - t$, and we are done.

Thus, suppose no such x in X, as in the previous paragraph, exists. Let $X = \{x_1,x_2,x_3,x_4,x_5\}$. Then, we have the following cases to consider:

Case (i). $T(X)$ has a vertex, u, of degree 5 (and so every element of X is a leaf). Clearly, in every coordinate position, u differs from at least two elements of X. Thus, $|T(X)| = \sum_{i=1}^{5} d(u,x_i) \geq 2n$, a contradiction of $L(n,X) \leq 2n - 1$.

Case (ii). $T(X)$ has exactly one branch point, u, of degree 4:

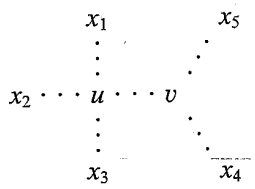


In this case, without loss of generality, we may assume that x_1, x_2, x_3, and x_4 are leaves and that x_5 lies on the path of $T(X)$ from u to x_4 (u could be x_5). See the diagram above. By Corollary 2.1, u must be the centroid of $\{x_1,x_2,x_3\}$, so that $\sum_{i=1}^{3} d_T(u, x_i) = n - |N[x_1,x_2,x_3]|$, where by d_T we mean the distance as measured in $T(X)$.

Clearly, $d_T(x_4,x_5) = n - |N[x_4,x_5]|$. Now since we are assuming $r = s = 0$, $N[x_1,x_2,x_3] \subseteq N[x_4,x_5]$. Also, if $u = x_5$, then $|N[x_4,x_5]| = 0$, while if $u \neq x_5$, then u disagrees with x_5 on $N[x_4,x_5]$. Hence, $d_T(u,x_5) \geq |N[x_4,x_5]|$, so that $d_T(u,x_5) + d_T(x_4,x_5) \geq n$.

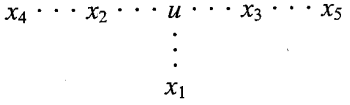
Thus, $|T(X)| \geq 2n - |N[x_1,x_2,x_3]| \geq 2n - t$.

Case (iii). $T(X)$ has two branch points, one, u, of degree 4 and the other, v, of degree 3:



As in Case (ii), u is the centroid of $\{x_1,x_2,x_3\}$ and $\sum_{i=1}^{3} d_T(u,x_i) = n - |N[x_1,x_2,x_3]|$. Again, using Corollary 2.1, v agrees with x_4 and x_5 on $N[x_4,x_5]$ and, thus, $d_T(v,x_4) + d_T(v,x_5) = n - |N[x_4,x_5]|$. Also, v disagrees with u on $N[x_4,x_5]$, so $d_T(u,v) \geq |N[x_4,x_5]|$. Thus, $|T(X)| \geq 2n - |N[x_1,x_2,x_3]| \geq 2n - t$.

Case (iv). $T(X)$ has exactly one branch point, u, of degree 3:



For vertices a, b, and c, denote by $N[\bar{a},b,c]$ the set of coordinates where b and c agree with each other but disagree with a, i.e., $N[\bar{a},b,c] = N[b,c] - N[a,b,c]$.

Then, $N[x_2,x_4] = N[x_1,x_2,x_4] \cup N[\overline{x_1},x_2,x_4]$, and $N[x_3,x_5] = N[x_1,x_3,x_5] \cup N[\overline{x_1},x_3,x_5]$, where in both cases, the unions are disjoint. Thus, for $i = 2$ and 3, we have

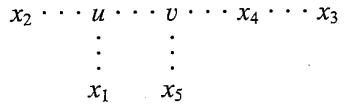
$$d_T(x_i,x_{i+2}) = n - |N[x_i,x_{i+2}]| = n - |N[x_1,x_i,x_{i+2}]| - |N[\overline{x_1},x_i,x_{i+2}]|.$$

Now observe that, since $s = 0$, $N[\overline{x_1},x_2,x_4] \cap N[\overline{x_1},x_3,x_5] = \varnothing$, while if $j \in N[\overline{x_1},x_2,x_4] \cup N[\overline{x_1},x_3,x_5]$, then exactly two of $\{x_1,x_2,x_3\}$ agree on position j. Thus,

$$\sum_{i=1}^{3} d_T(u,x_i) \geq |N[\overline{x_1},x_2,x_4]| + |N[\overline{x_1},x_3,x_5]|.$$

Hence, $|T(X)| \geq 2n - |N[x_1,x_2,x_4]| - |N[x_1,x_3,x_5]| \geq 2n - t$.

Case (v). $T(X)$ has two branch points, u and v, both of degree 3:



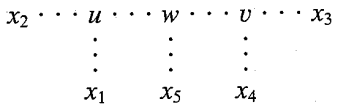
Using the notation above, $d_T(x_3,x_4) = n - |N[x_3,x_4]| = n - |N[x_5,x_3,x_4]| - |N[\overline{x_5},x_3,x_4]|$, and $d_T(u,x_1) + d_T(u,x_2) = n - |N[x_1,x_2]| = n - |N[x_5,x_1,x_2]| - |N[\overline{x_5},x_1,x_2]|$. We again have that $N[\overline{x_5},x_3,x_4] \cap N[\overline{x_5},x_1,x_2] = \varnothing$. On the other

hand, if $j \in N[\overline{x_5},x_3,x_4]$, then v disagrees with one of x_4 or x_5 in position j, while if $j \in N[\overline{x_5},x_1,x_2]$, then v disagrees with one of u or x_5 in position j. Thus,

$$d_T(u,v) + d_T(v,x_4) + d_T(v,x_5) \geq |N[\overline{x_5},x_3,x_4]| + |N[\overline{x_5},x_1,x_2]|,$$

leading to $|T(X)| \geq 2n - t$ as before.

