SIAM J. APPL. MATH.
Vol. 32, No. 4, June 1977

THE COMPLEXITY OF COMPUTING STEINER MINIMAL TREES*

M. R. GAREY, R. L. GRAHAM AND D. S. JOHNSON†

Abstract. It is shown that the problem of computing Steiner minimal trees for general planar point sets is inherently at least as difficult as any of the *NP*-complete problems (a well known class of computationally intractable problems). This effectively destroys any hope for finding an efficient algorithm for this problem.

1. Introduction. Let X denote a finite set of n points in the plane. A *spanning tree* $T(X)$ for X is any tree structure that includes every point of X and consists solely of straight line segments (called *edges*) having both endpoints in X. The *length* of $T(X)$, denoted by $l(T(X))$, is defined to be the sum of the (Euclidean) lengths of the edges of $T(X)$. If $T^*(X)$ is a spanning tree that satisfies $l(T^*(X)) \leqq l(T(X))$ for all spanning trees $T(X)$ for X, then $T^*(X)$ is called a (Euclidean) *minimal spanning tree* for X. If $X \subseteq Y$, any spanning tree $T(Y)$ for Y is called a *Steiner tree* for X. It is often possible to choose a superset Y of X in such a way that $l(T^*(Y)) < l(T^*(X))$. If

$$l(T^*(Y)) \leqq l(T^*(Y')) \tag{1}$$

for all sets Y' containing X, then the tree $T^*(Y) = S^*(X)$ is called a (Euclidean) *Steiner minimal tree* (abbreviated by ESMT) for X. An example is shown in Fig. 1.

Minimal spanning trees and Steiner minimal trees arise frequently in problems concerning network design [6], optimal location of facilities [17], and component placement on circuit boards [10], to name a few applications, and considerable effort has gone into developing efficient algorithms for constructing these trees. For constructing a minimal spanning tree on n points in the plane, procedures are now known [16] that require at most $\mathcal{O}(n \log n)$ operations (more precisely, $\mathcal{O}(b^2 n \log n)$ where b is the maximum number of bits used to express a coordinate of a point in X, a bound which takes account of the complexity of arithmetic operations). In contrast, no proposed algorithm for constructing an ESMT for X has been shown to require fewer than exponentially many (in terms of n and b) operations in the worst case. In fact, this was not even known to be a *finite* problem until 1961 [14]. Subsequent work by Cockayne and Schiller [4], Boyce and Seery [3], and others has made it feasible to compute $S^*(X)$ for general

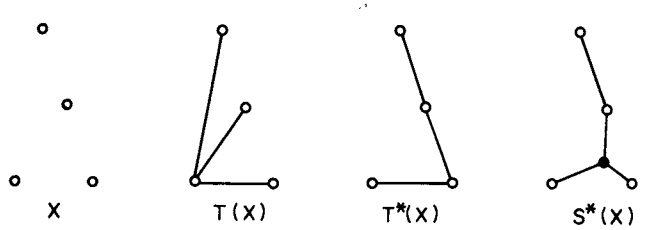

FIG. 1. *Examples of tree types*

* Received by the editors May 10, 1976.
† Bell Laboratories, Murray Hill, New Jersey 07974.

sets X having up to about 15 points. However, these methods are hopelessly inadequate for sets X with 50 or 100 points, and it appears that the ESMT problem may well be inherently intractable.

Thus, it is natural to ask whether the ESMT problem is "*NP*-complete." As described in [2], [12], [13], the *NP*-complete problems form a wide and varying class with the following important properties:

(A) No *NP*-complete problem is known to be solvable by a polynomial time-bounded algorithm.

(B) If any *one* of the *NP*-complete problems could be solved by a polynomial time-bounded algorithm, then *all* *NP*-complete problems could be solved by polynomial time-bounded algorithms.

(Here the execution time of an algorithm is expressed as a function of the number of bits required to express the input.) The class of *NP*-complete problems includes many members notorious for their computational difficulty, such as the traveling salesman problem, the graph chromatic number problem, tautology testing, and clique finding. It is widely believed (though not yet proved) that all its members require exponential time solution algorithms. Hence, *NP*-completeness is a strong argument for inherent intractability.

In this paper we shall show that the ESMT problem is at least as difficult as any of the *NP*-complete problems. For technical reasons, we will not show that the ESMT problem is itself *NP*-complete. Indeed the ESMT problem is perhaps not the problem we *should* show to be *NP*-complete. As defined, it involves idealizations that separate it from the real world of computing. The key observation is that computers cannot manipulate infinite-precision numbers; all numbers in a computation are presented with a limited number of bits and hence are rational. By suitable scaling we may even think of them as integers. Of course, a computation can manipulate symbolic irrational numbers, such as "π" or "$\sqrt{5}$", but whenever an expression involving these symbolic irrationals is evaluated, they must be rounded to rational numbers in order for the computation to terminate.

There are two sources of irrational numbers in the ESMT problem. The first is that, even if all points in X have integer coordinates, it is possible that in the Steiner minimal tree $S^*(X) = T^*(Y)$ one or more points of Y may have irrational coordinates. A second and more crucial source of irrationality is the Euclidean metric, where the distance $d(x, y)$ between two points $x = (x_1, x_2)$ and $y = (y_1, y_2)$ is defined by

$$d(x, y) = ((x_1 - y_1)^2 + (x_2 - y_2)^2)^{1/2}.$$

Since the length of an edge may be irrational, even if both endpoints have integer coordinates, the length of a particular tree can be a sum of many different irrational numbers. This might make it difficult to compare the length of two different trees, since no polynomial bound is known on the accuracy required for correctly making the comparison.

In view of these comments, one might wonder whether the *only* source of difficulty in the ESMT problem involves computing with irrational numbers. In this case a proof that the ESMT problem is difficult would have little practical significance, since one would be perfectly willing to specify in advance the

precision with which computations are to be carried out. To avoid this criticism, we shall modify the ESMT problem in a manner more appropriate to real-world computing and it is this modified problem that we prove to be *NP*-complete. It will follow from our proof that the ESMT problem itself is at least as difficult as any *NP*-complete problem and that irrational numbers are *not* the sole source of its difficulty. Furthermore, in the unlikely event that all *NP*-complete problems can be solved with polynomial time algorithms, we will know that Steiner minimal trees can be determined to within any desired accuracy in polynomial time, even though the ESMT problem might still be hard.

There are two points at which we introduce approximations into the problem. First, we require that all points of X and the additional points of $Y-X$ have only integer coordinates. Second, we replace the Euclidean metric by a "discretized" Euclidean metric in which the length of an edge joining two points x and y is taken to be $d'(x, y) = \lceil d(x, y) \rceil$ (where $\lceil \alpha \rceil$ is the least integer not less than α). The discrete Euclidean length $l'(T)$ of a tree T is then the sum of the discrete Euclidean lengths of its edges.

Thus we obtain a discrete version which, by appropriate scaling, can be made to approximate the original ESMT problem to within any desired degree of accuracy and which accurately reflects the manner in which Steiner minimal trees must be computed in practice. Stated as a feasibility problem, as required for proofs of *NP*-completeness [2], [12], [13], it becomes the following discrete Euclidean Steiner minimal tree (DESMT) problem:

> Given a set X of integer-coordinate points in the plane and a positive integer L, does there exist a set $Y \supseteq X$ of integer-coordinate points such that some spanning tree T for Y satisfies $l'(T) \leqq L$?

It can be shown that this feasibility question can be answered in polynomial time if and only if the corresponding optimization problem can be solved in polynomial time.

The main result of this paper is that the DESMT problem is *NP*-complete. We prove this by reducing to it a known *NP*-complete problem, that of "exact cover by 3-sets". However, having emphasized all the practical reasons for directing our attention to the discrete problem, we must now point out that it will be more convenient in our proof to work in the original, infinite-precision Euclidean domain. This allows us to simplify the arguments by using a number of geometrical lemmas about the ESMT problem which would be more difficult to prove for the DESMT problem. Thus the reader should be warned that our basic construction will involve points having fractional and even irrational coordinates. However, we shall prove a result strong enough to insure that the scaling and rounding needed to translate our construction back to the discrete case cannot change the outcome of the comparison to L.

We begin with a discussion of some elementary properties of ESMT's and then proceed to the more powerful lemmas and our construction.

2. Preliminaries. When there is no danger of confusion, we shall usually abbreviate $T(X)$, $T^*(X)$, and $S^*(X)$ by T, T^*, and S^*. If T is a spanning tree for X and $u, v \in X$, the (unique) path between u and v in T will be denoted by $P_T(u, v)$. The *maximum length* of an edge in $P_T(u, v)$ is denoted by $m(P_T(u, v))$. The

maximum edge length in T is denoted by $m(T)$. The following result appears in [8].

LEMMA 1. *For any set X and points $u, v \in X$,*

$$m(P_T(u, v)) \geqq m(P_{T^*}(u, v)) \geqq m(P_{S^*}(u, v)). \tag{2}$$

Proof. Suppose there exists a spanning tree T for X and $u, v \in X$ such that

$$m(P_T(u, v)) < m(P_{T^*}(u, v)).$$

Then there exists an edge e of $P_{T^*}(u, v)$ which is longer than every edge of $P_T(u, v)$. If we delete this edge e from T^*, the resulting set consists of exactly two connected components. It is easily seen that the addition of some edge e' from $P_T(u, v)$ must rejoin these two components, forming a spanning tree T' for X with

$$\begin{aligned} l(T') &= l(T^*) - l(e) + l(e') \\ &\leqq l(T^*) - m(P_{T^*}(u, v)) + m(P_T(u, v)) \\ &< l(T^*) \end{aligned}$$

which contradicts the definition of a minimal spanning tree T^* for X. This proves the first inequality in (2). The second inequality follows similarly. □

Let $S^* = T^*(Y)$ be a Steiner minimal tree for X with $|Y|$ as small as possible. A point $x \in X$ is called a *regular point* of S^*; a point $s \in Y - X$ is called a *Steiner point* of S^*. The following basic results about such Steiner minimal trees are easily proved [7].