Case (vi). $T(X)$ has three branch points, u, v and w, all of degree 3:



As in Case (v), we have

$$d_T(u,x_1) + d_T(u,x_2) = n - |N[x_1,x_2]| = n - |N[x_5,x_1,x_2]| - |N[\overline{x_5},x_1,x_2]|$$

and

$$d_T(v,x_3) + d_T(v,x_4) = n - |N[x_3,x_4]| = n - |N[x_5,x_3,x_4]| - |N[\overline{x_5},x_3,x_4]|.$$

Now if $j \in N[\overline{x_5},x_3,x_4]$, then w disagrees with one of v or x_5 in position j, while if $j \in N[\overline{x_5},x_1,x_2]$, then w disagrees with one of u or x_5 in position j. Thus,

$$d_T(u,w) + d_T(w,x_5) + d_T(v,w) \geq |N[\overline{x_5},x_3,x_4]| + |N[\overline{x_5},x_1,x_2]|,$$

leading to $|T(X)| \geq 2n - t$ as before.
Hence, $|T(X)| = 2n - t$.

All that remains to be considered is when one, or both, of r or s is not zero. Let $m = n - r - s = |N_3|$ and, without loss of generality, assume that $N_3 = \{1, 2, \ldots, m\}$. Let X' be the set of vertices in $Q(m)$ obtained by deleting the last $r + s$ coordinates from the elements of X. Then, by what we did above, $L(m,X') = 2m - t$, where t is as before. Then, by applying Lemma 2.3(i) r times and Lemma 2.3(ii) s times, we see that $L(n,X) = 2m - t + s = 2n - 2r - s - t$, as required. ■

Using Theorem 2.2, we can now determine $L(n,5)$.

Corollary 2.2. For $n \geq 3$, $L(n,5) = 2n - \lceil n/10 \rceil - \lceil (n-4)/10 \rceil$.

Proof. We first show that $L(n,5) \leq 2n - \lceil n/10 \rceil - \lceil (n-4)/10 \rceil$. Let $X \subseteq \Omega$ with $|X| = 5$, and let r, s, t, etc., be as in Theorem 2.2. We will assume $n \geq 3$ and prove that $t \geq \lceil n/10 \rceil + \lceil (n-4)/10 \rceil$. As was pointed out in the proof of Theorem 2.2, $L(n,X) \leq 2n - 1$, so we are done if $n = 3$ and 4.

First suppose that $r = s = 0$. Then, given any coordinate position, exactly three vertices of X agree on that position, so if we let $\Sigma = \{A_1, \ldots, A_{10}\}$ be the 10 distinct subsets of X of size three, there is a mapping θ from the set $N = \{1, 2, \ldots, n\}$ of coordinate positions to Σ given by $\theta(i)$ = the subset of size three of X that agrees on position i. Then, given $A, B \subseteq X$ with $|A| = |B| = 3$ and $X = A \cup B$, $t(A,B)$ can be interpreted as follows: $t(A,B) = |\theta^{-1}(A)| + |\theta^{-1}(B)|$. We will show that there exist subsets $A, B \subseteq X$ with $|A| = |B| = 3$, and $X = A \cup B$, such that $t \geq t(A,B) \geq \lceil n/10 \rceil + \lceil (n-4)/10 \rceil$.

Consider the restriction $\bar{\theta} : \{1,2,\ldots,n-1\} \to \{A_1, \ldots, A_{10}\}$ of the map θ to the coordinate positions 1 through $n-1$. In analogy with the functions $t(A,B)$ and t discussed above, we let $\bar{t}(A,B) = |\bar{\theta}^{-1}(A)| + |\bar{\theta}^{-1}(B)|$, when $|A| = |B| = 3$ and $X = A \cup B$, and we let $\bar{t} = \max\{\bar{t}(A,B): |A| = |B| = 3, \text{ and } X = A \cup B\}$.

So suppose, as an inductive hypothesis, that there exist i and j with $1 \leq i \neq j \leq n-1$, such that $\theta(i) \cup \theta(j) = X$ and $\bar{t}(\theta(i), \theta(j)) \geq \lceil (n-1)/10 \rceil + \lceil (n-5)/10 \rceil$.

If n is congruent to neither 1 nor 5 (mod 10), then $\lceil (n-1)/10 \rceil = \lceil n/10 \rceil$ and $\lceil (n-5)/10 \rceil = \lceil (n-4)/10 \rceil$, so since $\bar{t} \geq \lceil n/10 \rceil + \lceil (n-4)/10 \rceil$, by the inductive hypothesis, we are done because $t \geq \bar{t}$.

So suppose $n \equiv 1 \pmod{10}$ and write $n = 1 + 10q$. Then, by the inductive hypothesis, we know that there are subsets A_i and A_j in Σ such that $\bar{t} = \bar{t}(A_i, A_j) \geq \lceil (n-1)/10 \rceil + \lceil (n-5)/10 \rceil = 2q$. If $\bar{t}(A_i,A_j) > 2q$, then $t \geq \bar{t} \geq \bar{t}(A_i,A_j) \geq \lceil n/10 \rceil + \lceil (n-4)/10 \rceil$ and we are done. So we may assume that $\bar{t}(A_i,A_j) \leq \bar{t} = 2q$ for all $1 \leq i, j \leq 10$. It suffices to show that $t(\theta(n),A_i) = 2q + 1$ for some i since then $t \geq t(\theta(n),A_i) = 2q + 1 = \lceil n/10 \rceil + \lceil (n-4)/10 \rceil$.