Fact 1. Every Steiner point s of S^* has degree 3 and each of the three edges of S^* incident to s meets the other two at angles of 120°.

Fact 2. If X has n points, then S^* has at most $n-2$ Steiner points.

Fact 3. If S^* has any Steiner points, then S^* has some Steiner point adjacent to at least two regular points of S^*.

Fact 4. No two edges of S^* that share a common endpoint meet at an angle of less than 120°.

Fact 5. Two edges of S^* intersect only at a common endpoint.

3. Geometrical considerations. Before proceeding to the main construction, we shall first prove several useful geometrical results concerning Steiner minimal trees (cf. [8]). Define the semi-infinite strip W (shown in Fig. 2) by

$$W = \{(x, y) : |x| \leqq 1,\ y \geqq |x|/\sqrt{3}\}.$$

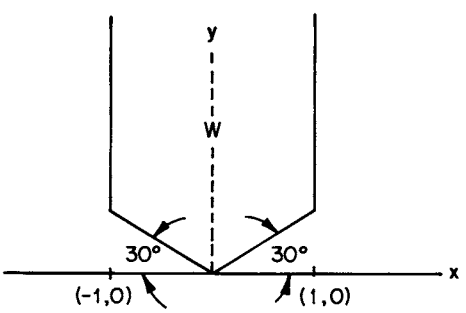

FIG. 2. *The semi-infinite strip W*

LEMMA 2. *Let S^* be a Steiner minimal tree for X with $m(S^*) \leqq 1$ and suppose $s_0 = (0, 0)$ is a Steiner point of S^*. Then for some regular point x of S^*,*

$$P_{S^*}(s_0, x) \subseteq W.$$

Proof. Since $s_0 = (0, 0)$ is a Steiner point, let us extend each of the three edges of S^* incident to s_0 infinitely in both directions, forming six directed half-lines all emanating from s_0. We denote these by L_i, $0 \leqq i \leqq 5$, where the clockwise angle between L_i and L_{i+1} is 60°. We may assume that L_0 is chosen in the first quadrant so that the angle θ between L_0 and the positive y-axis satisfies $0 \leqq \theta < 60°$ (see Fig. 3).

We partition W into three sets, as illustrated in Fig. 4, based on the angle θ. W_1 is the set of points in W which would be moved out of W if they were translated in the direction of L_5 for a distance of 1. W_2 is the set of points in W which would be moved out of W if they were translated in the direction of L_0 for a distance of 1. W_3 is the set of points in W which would *remain* in W if translated for a distance of 1 in the direction of either L_0 or L_5. The precise specifications are as follows (see Fig. 4):

$$W_1 = \{(x, y) \in W : x < -1 + \cos(\theta + 30°)\},$$

$$W_2 = \{(x, y) \in W : x > 1 - \sin\theta\},$$

$$W_3 = W - (W_1 \cup W_2).$$

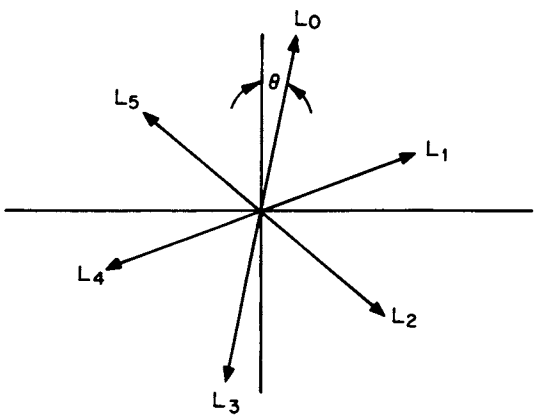

FIG. 3. *Half-lines from s_0*

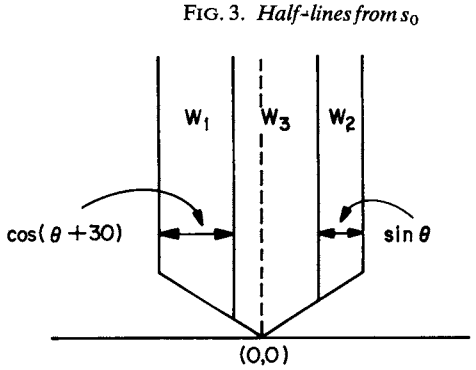

FIG. 4. *Partitioning W*

Observe for future reference that the width of the strip W_3 is $2-\cos(\theta+30°)-\sin\theta$, which is at least 1, with that minimum value occurring at $\theta=30°$.

We now show how to generate a sequence of points which form a path of S^* contained entirely within W. This sequence will terminate with a regular point. The first point in the sequence is the Steiner point $s_0=(0,0)$. In general, suppose the sequence generated so far is $(s_0, s_1, \ldots, s_t)$ where each s_i, $0\leqq i\leqq t$, is a distinct Steiner point of S^* in W and each line segment $\langle s_i, s_{i+1}\rangle$, $0\leqq i<t$, is an edge of S^*. An inductive application of Fact 1 implies that the direction of the edges leaving s_t are all from the set $\{L_i : 0\leqq i\leqq 5\}$. We use the following rules for choosing the next point $s_{t+1}=v$ from those points that are adjacent to s_t in S^*:

If $s_t\in W_1$, choose v so that $\langle s_t, v\rangle$ is an edge of S^* with the same direction as L_0 or L_1.

If $s_t\in W_2$, choose v so that $\langle s_t, v\rangle$ is an edge of S^* with the same direction as L_4 or L_5.

If $s_t\in W_3$, choose v so that $\langle s_t, v\rangle$ is an edge of S^* with the same direction as L_5 or L_0.

Note that, by Fact 1, such a v must exist and is uniquely determined. It remains to be shown that (a) $v\in W$ and (b) $v\notin\{s_0, s_1, \cdots, s_t\}$.

(a) By the selection rules and our definition of W_1, W_2, and W_3, the only way v could be outside W would be if $v=(v_x, v_y)$ with $|v_x|\leqq 1$ and $v_y/|v_x|<1/\sqrt{3}$. If $v_x>0$, this would imply that $s_t\in W_1$ and $\langle s_t, v\rangle$ has direction L_1. If $v_x<0$, the implication would be that $s_t\in W_2$ and $\langle s_t, v\rangle$ has direction L_4. We shall treat only the former case, as the latter is symmetric.

So suppose $s_t\in W_1$ and $v\notin W$. Consider the unit length line segment in the direction of L_1 from c to d, where $c=(-1+\cos(\theta+30°), (1-\cos(\theta+30°))/\sqrt{3})$. Since this segment is parallel to $\langle s_t, v\rangle$, is at least as long, and since c is below and to the right of every point in W_1, d must be below and to the right of v. Hence the coordinates of $d=(d_x, d_y)$ must also satisfy $d_y/d_x<1/\sqrt{3}$. However, a simple calculation yields

$$d_x=-1+\cos(\theta+30°)+\cos(\theta-30°),$$

$$d_y=\frac{1}{\sqrt{3}}-\frac{1}{\sqrt{3}}\cos(\theta+30°)-\sin(\theta-30°).$$

The reader may readily verify that these values satisfy $d_y/d_x\geqq 1/\sqrt{3}$, with equality only at $\theta=30°$. Thus we have a contradiction and $v\in W$ as desired.

(b) If $v\in\{s_0, s_1, \cdots, s_t\}$, then it must be the case that $v=s_{t-1}$, since the Steiner tree S^* contains no cycles. But this implies that $\langle s_t, v\rangle$ must be in the opposite direction to $\langle s_{t-1}, s_t\rangle$. The only opposing directions that can be chosen by our rules are L_4 and L_1. Our rules then imply that $s_{t-1}\in W_2$ and $s_t\in W_1$ or *vice versa*. However, this implies that the edge $\langle s_{t-1}, s_t\rangle$ is longer than 1, the minimum possible width for W_3. This is a contradiction, since $\langle s_{t-1}, s_t\rangle$ is an edge of S^* and $m(S^*)\leqq 1$.

Thus, s_{t+1} is a distinct new point of S^* in W. By induction, we can continue adding new points to our sequence until a regular point $v=x\in X$ is reached. Since X is finite and, by Fact 2, S^* can have at most $|X|-2$ Steiner points, the sequence must terminate with such a point. Hence, W contains the desired path $P_{S^*}(s_o, x)$. □

Let $\alpha \geqq 0$ be arbitrary and let us truncate W to form $W(\alpha)$ defined by $W(\alpha) = \{(x, y) \in W: y \leqq \alpha\}$. The preceding arguments actually imply the following more precise result.

COROLLARY 1. *Let S^* be a Steiner minimal tree for X with $m(S^*) \leqq 1$ and suppose $s_0 = (0, 0)$ is a Steiner point of S^*. Then for any $\alpha \geqq 0$ either $W(\alpha+1)$ contains a regular point of S^* or $W(\alpha+1) - W(\alpha)$ contains a Steiner point of S^*.*

If s is a Steiner point of a Steiner minimal tree S^* for X and $\langle s, v\rangle$ is an edge of S^*, let $H(s, v)$ denote the closed regular hexagonal region of side 1 which is bisected by the line through s and v and which intersects $\langle s, v\rangle$ in the single point s (see Fig. 5).

LEMMA 3. *Let S^* be a Steiner minimal tree for X with $m(S^*) \leqq 1$. If s is a Steiner point of S^* and $\langle s, v\rangle$ is an edge of S^*, then $H(s, v)$ contains a regular point of S^*.*

Proof. Suppose $H(s, v)$ contains no regular point of S^*. By Fact 1, S^* must have edges $\langle s, s_1\rangle$ and $\langle s, s_2\rangle$ that lie on the boundary of $H(s, v)$, since by hypothesis no edge of S^* is longer than 1. Also, the endpoints, s_1 and s_2, of these edges must be Steiner points of S^* since, by assumption, $H(s, v)$ contains no regular point of S^*. Because s_1 and s_2 are Steiner points, there must exist Steiner points t_1 and t_2 of S^* such that $\langle s_1, t_1\rangle$ and $\langle s_2, t_2\rangle$ are edges of S^* lying entirely inside $H(s, v)$ and which are parallel to $\langle v, s\rangle$. Finally, there must exist Steiner points u_1 and u_2 of S^* such that $\langle t_1, u_1\rangle$ and $\langle t_2, u_2\rangle$ are edges of S^* lying entirely inside $H(s, v)$ and which are parallel to $\langle s, s_2\rangle$ and $\langle s, s_1\rangle$ respectively (see Fig. 5).

For $i \in \{1, 2\}$, let L_i denote the line passing through u_i parallel to $\langle s, s_i\rangle$. By Fact 5, the edges $\langle t_1, u_1\rangle$ and $\langle t_2, u_2\rangle$ cannot intersect. Hence, either u_2 lies strictly above L_1 or u_1 lies strictly above L_2 or both. Suppose without loss of generality that u_2 lies above L_1. We apply Lemma 2 to the Steiner point u_1. Since u_2 (and therefore t_2) lies *strictly* above L_1, there is an orientation of the region W (with u_1 playing the role of $s_0 = (0, 0)$) so that $t_1 \notin W$, $t_2 \notin W$, and $s \in W$. By Lemma 2, there must exist a regular point v' of S^* such that the path $P_{S^*}(u_1, v')$ lies entirely in W. However, since $t_1, t_2 \notin W$ then by Fact 5, no edge of this path can intersect any of

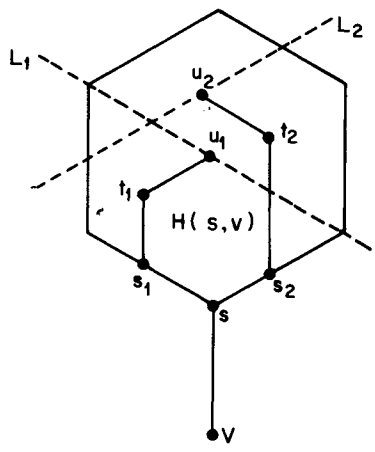

FIG. 5. *The region $H(s, v)$*

the edges $\langle s_1, t_1\rangle$, $\langle s, s_1\rangle$, $\langle s, s_2\rangle$, $\langle s_2, t_2\rangle$. Thus $P_{S^*}(u_1, v') \subseteq H(s, v)$ which implies $v' \in H(x, v)$. This contradicts the assumption that no regular point lies in $H(s, v)$ and the lemma is proved. □

For a point x in the plane, let D_x denote the disc of all points at a distance of at most 2 from x. The following corollary is an immediate consequence of Lemma 3.

COROLLARY 2. *Let s be a Steiner point of a Steiner minimal tree S^* for X with $m(S^*) \leqq 1$. Then D_s contains a regular point of S^*.*

Finally, consider the region P, called a *probe*, shown in Fig. 6. P is formed by taking the union of a copy of $W(10)$ together with the set of all points at a distance of at most 2 from some point of $W(10)-W(9)$. The point $t(P)$ at the (central) 120° angle of the probe is called the *tip* of the probe.

LEMMA 4. *For any placement (i.e., translation and rotation) of a probe P in the plane which contains no point of X, the tip $t(P)$ of P cannot be a Steiner point of S^*.*

Proof. Suppose $t(P)$ is a Steiner point of S^* but that P contains no point from X. By Corollary 1, since the corresponding copy of $W(10) \subseteq P$ contains no regular point of S^* (i.e., point of X), the region $W(10)—W(9)$ must contain some Steiner point s of S^*. By Corollary 2, some regular point $x \in X$ of S^* must be within distance 2 from s, so that $x \in P$. This contradicts the hypothesis that $X \cap P = \varnothing$ and Lemma 4 is proved. □

(Note that the length 10 of the truncated strip $W(10)$ is not essential for this proof, but has been chosen merely for convenience in what follows.)

4. The configuration $X(\mathcal{F})$. In this section we begin the reduction of a known *NP*-complete problem to the discrete Euclidean Steiner minimal tree problem. The problem we reduce is that of *exact cover by 3-sets*, abbreviated X3C. As input to this problem we are given a family $\mathcal{F}$ of 3-element subsets $F_1, F_2, \cdots, F_t$ of a set F of $3n$ elements, which we take without loss of generality to be $\{1, 2, 3, \cdots, 3n\}$. The problem X3C is to decide whether there is a subfamily $\mathcal{F}' \subseteq \mathcal{F}$ such that

(i) distinct elements of $\mathcal{F}'$ are disjoint; and,

(ii) the elements of $\mathcal{F}'$ cover F, i.e.,

$$\bigcup_{F' \in \mathcal{F}'} F' = F.$$

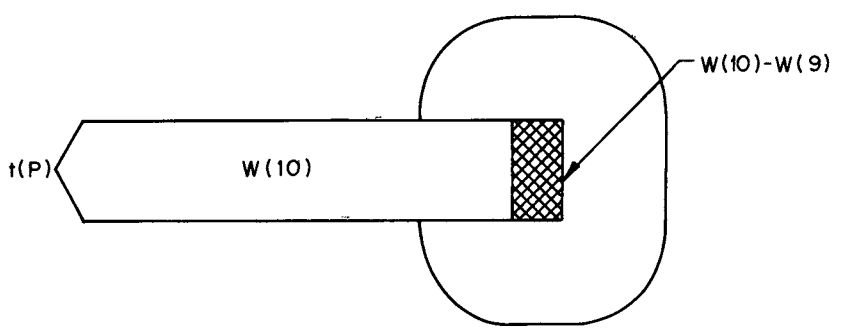

FIG. 6. *The probe P*

The problem X3C is known to be *NP*-complete, and in fact contains as a special case the problem of 3-dimensional matching shown to be *NP*-complete in [12]. In the remainder of this section we shall describe a configuration $X = X(\mathcal{F})$ of points in the plane which will serve as an intermediary in our eventual transformation of an instance $\mathcal{F}$ of X3C into an instance of DESMT. We shall remain with X in the ordinary Euclidean domain for quite some time, with the translation to the discrete domain not coming until § 7, after all of the important properties of X have been derived. We begin by assuming that both n and t exceed 1, since otherwise the X3C problem for $\mathcal{F}$ would be trivial.

We shall build $X(\mathcal{F})$ up in stages. A basic unit in the construction is shown in Fig. 7. This configuration is called a *standard row* and consists of 100 equally spaced points lying in a straight line with adjacent points separated by a distance of 1/10. The points a and $\bar{a}$ are called the *endpoints* of the row (see Fig. 7(a)). We shall denote a standard row schematically as in Fig. 7(b).

We now combine standard rows in various ways to form the next units in the construction. In Fig. 8(a) we show a configuration Q composed of 4 standard rows. The endpoints a, b, c, d, called the *active* points of Q, are the vertices of a square of side 1 and the standard rows lie on the extended diagonals of that square. We shall call Q a *square* and denote it schematically as in Fig. 8(b).

In Fig. 9(a), we show a configuration $R(\varepsilon)$ composed of 3 standard rows. The endpoints a, b, c are the *active* points of $R(\varepsilon)$ and form the vertices of an equilateral triangle with side length $1-\varepsilon$ for some ε, $0 \leqq \varepsilon < 1/200$, to be specified later. The three standard rows radiate out from the three vertices of the triangle and lie on the half-lines bisecting the exterior angles of the triangle. When $\varepsilon > 0$, $R(\varepsilon)$ is called a *small triangle* and is denoted schematically as in Fig. 9(b). When $\varepsilon = 0, R(0) = R$ is called a *standard triangle* and is denoted schematically as in Fig. 9(c).

FIG. 7. *A standard row*

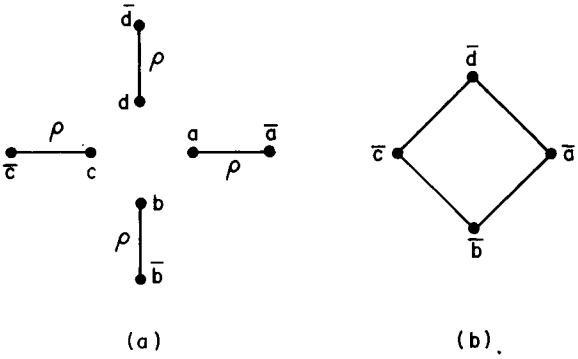

FIG. 8. *A square Q*

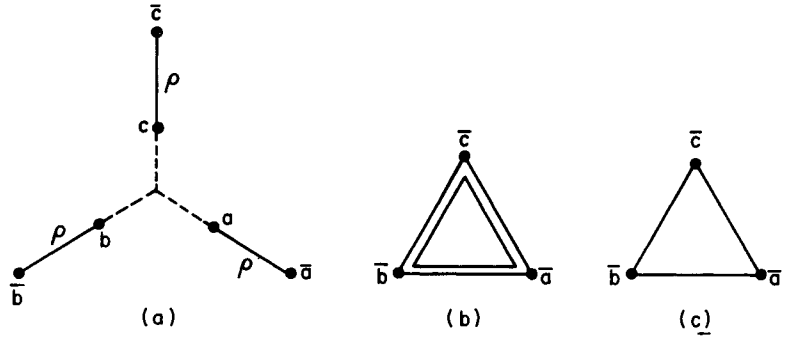

FIG. 9. *A triangle $R(\varepsilon)$*

Next we consider two configurations which have no active points. An *angle* A, shown in Fig. 10(a), consists of two standard rows having a common endpoint and which lie on lines meeting at 30°. A junction J, shown in Fig. 10(c), consists of 3 standard rows having a common endpoint and which lie on lines that all meet each other at 120°. These are represented schematically as shown in Figs. 10(b) and 10(d).