Now define a graph P with vertices the elements of Σ, with two elements of Σ being adjacent iff their union is X. Then, it is not difficult to see that P is isomorphic to the Petersen graph. For $1 \leq k \leq 10$, let $n_k = |\{i : 1 \leq i \leq n-1, \bar{\theta}(i) = A_k\}|$. Then, since $\bar{t} = 2q$, we have $n_k + n_p \leq 2q$ for all $1 \leq k \neq p \leq 10$ with A_k adjacent to A_p in P. Further, $\sum_{k=1}^{10} n_k = n - 1 = 10q$.

We claim that $n_k = q$ for all $1 \leq k \leq 10$. Without loss of generality, suppose first that $n_1 = q + a$, for some $a \geq 0$, and suppose that A_8, A_9, and A_{10} are the three vertices adjacent to A_1. Then, $n_i \leq q - a$, for $i = 8, 9$, and 10. Also, it can be seen that on the remaining six vertices of the Petersen graph, there is a matching of three edges. Thus,

$$10q = \sum_{k=1}^{10} n_k \leq (q + a) + 3(q - a) + 3 \cdot 2q,$$

whence $a = 0$. Thus, $n_k \leq q$ for all $1 \leq k \leq 10$. Then, from the fact that $\sum_{k=1}^{10} n_k = 10q$, we must have $n_k = q$ for all $1 \leq k \leq 10$. Then, $|\theta^{-1}(\theta(n))| = |\bar{\theta}^{-1}(\theta(n))| + 1 = q + 1$, and, thus, $t(\theta(n),A_i) = 2q + 1$ for any i for which $\theta(n) \cup A_i = X$, as required.

Finally, suppose $n \equiv 5 \pmod{10}$ and write $n = 5 + 10q$. Then, by the inductive hypothesis with $m = n - 1$, we know that there are subsets in Σ, A_i and A_j, say, such that $\bar{t} = \bar{t}(A_i,A_j) \geq \lceil (n-1)/10 \rceil + \lceil (n-5)/10 \rceil = 2q + 1$. If $\bar{t}(A_i,A_j) > 2q + 1$, then $t \geq \bar{t}(A_i,A_j) \geq \lceil n/10 \rceil + \lceil n - 4/10 \rceil$ and as above we are done. So we may assume that $\bar{t} = \bar{t}(A_i,A_j) = 2q + 1$. Also we are reduced to

showing that for some A_i adjacent to $\theta(n)$ we have $t(\theta(n),A_i) > 2q + 1$, since it would then follow that $t \geq t(\theta(n),A_i) \geq \lceil n/10 \rceil + \lceil (n-4)/10 \rceil$. Again, letting n_k be as above, we have $n_k + n_p \leq 2q + 1$ for all $1 \leq k \neq p \leq 10$ with A_k adjacent to A_p in the Petersen graph defined above and $\sum_{k=1}^{10} n_k = n - 1 = 10q + 4$.

We claim that $n_k = q + 1$ for four values of k with $1 \leq k \leq 10$ and $n_k = q$ for the remaining six values of k. Without loss of generality, suppose first that $n_1 = q + 1 + a$, for some $a \geq 0$, and suppose that A_8, A_9, and A_{10} are the three vertices adjacent to A_1. Then, $n_i \leq q - a$, for $i = 8$, 9, and 10. Again, on the remaining six vertices of the Petersen graph, there is a matching of three edges. Thus,

$$10q + 4 = \sum_{k=1}^{10} n_k \leq (q + 1 + a) + 3(q - a) + 3(2q + 1),$$

whence $a = 0$. Thus, no n_k is greater than $q + 1$, and if $n_k = q + 1$ for some k, then $n_p \leq q$ for each A_p adjacent to A_k. Now the maximum size of an independent set in the Petersen graph is 4, so $10q + 4 = \sum_{k=1}^{10} n_k \leq 4(q + 1) + 6q$, whence equality holds, and the claim is true. Without loss of generality, $n_k = q + 1$, for $1 \leq k \leq 4$, and $n_k = q$, for $5 \leq k \leq 10$.

Let $\theta(n) = A_s$, and note that $|\theta^{-1}(A_s)| = |\bar{\theta}^{-1}(A_s)| + 1 = n_s + 1$. If $1 \leq s \leq 4$, then for any A_i that is adjacent to A_s in P (so that $i \geq 5$), we have

$$t(A_s,A_i) = |\theta^{-1}(A_s)| + |\theta^{-1}(A_i)| = n_s + 1 + q + 1 = 2q + 2.$$

If $s \geq 5$, then there is an A_i, $1 \leq i \leq 4$, which is adjacent to A_s in P, so that again

$$t(A_s,A_i) = |\theta^{-1}(A_s)| + |\theta^{-1}(A_i)| = n_s + 1 + q + 1 = 2q + 2,$$

as required.

If one (or both) of r and s is not zero, then a similar argument as above, with N replaced by N_3, and n replaced by $m = |N_3| = n - r - s$, shows that $t \geq \lceil m/10 \rceil + \lceil m - 4/10 \rceil$. Hence, by Theorem 2.2, $L(n,X) \leq 2n - (2r + s + \lceil m/10 \rceil + \lceil (m-4)/10 \rceil) \leq 2n - \lceil n/10 \rceil - \lceil (n-4)/10 \rceil$.