Finally we form a fundamental configuration C, called a *crossover* (for reasons that will become clear later) by combining some of the previous units. The crossover C consists of two standard triangles R, two junctions J, four angles A, and a number of rows of points (denoted by $\bar{\rho}$'s) used to interconnect these components (see Fig. 11(a)). Each $\bar{\rho}$ is called a *long row* and consists of at least 1000 (not necessarily equally spaced) points lying on a straight line with distances between consecutive points ranging between 1/11 and 1/10. A long row $\bar{\rho}$ shares an endpoint with the corresponding row ρ of a component it meets and ρ and $\bar{\rho}$ are collinear (see Fig. 12 for a typical connection). The exact positions of points in each $\bar{\rho}$ are chosen so that the topological arrangement of C shown in Fig. 11(a) is geometrically possible. The schematic representation for C is shown in Fig. 11(b). If C is constructed using a small triangle $R(\varepsilon)$ in place of the *upper* standard triangle R (with the lower triangle still standard), the corresponding configuration $C(\varepsilon)$ is called a *warped* crossover and is denoted as in Fig. 11(c).

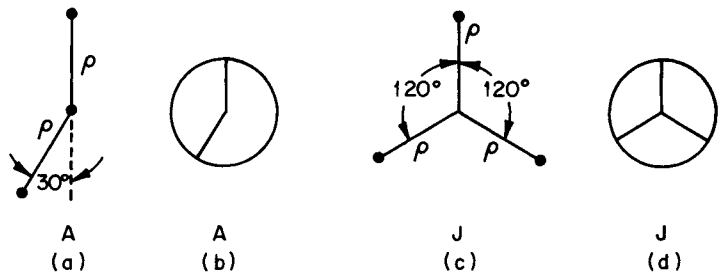

FIG. 10. *An angle A and junction J*

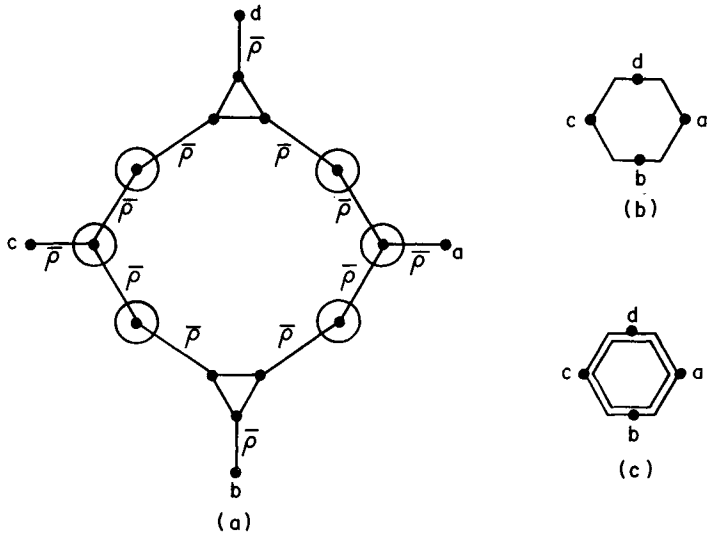

FIG. 11. *A crossover C*

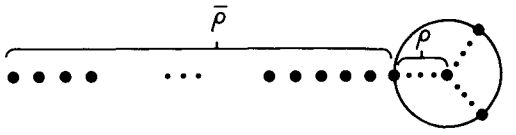

FIG. 12. *A typical connection*

We can combine two angles, a junction, and five long rows to form a *terminator* Ω, as shown in Fig. 13(a). We classify these as *downward* and *upward* terminators, abbreviated respectively as in Figs. 13(b) and 13(c), depending on whether the point a is above or below the other points of the terminator.

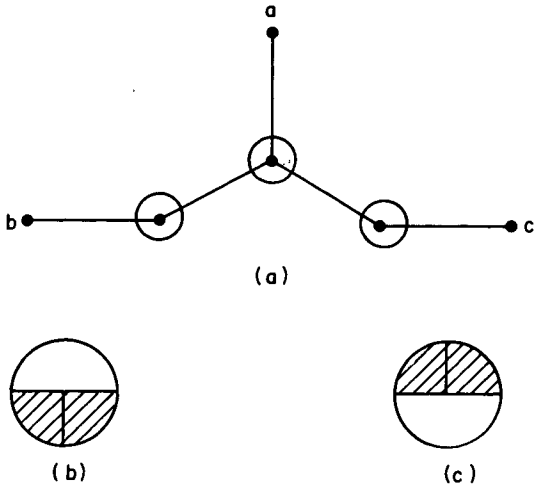

FIG. 13. *A terminator* Ω

We are now ready to construct $X(\mathcal{F})$. It will be formed by connecting crossovers, squares, and terminators by long rows $\bar{\rho}$. To begin with, $X(\mathcal{F})$ has a chain of $t+1$ upward terminators Ω_i, $0 \leqq i \leqq t$, joined by long rows as shown in Fig. 14. Each terminator Ω_k, $1 \leqq k \leqq t$, is connected to a square Q_k which itself is connected to two other upward terminators Ω'_k and Ω''_k as in Fig. 15.

We associate with each square Q_k the subset $F_k = \{a_k, b_k, c_k\} \in \mathcal{F}$. Going up from Q_k will be a chain of crossovers $C_k(i)$ for $0 \leqq i < a_k$ connected by long rows. All the $C_k(i)$ are ordinary crossovers except for $C_k(a_k - 1)$, which is a *warped* crossover $C(\varepsilon)$. Located above $C_k(a_k - 1)$ is a downward terminator $\bar{\Omega}_k$. Similarly, Ω'_k and Ω''_k each have rising chains of crossovers $C'_k(i)$, $0 \leqq i < b_k$, and $C''_k(j)$, $0 \leqq j < c_k$, respectively, with $C'_k(b_k - 1)$ and $C''_k(c_k - 1)$ being the warped crossovers in these chains, which themselves are connected above to downward terminators $\bar{\Omega}'_k$ and $\bar{\Omega}''_k$. All the $C_k(i)$, $C'_k(i)$, $C''_k(i)$ are at the same horizontal level, called the ith *level*, for $1 \leqq k \leqq t$. Also $\bar{\Omega}_k$ (lying above $C_k(a_k - 1)$) is at the a_kth level, with $\bar{\Omega}'_k$ and $\bar{\Omega}''_k$ located at levels b_k and c_k, respectively. A single downward terminator Ω'_0 at level 0 is connected directly to the upward terminator Ω_0 below. Finally, all components at the same level are connected in a chain by long rows. We illustrate a partition of the general format in Fig. 16.

As an example, we show in Fig. 17 a schematic representation of $X(\mathcal{F}_0)$ for the family $\mathcal{F}_0 = \{\{1, 2, 4\}, \{2, 3, 6\}, \{3, 5, 6\}\}$. The interconnecting lines all represent suitably chosen long rows $\bar{\rho}$.

In general $X(\mathcal{F})$ will contain $6t+2$ terminators, t squares, $3t$ warped crossovers and at most $9nt - 3t$ ordinary crossovers. It is not hard to see that the placement of the various components can be arranged so that the number of points in $X(\mathcal{F})$ is bounded by a polynomial in n and t.

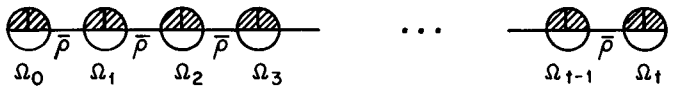

FIG. 14. *A chain of terminators*

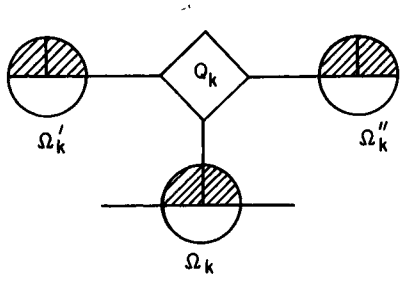

FIG. 15

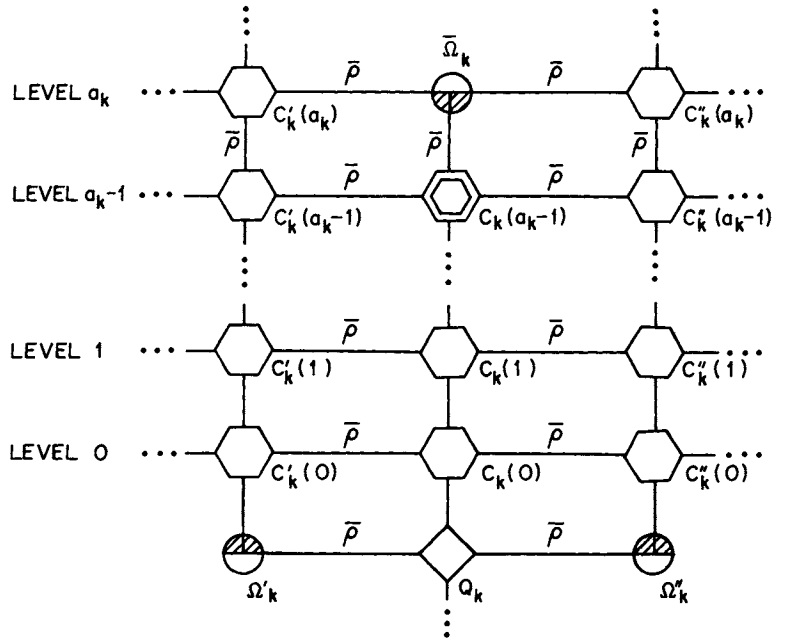

FIG. 16. *A portion of* $X(\mathcal{F})$

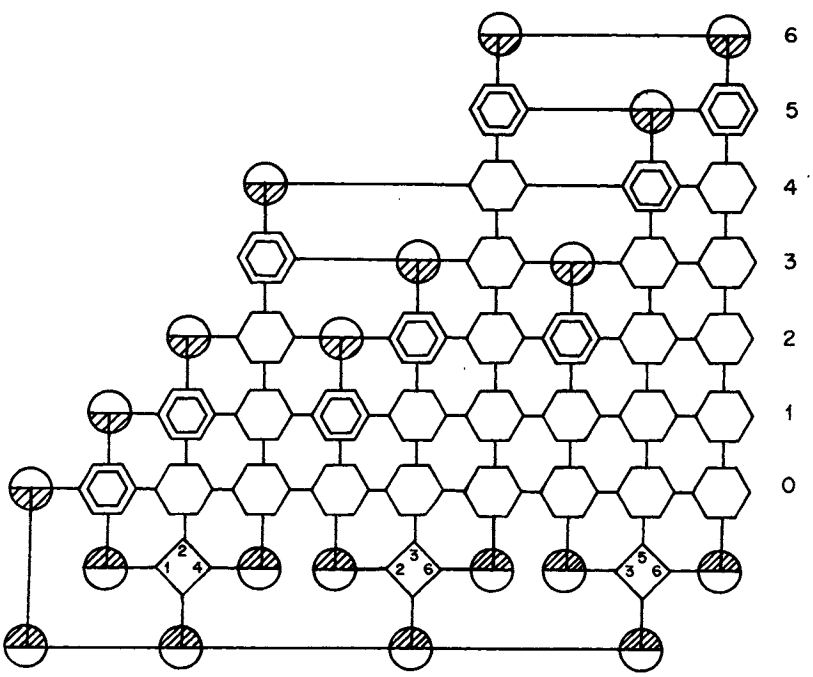

FIG. 17. $X(\mathcal{F}_0)$

5. Properties of $S^*(X(\mathcal{F}))$. We assume an arbitrary (fixed) family $\mathcal{F} = \{F_1, F_2, \cdots, F_t\}$ of 3-sets $F_k \subseteq \{1, 2, \cdots, 3n\}$ is given. Let S^* be a Steiner minimal tree for $X = X(\mathcal{F})$. In this section we shall show that S^* must have a very restricted structure. In particular, the only points of X that can be adjacent to a Steiner point in S^* are the *active* points of the triangles and squares.

To begin with, it is easy to see from the construction of X that the longest edge in a minimal spanning tree T^* for X has length 1, i.e., $m(T^*) = 1$. Hence, by Lemma 1,

$$m(S^*) \leqq 1. \tag{3}$$

Thus, Lemma 4 can be applied to delimit the possible locations of Steiner points in S^*. In fact, for almost every point p of the plane not belonging to $X(\mathcal{F})$, the probe of Lemma 4 can be placed with p at the probe tip and no points of X inside the probe. In particular, it follows that there are just two types of possibilities for a Steiner point s of S^*:

(i) s is the 120° point of the three active points of some triangle R or $R(\varepsilon)$ (see Fig. 18(a));

(ii) s is located in a certain region $\mathcal{R}$ (bounded by 4 elliptical arcs) that lies in the unit square determined by the four active points of some square Q (see Fig. 18(b)).

We can determine the boundary of $\mathcal{R}$ in case (ii) by calculating the locus of locations of the vertex v of the Lemma 4 probe as it rotates through 30° while the two corners adjacent to v move on the two coordinate axes (see Fig. 19).

A simple calculation will verify that the arc in the first quadrant determined by v is a portion of an ellipse given by the equation

$$x^2 + xy\sqrt{3} + y^2 = 1/3. \tag{4}$$

The intercepts of the arc are the points $(0, 1/\sqrt{3})$ and $(1/\sqrt{3}, 0)$.

Having located the potential Steiner points of S^*, it is now possible to draw strong conclusions about which regular points are linked together by edges.

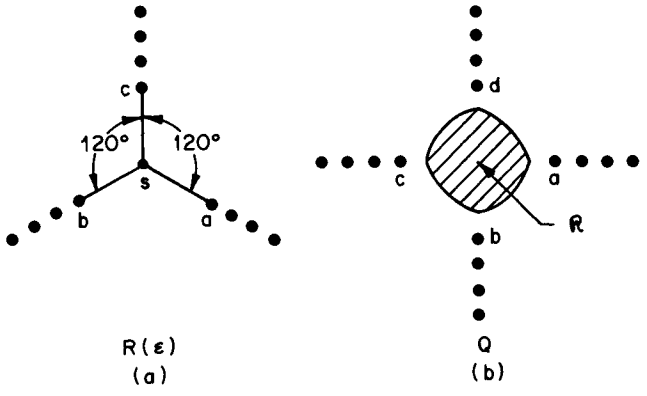

FIG. 18. *Potential Steiner points*

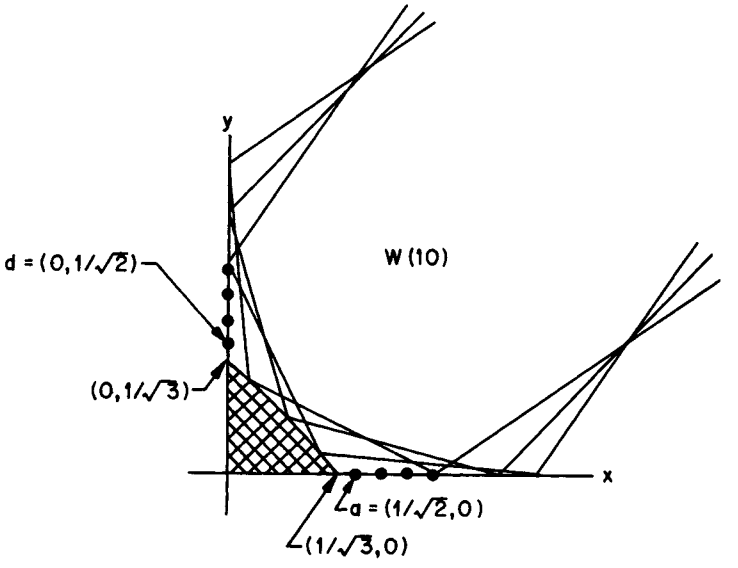

FIG. 19

LEMMA 5. *If $x, y \in X$ and the distance between x and y does not exceed* $1/10$, *then $\langle x, y\rangle$ is an edge of S^*.*

Proof. Since there *is* a spanning tree for X which contains the edge $\langle x, y\rangle$, we know by Lemma 1 that $m(P_{S^*}(x, y)) \leqq 1/10$. If $\langle x, y\rangle$ is not contained in S^*, then x and y are joined by a path in S^* that contains no edge longer than $1/10$ and that does not use $\langle x, y\rangle$. By the construction of X, no such path can consist solely of edges joining pairs of regular points. Thus $P_{S^*}(x, y)$ must contain a Steiner point and hence an edge joining some regular point to a Steiner point. However, by our limitations on the locations of possible Steiner points, the length of such an edge must be at least $1/\sqrt{2} - 1/\sqrt{3} = 0.1297\cdots > 1/10$. This contradicts the fact that $m(P_{S^*}(x, y)) \leqq 1/10$, proving the lemma. □

LEMMA 6. *Any edge of S^* longer than* $1/10$ *that joins two points of X must join two active points of the same square or triangle.*

Proof. The edges that are known to be in S^* by Lemma 5 form disjoint subtrees or "components" of S^* that include all points of X. No two points of X belonging to the same component can be joined by an additional edge (longer than $1/10$), since that would form a cycle. But the construction of X, since $\varepsilon < 1/200$ (see Fig. 20), insures that the only pairs of points in different components which are separated by distance 1 or less are active points of the same square or triangle. Since $m(S^*) \leqq 1$, the lemma follows. □

LEMMA 7. *If a point $x \in X$ is adjacent to a Steiner point of S^*, then x is an active point of either a square Q or a triangle R or $R(\varepsilon)$.*

Proof. Since $m(S^*) \leqq 1$, any point x adjacent to a Steiner point s must be within distance 1 of one of the locations where a Steiner point can occur. These possibilities are shown in Fig. 21.

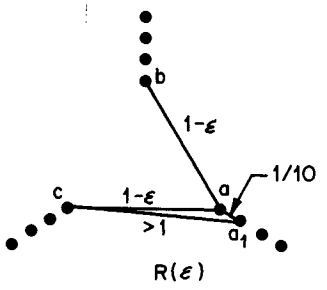

FIG. 20. *Edge* $\langle a_1, c\rangle$ *longer than* 1

If x is not an active point, then there must be a point $y \in X$ in the same standard row at distance $1/10$ from x and closer to the corresponding active point than x is. By Lemma 5, S^* contains the edge $\langle x, y\rangle$. But then, in either case, S^* contains two edges, $\langle x, y\rangle$ and $\langle x, s\rangle$, that meet at a common point x at an angle less than 120°. This contradicts Fact 4 and it follows that x must be an active point. □

In light of Lemmas 6 and 7, let us call the triangles and squares of X the *active regions*, as these are the only regions in which Steiner points and edges longer than $1/10$ can occur. Since $m(S^*) \leqq 1$, then by Fact 1, we see that there are only a very limited number of possible configurations of S^* within each active region. We catalog these possibilities in Fig. 22, where we show the possible configurations of edges joining active points and (possibly) Steiner points within each type of active region. Only one representative is given for symmetric configurations.

In Table 1 we list the total length of the edges in each configuration shown in Fig. 22.