Now write $n = 10d + e$, with $0 \leq e < 10$. Let $X = \{x_1,x_2,x_3,x_4,x_5\}$, where

$$x_1 = (0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0;\ d,\ e) = 0,$$

$$x_2 = (0\ 1\ 1\ 1\ 0\ 1\ 1\ 1\ 0\ 0;\ d,\ e),$$

$$x_3 = (1\ 1\ 1\ 0\ 0\ 0\ 0\ 1\ 1\ 1;\ d,\ e),$$

$$x_4 = (1\ 1\ 0\ 1\ 1\ 1\ 0\ 0\ 0\ 1;\ d,\ e), \text{ and}$$

$$x_5 = (1\ 0\ 1\ 1\ 1\ 0\ 1\ 0\ 1\ 0;\ d,\ e).$$

Then, it can be checked that, with $A = \{x_1,x_2,x_3\}$ and $B = \{x_3,x_4,x_5\}$, we have $t = t(A,B)$, and then $L(n,X) = 2n - \lceil n/10 \rceil - \lceil (n-4)/10 \rceil$, by Theorem 2.2, finishing the proof. ■

Using the above results, we can consider the case $|X| = 4$ in more detail. (See the MCUB earlier).

Let $X \subseteq Q(n)$ with $|X| = 4$. As notation, we let $N_1 = N \backslash N[X]$ and $N_2 = \cup_{A \subseteq X \text{ and } |A|=3} N_1[A]$. Now we let $N_3 = N \backslash (N[X] \cup N_2)$, and for $A \subseteq X$ with $|A| = 2$, we let $t(A) = |N_3[A]|$. Finally, let $t = \max\{t(A): A \subseteq X, |A| = 2\}$.

Theorem 2.3. Let $X \subseteq Q(n)$ with $|X| = 4$. Let $r = |N[X]|$, $s = |N_2|$, and t as above. Then, $L(n,X) = 2n - 2r - s - t$.

Proof. As before, we may assume without loss of generality that $r = s = 0$. Let $X = \{x_1,x_2,x_3,x_4\}$ and let $A \subseteq X$, $|A| = 2$, such that $t = t(A)$; say $A = \{x_3,x_4\}$. Let $x_5 = x_4$ and apply Theorem 2.2 to the set $\{x_1,x_2,x_3,x_4,x_5\}$. It is not difficult to show that the maximum value of $t(A,B)$, as defined just before Theorem 2.2, is $t(A',B')$, where $A' = \{x_1,x_2,x_5\}$ and $B' = \{x_3,x_4,x_5\}$, and that this in turn is $t(A)$ as defined above.

Hence, $L(n,X) = 2n - 2r - s - t$, as required. ∎

Corollary 2.3. For $n \geq 3$, $L(n,4) = \lfloor (5/3)n \rfloor$, the value given by the MCUB.

Proof. Let $X \subseteq Q(n)$ with $|X| = 4$. Again, we may suppose $r = s = 0$. There are exactly six distinct subsets of size two that can be chosen from X. Given any coordinate position, there are exactly two 0's and two 1's (since $r = s = 0$), and so for each $A \subseteq X$, with $|A| = 2$, we have $N[A] = N[X \backslash A]$. Thus, $t \geq \lceil n/3 \rceil$. Hence, by Theorem 2.3, $L(n,X) \leq 2n - \lceil n/3 \rceil = \lfloor (5/3)n \rfloor$.

Now write $n = 3d + e$, with $0 \leq e \leq 2$. Let $X = \{x_1,x_2,x_3,x_4\}$, where

$$x_1 = (0\ 0\ 0;\ d,\ e) = 0,$$

$$x_2 = (0\ 1\ 1;\ d,\ e),$$

$$x_3 = (1\ 0\ 1;\ d,\ e), \text{ and}$$

$$x_5 = (1\ 1\ 0;\ d,\ e).$$

Then, we see, using Theorem 2.3, that $L(n,X) = 2n - \lceil n/3 \rceil$, and we are done. ∎

3. SETS OF CONSTANT WEIGHT AND THE TURAN PROBLEM

Let $W(k + 1,n)$ denote the set of all weight $k + 1$ points in $Q(n)$. In this section, we describe an upper bound for $L(n, W(k + 1,n))$ that follows by applying the work of Frankl and Rodl [5] on the generalized Turan problem.

Observe first that a trivial upper bound for $L(n,W(k + 1,n))$ is obtained by noting that the subgraph of $Q(n)$ induced by $W(k + 1,n) \cup W(k,n)$ is connected. It follows that $L(n,W(k + 1,n)) \leq |W(k + 1,n)| + |W(k,n)| - 1 = \binom{n}{k+1} + \binom{n}{k} - 1$. Of course, one would expect to improve this bound since there are surely more economical ways of dominating $W(k + 1,n)$ by weight k words than to use all of

$W(k,n)$. Nevertheless, the dramatic improvement implied by the result of [5] still comes as a surprise; assuming certain growth conditions on k (relative to n), we have

$$L(n,W(k+1,n)) \leq \binom{n}{k+1} + (2 + o(1)) \left(\frac{\log(k)}{k}\right)\binom{n}{k}$$

as $k \to \infty$.

As background, recall the Turan problem for graphs. This is to find the maximum number of edges a graph G on n points can have so that G contains no subgraph isomorphic to K_t (the complete graph on t points). This problem was solved in [9]. Equivalently, one may ask for the smallest size of a set S of edges in K_n such that every K_t subgraph of K_n contains some edge from S.

Generalizing to hypergraphs, one may ask for the minimum possible size $C(n,k+a,k)$ of a collection S of k-sets from a ground set X on n points so that every $(k+a)$-subset of X contains at least one element from S. The connection with hypercubes is then natural. First, we view the elements of X as the coordinate positions used in describing the points of $Q(n)$. Each subset G of X then corresponds to a point, which we denote $p(G)$, of $Q(n)$ having 1's in the coordinate belonging to G and 0's in the remaining coordinates. Similarly, for a collection S of subsets of X, we let $p(S)$ be the subset $\{p(G)\colon G \varepsilon S\}$ of the point set of $Q(n)$. We shall be interested in the case $a = 1$, since under this correspondence, the number $C(n,k+1,k)$ will be the fewest number of weight k points of $Q(n)$ needed to dominate $W(k+1,n)$. It was shown in [3] that $C(n,2k+1,2k)/\binom{n}{2k} \leq \frac{1}{4} + 4^{-k}$ by using Hadamard designs, among many other interesting results on the Turan problem for hypergraphs [2]. This result is good for small k, but for large k, the following is stronger.