6. The value of $l(S^*)$. Before providing estimates on the total length of S^*, we must first pick a specific value for ε. Let $\hat{C} = \hat{C}(X)$ denote the number of crossovers in $X = X(\mathcal{F})$. As noted earlier, $\hat{C} \leqq 9nt$. We specify ε as follows:

$$\varepsilon = 1/(200\, nt). \tag{5}$$

A simple calculation shows that this choice implies that

$$(L(\beta_1) - L(\alpha_1))\hat{C} < \tfrac{2}{3}L(\gamma_1) - L(\alpha_1), \tag{6}$$

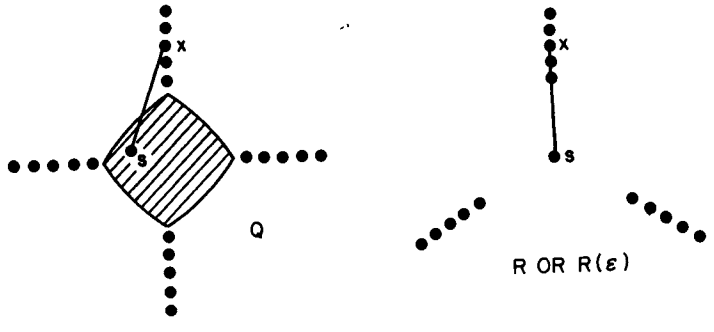

FIG. 21. *Possibilities for Lemma* 7

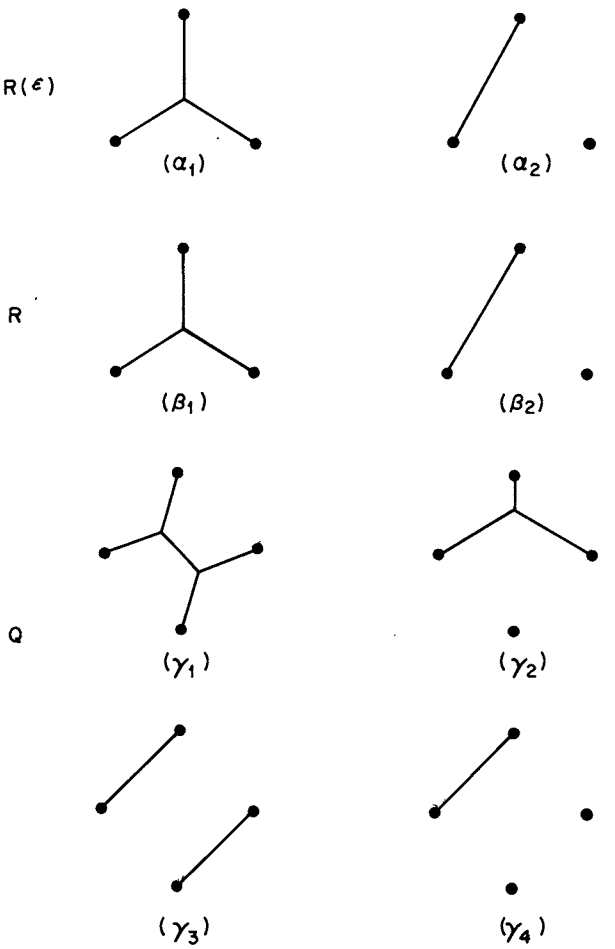

FIG. 22. *Possible configurations in active regions*

TABLE 1
Lengths of configurations in Fig. 22

Configuration	Total Length
α_1	$L(\alpha_1)=(1-\varepsilon)\sqrt{3}$
β_1	$L(\beta_1)=\sqrt{3}$
γ_1	$L(\gamma_1)=1+\sqrt{3}$
α_2	$L(\alpha_2)=1-\varepsilon$
β_2	$L(\beta_2)=1$
γ_2	$L(\gamma_2)=(1+\sqrt{3})/\sqrt{2}$
γ_3	$L(\gamma_3)=2$
γ_4	$L(\gamma_4)=1$

and also satisfies the inequality $\varepsilon < 1/200$ required by the construction of $R(\varepsilon)$.

Now, by Lemma 5, S^* contains all the edges joining adjacent points in the same standard or long row. Let L_0 denote the total length of all these edges. If we consider the graph composed of just these edges, we observe that it is made up of $N = 2\hat{C} + 3n + 1$ connected components. In S^* all these N components must be joined together. Lemmas 6 and 7 tell us that such interconnections can only be made in the active regions using the configurations of Fig. 22. We use L^+ to denote the total length of these additional edges, so that $l(S^*) = L_0 + L^+$.

THEOREM 1. *If $\mathcal{F}$ has an exact cover, then*

$$l(S^*) \leqq 3nL(\alpha_1) + (\hat{C} - 3n)L(\beta_1) + nL(\gamma_1) + L_0; \tag{7}$$

and if $\mathcal{F}$ does not contain an exact cover, then

$$l(S^*) \geqq 3nL(\alpha_1) + (\hat{C} - 3n)L(\beta_1) + nL(\gamma_1) + L_0 + \varepsilon. \tag{8}$$

Proof. Suppose there exists an exact cover $\mathcal{F}' = \{F_{i_1}, F_{i_2}, \cdots, F_{i_n}\}$. We shall construct a Steiner tree S for X satisfying the bound of (7) and thus implying that (7) also holds for S^*. We begin by including all the edges joining adjacent points in the same standard or long row, for a total length of L_0. The remaining edges are constructed as follows:

(i) Form a type γ_1 configuration in each square Q_{i_k}, $1 \leqq k \leqq n$.

(ii) In each crossover belonging to one of the three columns of crossovers extending up from some Q_{i_k}, $1 \leqq k \leqq n$, form a type α_1 or type β_1 configuration in the *upper* triangle of that crossover.

(iii) Form a type β_1 configuration in the *lower* triangle of each crossover not considered in (ii).

Since each column of crossovers contains exactly one warped crossover, exactly $3n$ type α_1 configurations are formed under (ii) and (iii). Thus the total length of the additional edges is $3nL(\alpha_1) + (\hat{C} - 3n)L(\beta_1) + nl(\gamma_1)$, as desired.

To see that we have indeed constructed a Steiner tree for X, it is sufficient to show that each of the N components is joined by a path to the *backbone* component composed of terminators $\Omega_0', \Omega_0, \Omega_1, \cdots, \Omega_t$ and the long rows joining them (see Figs. 14–17).

First we observe that, since every crossover has either a type α_1 or type β_1 configuration in one of its triangles, all the components on level i, $0 \leqq i \leqq 3n$, are joined together into an overall *level i component.* Moreover the level 0 component is joined to the backbone since terminator Ω_0' is both at level 0 and in the backbone. If $F_k = \{a_k, b_k, c_k\} \in \mathcal{F}'$, we have by (ii) that the downward terminators $\bar{\Omega}_k$, $\bar{\Omega}_k'$ and $\bar{\Omega}_k''$ are joined to crossovers $C_k(a_k - 1)$, $C_k'(b_k - 1)$ and $C_k''(c_k - 1)$, respectively, by type α_1 configurations in the upper triangles of those crossovers. Thus the level a_k component is connected to the level $a_k - 1$ component, the level b_k component is connected to the level $b_k - 1$ component, and the level c_k component is connected to the level $c_k - 1$ component. Since each integer i, $1 \leqq i \leqq 3n$, belongs to some $F \in \mathcal{F}'$, it follows by induction that all level components are connected to the backbone. The connected component formed by the backbone plus all the level components forms a single *skeleton* component. The only components remaining to be accounted for are those running between successive levels, such as the long row component that runs from crossover $C_k(i)$

to $C_k(i+1)$, and those that run from the squares Q_k and level 0. However, it is easy to see that (ii) and (iii) guarantee that each of these is connected to the skeleton in exactly one place, the connection being at the "bottom" of the component if it is in one of the columns associated with a Q_k for which $F_k \in \mathcal{F}'$ and through the "top" otherwise. The reader may verify that since $\mathcal{F}'$ is an *exact* cover, the graph we have constructed contains no cycles, and hence is a Steiner tree for X.

Thus we have shown that (7) holds if $\mathcal{F}$ contains an exact cover. We shall now complete the proof by showing that if (8) fails to hold, then $\mathcal{F}$ must contain an exact cover. If (8) does not hold, we must have

$$L^+ < 3nL(\alpha_1) + (\hat{C} - 3n)L(\beta_1) + nL(\gamma_1) + \varepsilon. \tag{9}$$

For each configuration w shown in Fig. 22, let $N(w)$ denote the number of active regions that contain a type w configuration in S^*. We then have

$$L^+ = \sum_w N(w) \cdot L(w). \tag{10}$$

Recall that if we consider only that part of S^* made up of edges with length $1/10$ or less, there are exactly $N = 2\hat{C} + 3n + 1$ connected components. These N components are joined together in S^* by the configurations from Fig. 22 that occur in certain active regions. By observing the number of components that are joined together by each type of configuration, we see that we must have

$$\begin{aligned} N - 1 = {} & 2N(\alpha_1) + 2N(\beta_1) + 3N(\gamma_1) + N(\alpha_2) \\ & + N(\beta_2) + 2N(\gamma_2) + 2N(\gamma_3) + N(\gamma_4). \end{aligned} \tag{11}$$

The integer multiplier in (11) for each configuration type is the reduction in total number of components obtained by using one instance of that configuration, i.e., one less than the number of components joined together by that configuration. From Table 1, we can obtain the following inequalities on the average length per single component reduction using each configuration type:

$$\tfrac{1}{2}L(\alpha_1) < \tfrac{1}{2}L(\beta_1) < \tfrac{1}{3}L(\gamma_1) < \tfrac{1}{2}L(\gamma_2) < L(\alpha_2) < L(\beta_2) = L(\gamma_4) = \tfrac{1}{2}L(\gamma_3). \tag{12}$$

Now, although there are $\hat{C}$ crossovers in X and hence $2\hat{C}$ triangles, there can be at most $\hat{C}$ configurations of types α_1 and β_1. This is because no crossover can contain two such configurations, since that would cause a cycle to occur in S^*. Thus

$$N(\alpha_1) + N(\beta_1) \leqq \hat{C}. \tag{13'}$$

We claim that for (9) to hold we must have equality in (13').