Theorem 3.1 [5]. For $k \to \infty$, one has

$$C(n,k+1,k) \leq (1 + o(1)) \left(\frac{\log(k)}{k}\right)\binom{n}{k}.$$

Proof. We reproduce the construction underlying the theorem both for completeness and since we will need certain of its features later.

Let $X = \{1,2,\ldots,n\}$ be a ground set of n elements, and let $\binom{X}{m}$ denote the collection of m element subsets of X for any $1 \leq m \leq n$. Consider an integer $r < k$ to be chosen later, and assume without loss of generality that $n = rt$, where t is some positive integer. Now partition X as $X = X_0 \cup X_1 \cup X_2 \cup \ldots \cup X_{r-1}$, where $X_i = \{c\colon 1 \leq c \leq n,\ c \equiv i (\mathrm{mod}\ r)\}$. For any subset $G \subset X$, let $S(G) = \{i\colon G \cap X_i \neq \varnothing\}$, and $s(G) = |S(G)|$. For any integer e, $0 \leq e \leq r-1$, define

$$V_e = \left\{F \varepsilon \binom{X}{k}\colon e + \sum_{x \varepsilon F} x \equiv 0,1,2,\ldots, \text{or } r - s(F) (\mathrm{mod}\ r)\right\}.$$

We claim that for any e and any $G \varepsilon \binom{X}{k+1}$, there exists $F \varepsilon V_e$ such that $F \subset G$. To see this, for any $g \varepsilon G$, let $y(g) = e + \sum_{x \varepsilon G - \{g\}} x$. The numbers $\{y(g)\colon g \varepsilon G\}$

form a set of $s(G)$ distinct numbers modulo r. Hence, by the pigeon hole principle, one can choose a w such that $y(w) \equiv i \pmod r$ for some i satisfying $0 \le i \le r - s(G)$. The set $F = G - \{w\}$ then satisfies $F \varepsilon V_e$ and $F \subset G$, thereby proving the claim.

Finally, it can be shown (see [5]) that with $r = \lfloor k/\lfloor \log(k) \rfloor \rfloor$ there must exist an s such that

$$|V_s| < \binom{n}{k} \frac{1}{\lfloor k/\log(k) \rfloor} \frac{\log(k)}{\log(k) - 1}.$$

Taking $p(V_s)$ as our set of weight k points to dominate the weight $k + 1$ points, the theorem is proved. ■

We will need some notation to analyze connectedness in the subgraph of $Q(n)$ induced by $p(V_s) \cup \binom{X}{k+1}$. For a given $G \subset X$ with $|X| = n$, let $Cong(G)$ be the multiset of congruence classes (mod r) of all the elements of G. Now let W be any multiset of size k drawn from the set $\{0,1,2, \ldots, r-1\}$. We write $Block(W)$ for the set of all $G \subset X$ such that $\text{Cong}(G) = W$, and we shall refer to any collection B of subsets of X satisfying $B = \text{Block}(W)$ for some W as a *block*. For example, with $k = 4$, $r = 3$, and n large enough, we have $G \varepsilon \text{Block}(W)$, where $G = \{3,6,4,8\}$ and $W = \{0,0,1,2\}$. Note that any set V_e can be partitioned as a disjoint union of blocks since if $G \varepsilon V_e$, then any $H \varepsilon \binom{X}{k}$ satisfying $\text{Cong}(H) = \text{Cong}(G)$ must also belong to V_e. Finally, for any subset $M \subset W(k,n)$, we let $Nb(M) = \{u \varepsilon W(k+1,n)$: $uv \varepsilon E(Q(n))$ for some $v \varepsilon M\}$.

Lemma 3.1. For any block B, the subgraph of $Q(n)$ induced by $p(B) \cup Nb(p(B))$ is connected.

Proof. Let $G,H \varepsilon B$. We will construct a sequence of "exchanges" leading from G to H. Corresponding to this sequence is a path in $Q(n)$ from $p(G)$ to $p(H)$ whose points alternate between $p(B)$ and $Nb(p(B))$.

More precisely, let $G \oplus H$ be the symmetric difference of G and H, and say $G \oplus H = \{x_1,x_2, \ldots, x_c, y_1, y_2, \ldots, y_c\}$, where $x_i \varepsilon G \backslash H$, $y_i \varepsilon H \backslash G$, and $x_i \equiv y_i \pmod r$ for all $i \le c$. Now let $G_0 = G$, and inductively let $G_i = G_{i-1} \cup \{y_{\lceil i/2 \rceil}\}$ for i odd and $G_i = G_{i-1} \backslash \{x_{i/2}\}$ for i even, $1 \le i \le 2c$. An easy induction shows that $G_{2c} = H$, while $G_i \varepsilon B$ for i is even and $p(G_i)p(G_{i+1})$ is an edge of $Q(n)$ for all i. Hence, $p(G_0)p(G_1) \ldots p(G_{2c})$ is a path in $Q(n)$ from $p(G)$ to $p(H)$ alternating between $p(B)$ and $Nb(p(B))$, as required. ■

This lemma can now be applied to give our bound on $W(k+1,n)$.