Suppose $N(\alpha_1) + N(\beta_1) \leqq \hat{C} - 1$. Then by (10), (11), (12) and the fact that $N = 2\hat{C} + 3n + 1$, a lower bound on L^+ would be

$$L^+ \geqq (\hat{C} - 1)L(\alpha_1) + \left(\frac{3n+2}{3}\right)L(\gamma_1).$$

However, as noted in (6), the choice of ε insures that

$$\hat{C}(L(\beta_1) - L(\alpha_1)) < \tfrac{2}{3}L(\gamma_1) - L(\alpha_1)$$

which in turn implies

$$L^+ \geqq (\hat{C}-1)L(\alpha_1)+\left(\frac{3n+2}{3}\right)L(\gamma_1) > \hat{C}L(\beta_1)+nL(\gamma_1)$$

$$= 3nL(\alpha_1)+(\hat{C}-3n)L(\beta_1)+nL(\gamma_1)+3n\varepsilon\sqrt{3}$$

which clearly contradicts (9). Thus, if (9) holds

$$N(\alpha_1)+N(\beta_1)=\hat{C}. \tag{13}$$

Because of (13), every crossover in S^* contains exactly one configuration of type α_1 or type β_1. Thus, all components belonging to the same level are joined together, for each of the $3n+1$ levels. Next, note that we can have at most one type α_1 configuration at each level, since the presence of two such configurations in the same level would form a cycle with points in the level above it. (Type α_1 configurations only occur in small triangles, which are always the top triangles in their respective crossovers, and link directly to downward terminators in the next higher level. Two such links between a pair of adjacent levels creates a cycle.) Thus $N(\alpha_1) \leqq 3n$, since there are no crossovers in level $3n$ of our construction. As before, we can argue that equality must hold. For suppose $N(\alpha_1) \leqq 3n-1$. Then by (10), (11), (12), and (13) we would have

$$L^+ \geqq (3n-1)L(\alpha_1)+(\hat{C}-3n+1)L(\beta_1)+nL(\gamma_1)$$

$$= 3nL(\alpha_1)+(\hat{C}-3n)L(\beta_1)+nL(\gamma_1)+\varepsilon\sqrt{3}$$

another violation of (9). Thus if (9) holds, then

$$N(\alpha_1)=3n. \tag{14}$$

From this, (9), (12), and the fact that $\frac{1}{2}L(\gamma_2) > \frac{1}{3}L(\gamma_3)+\varepsilon$ it follows that

$$\begin{gathered} N(\beta_1)=\hat{C}-3n; \qquad N(\gamma_1)=n; \\ N(\alpha_2)=N(\beta_2)=N(\gamma_2)=N(\gamma_3)=N(\gamma_4)=0. \end{gathered} \tag{15}$$

Notice that if $C_k(i)$, $i>0$, is a crossover that has its Steiner point in its *upper* triangle, then $C_k(i-1)$ must also have a Steiner point in its upper triangle, for otherwise the vertical long row component between them would not be joined to the rest of the tree. By induction, this forces the base square Q_k to have *some* interconnected active points and hence a type γ_1 configuration, since by (15) no other configuration types can occur in a square Q.

However, for each k, $1<k\leqq 3n$, the $(k-1)$st level contains a type α_1 connection, that is, a (warped) crossover with a Steiner point in its upper (small) triangle, by (14). By the preceding remarks, this forces the square Q_{i_j} below the crossover to contain a type γ_1 configuration. Moreover, by our construction of X, k must be an element of the 3-set F_{i_j} corresponding to Q_{i_j}. Hence, the only way that just n squares Q_{i_j} can have interconnected active points is if the corresponding sets F_{i_j} *cover* $\{1, 2, \cdots, 3n\}$. This in turn implies that those sets must be *disjoint*. Therefore (9) implies that an exact cover exists and the theorem is proved. □

7. Conversion to the discrete problem. In this section we return to the DESMT problem, the subject of our *NP*-completeness result. The basic construction in § 4 did not restrict X to points with integer coordinates. Indeed, a detailed specification of that construction would even contain points with irrational coordinates. However, Theorem 1 gives us some slack, by providing a gap of ε between the lengths of an ESMT for $X = X(\mathcal{F})$ when an exact cover for $\mathcal{F}$ does and does not exist. By an appropriate scaling and rounding process, we shall convert that construction to the required discrete form in such a way that the length discrepancies introduced by this conversion are insufficient to make this gap disappear.

The conversion proceeds in two steps. In the first we perform the scaling. Let $M = |X(\mathcal{F})|$ and set

$$X' = \left\{ \left(\frac{12M}{\varepsilon} x_1, \frac{12M}{\varepsilon} x_2 \right) : x = (x_1, x_2) \in X(\mathcal{F}) \right\}.$$

Now the minimum distance between points of X' is at least $(12M/11\varepsilon) \gg 200$, and Theorem 1 has been scaled up as follows (where L_1 is defined to be $3nL(\alpha_1) + (\hat{C} - 3n)L(\beta_1) + nL(\gamma_1) + L_0$):

THEOREM 2. *Let S' be a Steiner minimal tree for X'. Then, if $\mathcal{F}$ has an exact cover,*

$$l(S') \leqq \frac{12M}{\varepsilon} L_1, \tag{16}$$

and otherwise,

$$l(S') \geqq \frac{12M}{\varepsilon} L_1 + 12M. \tag{17}$$

The second step of the conversion is to round up the coordinates of X', using the rounding function $f: \mathbb{R} \times \mathbb{R} \to \mathbb{Z} \times \mathbb{Z}$ defined so that, if $x = (x_1, x_2)$, $f(x) = (\lceil x_1 \rceil, \lceil x_2 \rceil)$. We specify the input to the DESMT problem which corresponds to $\mathcal{F}$ to be the set

$$X'' = \{ f(x) : x \in X' \}$$

and the length $L = \lceil (12M/\varepsilon) L_1 + 6M \rceil$. Note that because of the substantial distance between points in X', this rounding will not cause any distinct points of X' to coalesce, so that $|X''| = |X'| = |X| = M$.

As we shall show, the analogue of Theorems 1 and 2 now becomes the following:

THEOREM 3. *Let S^D be a discrete Euclidean Steiner minimal tree for X''. Then if $\mathcal{F}$ has an exact cover*

$$l'(S^D) < L, \tag{18}$$

and otherwise,

$$l'(S^D) > L + 2M. \tag{19}$$

Proof. Suppose first that $\mathcal{F}$ has an exact cover and let S' be an ESMT for X', with Y' being its set of vertices. By Theorem 2, $l(S') \leqq (12M/\varepsilon)L_1$. By Fact 2, we may assume that $|Y'| \leqq 2M-2$ and hence S' has at most $2M-3$ edges.

Now let $Y'' = \{f(x) : x \in Y'\}$ and consider the tree T'' with vertex set Y'' and an edge between two points if and only if their inverse images under f are joined by an edge in S'. For each edge $\{x, y\}$ of S', we have

$$d'(f(x), f(y)) \leqq \lceil d(x, y) + \sqrt{2} \rceil < d(x, y) + 3.$$

Hence $l'(T'') \leqq l(S') + (2M-3) \cdot 3 < (12M/\varepsilon)L_1 + 6M \leqq L$ and (18) is proved.

Before proving (19), we first observe that Fact 2 continues to hold in the discrete case. All that is required for Fact 2 to hold is that the distance metric obey the triangle inequality as d' does. For, if this is the case, there is no value in having a Steiner point with degree two or less in the Steiner minimal tree, and a simple counting argument then suffices to show that there need be at most $|X|-2$ Steiner points in a Steiner minimal tree for X.

Now, suppose $\mathcal{F}$ does not have an exact cover, but that S^D is a DESMT for X'' with $l'(S^D) \leqq L + 2M$. Let Y^D be the vertex set for S^D. By the previous remark, we may assume that $|Y^D| \leqq 2M-2$ and hence S^D has at most $2M-3$ edges. Define $g: \mathbb{Z} \times \mathbb{Z} \to \mathbb{R} \times \mathbb{R}$ by

$$g(x) = \begin{cases} y', & \text{if } y' \in X' \text{ and } f(y') = x, \\ x, & \text{otherwise.} \end{cases}$$

Because f is a 1–1 map of X' onto X'', g is a 1–1 map of Y^D onto $(Y^D - X'') \cup X' = Y'$. Consider the tree T' with vertex set Y' and an edge between two vertices if and only if there is an edge in S^D between their inverse images under g. For each edge $\{x, y\}$ of S^D we then have

$$d(g(x), g(y)) \leqq d'(x, y) + \sqrt{2} < d'(x, y) + 2.$$

Hence $l(T') \leqq L + 2M + 2(2M-3) < (12M/\varepsilon)L_1 + 12M$. But since $X' \subseteq Y'$, this implies by definition of an ESMT, that there exists an ESMT S' for X' such that $l(S') < (12M/\varepsilon) + 12M$, a contradiction to Theorem 2. Hence (19) holds and Theorem 3 is proved. □

THEOREM 4. *The DESMT problem is NP-complete.*

Proof. We first use Theorem 3 to show that the DESMT problem is *NP*-hard, that is, that a known *NP*-complete problem is polynomially reducible to it. The known *NP*-complete problem is of course, the X3C problem. Let $\mathcal{F}$ be an arbitrary input to X3C. As specified earlier, the corresponding input for the DESMT problem is the point set X'', as constructed via the intermediate set X of § 4, and the integer $L = \lceil (12M/\varepsilon)L_1 + 6M \rceil$. By Theorem 3, $\mathcal{F}$ has an exact cover if and only if X'' has a DESMT of length L or less. Thus the reduction of X3C to DESMT works as required. We now indicate why the reduction is a *polynomial* reduction, that is, why it can be accomplished in time bounded by a polynomial in the number of bits needed to specify $\mathcal{F}$.

Our first claim, which can be verified by a straightforward examination of the construction in § 4 is that $M = |X| = |X''|$ can be bounded by a polynomial in n and t, and hence in the number of bits specifying $\mathcal{F}$. Secondly, since X has a spanning tree with maximum edge length 1, X'' has a spanning tree with maximum edge length at most $(12M/\varepsilon)+3$. Hence if we let the left- and bottom-most point of X'' have coordinates (0, 0), all points have nonnegative integer coordinates bounded by $M((12M/\varepsilon)+3) = 2400ntM^2+3M$, still a polynomial in n and t. Thus X'' can be specified as a listing of its points using a number N of bits bounded by a polynomial in the number of bits specifying $\mathcal{F}$. Our next observation is that, although the details are not spelled out in § 4, it is possible to follow the schematic given there to construct X in time proportional to N, using symbolic square roots to describe coordinates, but in such a way that no point of X has more than one square root in any coordinate. Given a representation of X in this form, it then is possible to perform the scaling and rounding to obtain X'', again in time polynomially-bounded in N. Thus the whole process of constructing X'' from $\mathcal{F}$ can be accomplished in time bounded by a polynomial in the number of bits specifying $\mathcal{F}$ and our reduction *is* a polynomial reduction.

To complete the proof of *NP*-completeness, we must show that the DESMT problem is in *NP*, that is, it can be solved in polynomial time by a nondeterministic Turing machine. To do this it will be sufficient to show that given a set X of points in the plane, a DESMT for X can be described with a number of bits bounded by a polynomial in the number of bits specifying X. A nondeterministic algorithm for the DESMT problem could thus "guess" the DESMT, compute its length, and compare the result to L, all in polynomial time. So let X be a set of integer coordinate points, and set $M = |X|$, $m = \max\{|\beta|: \beta \text{ is a coordinate of a point in } X\}$. By the remark about Fact 2 in the proof of Theorem 3, we know that a DESMT for X need have at most $2M-2$ vertices and $2M-3$ edges. Also, there is no need for a vertex having coordinates outside the range $[-m, m]$. Hence a DESMT can be described with $\mathcal{O}(M \log m)$ bits. This is clearly bounded by a polynomial in the number of bits specifying X, which must be at least $M + \log_2 m$. □

Although we have been unable to prove that the ESMT problem itself is in *NP*, due to the obstacles posed by irrational numbers, we can prove that it is at least as hard as any *NP*-complete problem. This result will hold even when we allow the coordinates of the Steiner points to be output as symbolic expressions, so long as those expressions are amenable to approximation. By this we mean that an expression using m bits can be evaluated to within a range of $\pm\Delta$ in time bounded by a polynomial in m and $\log(\Delta)$. (Otherwise, the symbolic expression would in effect be computationally useless). For example, symbolic expressions formed from integers using the operations $+, -, \cdot, /$, and $\sqrt{\ }$ meet these requirements, and indeed are all that is needed to completely specify the coordinates of an ESMT. (In fact, all the coordinates are of the form $(n + n_2\sqrt{3})/n_3$ for appropriate integers n_i (cf. [2a])).

THEOREM 5. *Suppose A is a polynomial time algorithm which, when given a description of a finite set X of points in the plane, produces a representation of an ESMT for X, with the coordinates of the Steiner points given by symbolic expressions which are "amenable to approximation". Then the* X3C *problem, and hence all other NP-complete problems, can be solved in polynomial time.*

Proof. Suppose such an algorithm A exists. We show how to use A to solve X3C in polynomial time. Given an input $\mathcal{F}$ for X3C, construct X'' as before. Apply A to X'', yielding a symbolic ESMT S^* for X''. Evaluate $l(S^*)$ to an accuracy that guarantees a value within $\pm M$ of the actual value. Examining the proof of Theorem 3, we then note that $l(S^*) < L$ if $\mathcal{F}$ has an exact cover and $l(S^*) > L + 2M$ otherwise. Thus the approximation L^* to the length of S^* will satisfy $L^* \leqq L + M$ if and only if $\mathcal{F}$ has an exact cover, and this simple comparison is the last step involved in solving the X3C problem using A. Since each stage of this procedure operates in time bounded by a polynomial in the time for the preceding step, the whole process constitutes a polynomial time algorithm for X3C. □

Thus the ESMT problem is at least as difficult as the DESMT problem, even if irrational square roots are allowed to be represented symbolically. Moreover, our proofs only capture a part of the complexity of these Steiner minimal tree problems. For observe that in our construction the number of candidates for Steiner points was always much smaller than $|X|-2$, the maximum possible number which might have been required. We in fact showed that both problems are *NP*-hard, even if the number of potential Steiner points is severely limited by the structure of X. In general one must cope with the possibility that as many as $|X|-2$ Steiner points may be needed, and the number of candidates for those Steiner points could be an exponential function of $|X|$.

8. Some final comments. The results of this paper offer little hope to those who wish to construct general ESMT's or DESMT's. A related problem, arising in connection with wire routing on printed circuit boards, is the rectilinear Steiner minimal tree (RSMT) problem [5], [9], [11]. Here the Euclidean metric is replaced by the Manhattan (or rectilinear) metric, where the distance $d_1(x, y)$ between $x = (x_1, x_2)$ and $y = (y_1, y_2)$ is given by $d_1(x, y) = |x_1 - y_1| + |x_2 - y_2|$.

One might hope that the RSMT problem would be easier than either the ESMT or DESMT problems, since it has the property that one may always construct a minimum length Steiner tree by choosing a subset of the grid segments of the grid formed by drawing horizontal and vertical lines through each required point [9]. Unfortunately this is not the case, since the techniques of this paper can be used to show that the RSMT problem is also *NP*-complete. An alternative proof of this fact appears in [5].

An interesting open question, arising in connection with our attempts to prove that the ESMT problem is in *NP*, is whether the following number theoretic problem belongs to *NP*:

$$\text{Given positive integers } x_1, x_2, \cdots, x_n, \text{ and } L, \text{ is } \sum_{i=1}^{n} \sqrt{x_i} \leqq L?$$

In fact, without a result such as this, we are unable even to show that the Euclidean *minimal spanning tree* problem belongs to *NP*, an annoying fact in view of our ability to construct the minimal spanning tree explicitly in low-order polynomial time! One approach to showing that this problem belongs to *NP* would be to show that there exists a polynomial $p(n)$ such that either

$$L = \sum_{i=1}^{n} \sqrt{x_i} \quad \text{or} \quad \left| L - \sum_{i=1}^{n} \sqrt{x_i} \right| > 2^{-p(n)}$$

whenever each of the integers x_i can be expressed with at most n bits. At present, however, the best result of this type known [15] has $p(n)$ replaced by an exponential function of n.

REFERENCES

[1] A. V. AHO, M. R. GAREY AND F. K. HWANG, *Rectilinear Steiner trees: Efficient special case algorithms*, Networks, to appear.

[2] A. V. AHO, J. E. HOPCROFT AND J. D. ULLMAN, *The Design and Analysis of Computer Algorithms*, Addison-Wesley, Reading, MA., 1974, Chap. 10.

[2a] W. M. BOYCE, *An improved program for the full Steiner tree problem*, tech. memo., Bell Laboratories, Murray Hill, NJ, 1975.

[3] W. M. BOYCE AND J. B. SEERY, *STEINER 72, An improved version of Cockayne and Schiller's program STEINER for the minimal network problem*, tech. memo., Bell Laboratories, Murray Hill, NJ, 1973.

[4] E. J. COCKAYNE AND D. G. SCHILLER, *Computation of Steiner minimal trees*, Combinatorics, D. J. A. Welsh and D. R. Woodall, eds., Institute for Mathematics and Applications, Southend-on-Sea, Essex, England, 1972, pp. 53–71.

[5] M. R. GAREY AND D. S. JOHNSON, *The rectilinear Steiner tree problem is NP-complete*, this Journal, 32 (1977), pp. 826–834.

[6] E. N. GILBERT, *Minimum cost communication networks*, Bell System Tech. J., 9 (1967), pp. 2209–2227.

[7] E. N. GILBERT AND H. O. POLLAK, *Steiner minimal trees*, this Journal, 16 (1968), pp. 1–29.

[8] R. L. GRAHAM, *Some results on Steiner minimal trees*, tech. memo., Bell Laboratories, Murray Hill, NJ, 1967.

[9] M. HANAN, *On Steiner's problem with rectilinear distance*, this Journal, 14 (1966), pp. 255–265.

[10] M. HANAN AND J. M. KURTZBERG, *Placement techniques*, Design Automation of Digital Systems, vol. 1, M. A. Breuer, ed., Prentice-Hall, Englewood Cliffs, NJ, 1972, pp. 213–282.

[11] F. K. HWANG, *On Steiner minimal trees with rectilinear distance*, this Journal, 30 (1976), pp. 104–114.

[12] R. M. KARP, *Reducibility among combinatorial problems*, Complexity of Computer Computations, R. E. Miller and J. W. Thatcher, eds., Plenum Press, New York, 1972, pp. 85–104.

[13] ———, *On the computational complexity of combinatorial problems*, Networks, 5 (1975), pp. 45–68.

[14] Z. A. MELZAK, *On the problem of Steiner*, Canad. Math. Bull., 4 (1961), pp. 143–148.

[15] A. M. ODLYZKO, personal communication.

[16] M. I. SHAMOS AND D. HOEY, *Closest point problems*, 16th Annual Symposium on Foundations of Computer Science, IEEE, New York, 1975, pp. 151–162.

[17] J. SOUKUP, *On minimum cost networks with nonlinear costs*, this Journal, 29 (1975), pp. 571–581.