Theorem 3.2. $L(n,W(k+1,n)) \le \binom{n}{k+1} + (2 + o(1))\,([\log(k)]/k)\binom{n}{k}$ as $k \to \infty$, provided $(k^2/[\log(k)])^{1+1/k} \le n$.

Proof. A natural approach for constructing a subtree T of $Q(n)$ spanning $W(k+1,n)$ is as follows: We start with the subset $W(k+1,n) \cup V_s$ of $Q(n)$,

where $|V_s|$ satisfies the bound in the proof of Theorem 3.1. Now this set may not induce a connected subgraph of $Q(n)$. So write $V_s = \cup_i B_i$ as a disjoint union of blocks, and for each i, choose a point $u_i \varepsilon p(B_i)$. By the lemma it is then only necessary to "connect up" the u_i. We can do this by including in our set the origin 0 of $Q(n)$ and a shortest path, call it Path_i, from 0 to u_i for each i. Thus, the graph H induced by the set $W(k+1,n) \cup V_s \cup (\cup_i \text{Path}_i)$ of $Q(n)$ is a connected subgraph of $Q(n)$ spanning $W(k+1,n)$. Now just let T be any spanning tree of H.

It remains to estimate the number of points in H. The quantities $|W(k+1,n)|$ and $|V_s|$ being $\binom{n}{k+1}$ and the estimate of Theorem 3.1, respectively, we concentrate on $|\cup_i \text{Path}_i|$. Now an upper bound for the number of blocks in V_s is r^k since each block is specified by a selection of k elements (with repetitions allowed) from the set $\{0,1,2,\ldots,r-1\}$. Hence, we need at most r^k points u_i. Since each u_i has weight k, it follows that $|\cup_i \text{Path}_i| \leq kr^k$. Now using the value $r = \lfloor k/\lfloor \log(k) \rfloor \rfloor$ chosen in the proof of the theorem and the lower bound $\binom{n}{k} \geq (n/k)^k$, one can show that $kr^k \leq (\lceil \log(k) \rceil / k)\binom{n}{k}$ provided $(k^2/\log(k))^{1+1/k} \leq n$. Hence, with this bound on k, we find that the contribution to $|H|$ from $|\cup_i \text{Path}_i|$ is at most $(\lceil \log(k) \rceil / k)\binom{n}{k}$, and it follows that $|H| \leq \binom{n}{k+1} + (2 + o(1))(\lceil \log(k) \rceil / k)\binom{n}{k}$ as required. ■

4. NP-COMPLETENESS FOR THE WEIGHT 2 SUBPROBLEM

The complexity of the Steiner problem in the hypercube can be addressed by considering the following decision problem: Given a set $X \subset Q(n)$ and an integer k, decide if $L(n,X) \leq k$. This problem was shown to be NP-complete in [4]. In this section, we strengthen this result by showing that the following severely restricted subproblem, which we call ST(2), is NP-complete.

ST(2): Given a set $X \subset Q(n)$ consisting of points of weight at most 2, and an integer k, decide if $L(n,X) \leq k$.

Theorem 4.1. ST(2) is NP-Complete.

Proof. We describe a polynomial time reduction VERTEX COVER $\rightarrow$ ST(2). This suffices by the well-known NP-completeness of the vertex cover problem and the fact that ST(2) is clearly in NP. Recall that the vertex cover problem is defined as follows:

VERTEX COVER: Given a graph G and integer k, decide if G has a "vertex cover" of size $\leq k$, that is, decide if there exists a set $S \subset V(G)$ such that every edge of G is incident on at least one point of S, and $|S| \leq k$.

Let $\{G,k\}$ be an instance of VERTEX COVER, where G has n points and $V(G) = \{1,2,\ldots,n\}$. The corresponding instance of ST(2) will be a collection $w(G)$ of points in $Q(n)$ defined as follows: For each edge ij of G, $1 \leq i,j \leq n$, let $w(i,j)$ be the point of $Q(n)$ with 1's in the ith and jth coordinates and 0's

everywhere else, and as usual let 0 be the origin. Now, let $w(G) = \{w(i,j): ij \varepsilon E(G)\} \cup \{0\}$. The corresponding instance of ST(2) is then $\{w(G),|E(G)| + k + 1\} = \{w(G),|w(G)| + k\}$.

We claim that G has a vertex cover of size $\leq k \Leftrightarrow L(n,w(G)) \leq |w(G)| + k$.

The direction $\Rightarrow$ is straightforward. Let $S \subset V(G)$ be a vertex cover of G of size at most k. Then S corresponds naturally to a collection of weight 1 words in $Q(n)$, namely, $E(S) = \{e(i): i \varepsilon S\}$, where $e(i)$ is the point having a 1 in the ith coordinate and 0's elsewhere. Now the set $w(G) \cup E(S)$ induces a connected subgraph of $Q(n)$ since $E(S)$ dominates all the weight 2 points while the points of $E(S)$ are all joined to the origin. Since this set also spans $w(G)$, it follows that $L(n,w(G)) \leq |w(G)| + |S| \leq |w(G)| + k$.

The direction $\Leftarrow$ is more involved. We will need some notation and terms. For a set $X \subset Q(n)$, denote by $X(r)$ the set of weight r words in X, i.e., $X(r) = X \cap W(r,n)$. For any $v \varepsilon Q(n)$, we let *support*(v) be the set of coordinates of v that are 1. Now suppose every point of X has weight at least r, and let $k < r$. The *k-shadow*(X) is the set $\{w \varepsilon W(k,n)$: support$(w) \subset$ support(w') for some w' in $X\}$. Similarly, if every point of X has weight at most r and $k > r$, then we let *k-shade*(X) be the set $\{w \varepsilon W(k,n)$: support$(w) \supset$ support(w') for some w' in $X\}$.

Observe first that it suffices to show that for any connected subgraph $H \subset Q(n)$ containing $w(G)$ there exists a connected subgraph $H' \subset Q(n)$ containing $w(G)$ such that $|H'| \leq |H|$ and all points in H' have weight of at most 2. For if this were to hold, then $H'(1)$ would correspond to a vertex cover of G (by the correspondence $i \varepsilon G \leftrightarrow e(i) \varepsilon Q(n)$ described earlier in the proof). Assuming then that $|H| \leq |w(G)| + k$, it would follow from $|H'| \leq |H|$ that $|H'(1)| = |H'| - |w(G)| \leq |H| - |w(G)| = k$. Thus, G would have a vertex cover of size at most k, as required.

Our strategy for building H' is to find a set $D \subset W(1,n)$ satisfying

(Property A) 2-shade$(D) \supset H(2)$

and

(Property B) $|D| \leq |H \backslash w(G)|$.

For then, the graph H' induced by $w(G) \cup D$ would satisfy the requirements of the preceding paragraph [and, in fact, $D = H'(1)$].

The set D will be a disjoint union $D' \cup D''$. The first component D' will "simulate" $H(3)$ in the sense that 2-shade$(D') \supset$ 2-shadow$(H(3)) \supset H(2) \cap$ [2-shadow$(H(3))$]. The second component D'' will be $H(1) \backslash D'$, and will therefore satisfy 2-shade$(D'') \supset H(2) \backslash$ [2-shadow$(H(3))$]. Hence, D will satisfy Property A.

As notation, a subset $R \subset Q(n)$ will be called a *3-2 component of* H if it satisfies the following:

(a) $R \subset W(2,n) \cup W(3,n)$, $R(3) \subset H$, $|R(3)| \geq 2$,
(b) the graph induced by R in $Q(n)$ is connected, and
(c) R is maximal in $Q(n)$ with respect to (a) and (b).

A point $w\varepsilon H(3)$ will be called *isolated* if it is not contained in any 3-2 component of H, and we let $\Omega = \{w\varepsilon H(3)$: w isolated$\}$.

First, we show that for any 3-2 component R there is a set $D_R \subset$ 1-shadow($R(3)$) such that 2-shade(D_R) $\supset$ 2-shadow($R(3)$) and $|D_R| \leq |R(3)|$. We can construct D_R inductively as follows: Let $R(3) = \{w_1, w_2, \ldots, w_k\}$, $k \geq 2$. The elements of $R(3)$ may be considered as triangles in a graph on n points [the points of a triangle being the coordinates in which the corresponding element of $R(3)$ is 1] and the elements of $R(2)$ as edges of the graph contained in these triangles. The indexing of the w_i may be assumed to be such that we can "grow" $R(3)$ by starting with w_1, and inductively having formed the subset $R^i = \{w_1, w_2, \ldots, w_i\}$, the subset $R^{i+1} = R^i \cup \{w_{i+1}\}$ will have the property that w_{i+1} has an edge in common with some w_t, $t \leq i$. Finally, $R^k = R(3)$. We can then build D_R as follows: Let z_1 and z_2 be the two vertices [corresponding to points in $R(1)$] in the intersection of the triangles w_1 and w_2. Clearly, all 5 edges in $\{w_1, w_2\}$ are covered by z_1 and z_2 (equivalently, the corresponding 5 points in the 2-shadow of $\{w_1, w_2\}$ lie in the 2-shade of z_1 and z_2). Inductively having constructed the set $D^i = \{z_1, z_2, \ldots, z_i\}$ contained in R^i such that the 2-shade of D^i contains the 2-shadow of R^i, we let z_{i+1} be the vertex of w_{i+1} (if any) not lying in w_i. Then, let $D^{i+1} = D^i \cup \{z_{i+1}\}$. It follows by induction and the choice of z_{i+1} that $|D^{i+1}| \leq |R^{i+1}|$ and 2-shade(D^{i+1}) $\supset$ 2-shadow(R^{i+1}). Letting $D_R = D^k$, we have found our required set in 1-shadow($R(3)$).

Now we let D'(component) $= \cup_R D_R$, where the union is over all 3-2 components R. Then, 2-shade(D'(component)) $\supset$ 2-shadow($\cup_{R \text{ a 3-2 component}} R(3)$) = 2-shadow($H(3)\backslash\Omega$) and $|D'(\text{component})| \leq |\cup_{R \text{ a 3-2 component}} R(3)|$. We have thus simulated $H(3)\backslash\Omega$ by D'(component) and also shown that $|D'(\text{component})| \leq |H(3)\backslash\Omega|$. It remains to simulate Ω.

Consider then a point (or triangle) $w\varepsilon\Omega$. Since H is connected, it follows that either

(i) H contains some point $x(w)\varepsilon$1-shadow(w), or
(ii) if (i) does not hold, then H contains some point $y(w)\varepsilon$4-shade(w).

Let x_1 and x_2 be two points in 1-shadow(w) for which $|\{x_1, x_2\} \cap H(1)|$ is a maximum, and $x(w)\varepsilon\{x_1, x_2\}$ if (i) holds. Let $Y(w) = \{x_1, x_2\}$. Clearly, 2-shade($Y(w)$) $\supset$ 2-shadow(w); so we have simulated w by $Y(w)$.

We remark that in case (ii) there cannot exist a point $w'\varepsilon H(3) \cap \{3\text{-shadow}(y(w))\}$, since then w and w' would be in the same 3-2 component of H, contradicting $w\varepsilon\Omega$. Thus, we have

Property y(w). If $w_1 \neq w_2$, then $y(w_1) \neq y(w_2)$.

We can now build the required set D. For each $w\varepsilon\Omega$, let $Y = \cup_{w\varepsilon\Omega} Y(w)$. So by the above 2-shade(Y) $\supset$ 2-shadow(Ω). Now let $D' = Y \cup D'$(component), $D'' = H(1)\backslash D'$, and $D = D' \cup D''$. Then, since 2-shade(D'(component)) $\supset$ 2-shadow($H(3)\backslash\Omega$), we have

$$\begin{aligned} 2\text{-shade}(D) &\supset [2\text{-shadow}(H(3))] \cup [H(2)\backslash 2\text{-shadow}(H(3))] \\ &\supset H(2). \end{aligned}$$

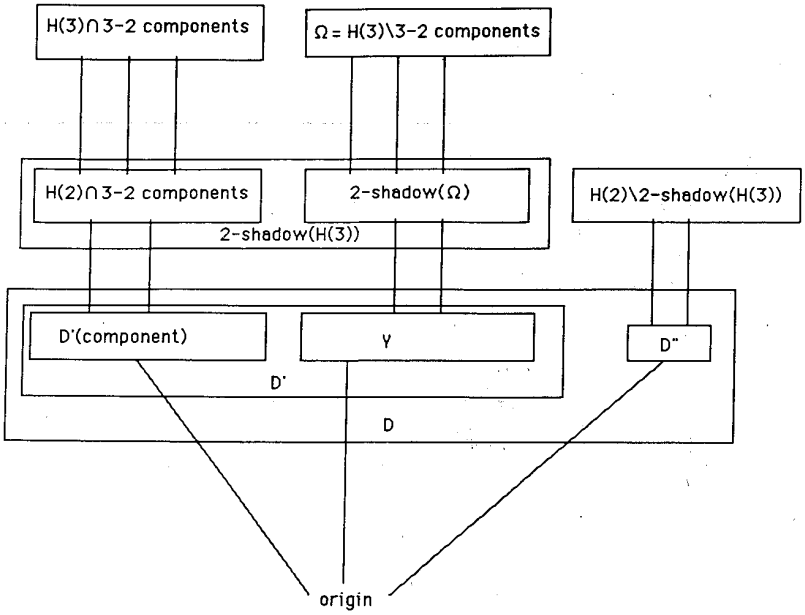


FIG. 1. The set D and property A.

Thus, D satisfies Property A above. A pictorial representation of how D satisfies Property A is given in Figure 1.

Toward proving that D has Property B, set $Z = [D'\backslash H(1)] \cap Y = Y\backslash H(1)$. We claim that it suffices to show that $|Z| \leq |\Omega| + |H(4)|$. For if this were true, then

$$\begin{aligned} |D'\backslash H(1)| &\leq |[D'\backslash H(1)] \cap D'(\text{component})| + |[D'\backslash H(1)] \cap Y| \\ &\leq |D'(\text{component})| + |\Omega| + |H(4)| \\ &\leq |H(3)| + |H(4)|. \end{aligned}$$

Hence, we would have

$$\begin{aligned} |D| &\leq |D' \cap H(1)| + |D''| + |D'\backslash H(1)| \\ &\leq |H(1)| + |H(3)| + |H(4)| \\ &\leq |H\backslash w(G)|, \end{aligned}$$

thereby proving Property B for D, and the proof would be completed.

We are therefore reduced to showing that $|Z| \leq |\Omega| + |H(4)|$. Pick $w \varepsilon \Omega$. If case (ii) holds for w, then $|\{w, y(w)\}| = |Y(w)| = |Y(w)\backslash H(1)|$. If case (i) holds for w, then $1 = |\{w\}| \geq |Y(w)\backslash H(1)|$. Thus, $|Z| = |Y\backslash H(1)| = \sum_{w \varepsilon \Omega} |Y(w)\backslash H(1)| \leq |\Omega| + |4\text{-shade}(w)|$ [by Property $y(w)$] $\leq |\Omega| + |H(4)|$, as required. ■

REFERENCES

[1] A. V. Aho, M. R. Garey, and F. K. Hwang, Rectilinear Steiner trees: Efficient special-case algorithms. *Networks* **7** (1977) 37–58.

[2] D. deCaen, On Turan's hypergraph problem. Ph.D. Thesis, University of Toronto (1982).

[3] D. deCaen, D. L. Kreher, and J. Wiseman, On constructive upper bounds for the Turan numbers T(n,2r+1,2r). *Congress. Numer.* **65** (1988) 277–280.

[4] L. R. Foulds and R. L. Graham, The Steiner problem in phylogeny is NP-complete. *Adv. Appl. Math.* **3** (1982) 43–49.

[5] P. Frank and V. Rodl, Lower bounds for Turan's problem. *Graphs and Combin.* **1** (1985) 213–216.

[6] M. R. Garey and D. S. Johnson, The rectilinear Steiner tree problem is NP-complete. *SIAM J. Appl. Math.* **32**(4) (1977) 826–833.

[7] F. K. Hwang, On Steiner minimal trees with rectilinear distance. *SIAM J. Appl. Math.* **30**(1) (1976) 104–114.

[8] F. K. Hwang and D. Richards, Steiner tree problems. *Networks,* in press.

[9] P. Turan, An extremal problem in graph theory (in Hungarian). *Mat. Fiz. Lapok* **48** (1941) 436–452.

[10] P. Winter, Steiner problem in networks: A survey. *Networks* **17** (1987) 129–167.

[11] Note added in proof: We have just learned that Theorem 4.1 was proven independently in the paper "The Computational Complexity of Inferring Rooted Phylogenies by Parsimony," by W. Day, D. Johnson, and D. Sankoff in *Mathematical Biosciences* **81** (1986) 33–42.

Received May 1990
Accepted May 1991