SOME NP-COMPLETE GEOMETRIC PROBLEMS

by

M. R. Garey, R. L. Graham and D. S. Johnson
Bell Laboratories
Murray Hill, New Jersey

ABSTRACT

We show that the STEINER TREE problem and TRAVELING SALESMAN problem for points in the plane are NP-complete when distances are measured either by the rectilinear (Manhattan) metric or by a natural discretized version of the Euclidean metric. Our proofs also indicate that the problems are NP-hard if the distance measure is the (unmodified) Euclidean metric. However, for reasons we discuss, there is some question as to whether these problems, or even the well-solved MINIMUM SPANNING TREE problem, are in NP when the distance measure is the Euclidean metric.

INTRODUCTION

Geometric optimization problems are both practically and theoretically intriguing. They are practically intriguing because, for instance, Euclidean space is the domain of the everyday world, the space in which problems actually arise and in which the solutions are to be applied. They are theoretically intriguing because, despite the attention paid to geometric problems since ancient times, little is known about their computational complexity.

It is only recently that results in what might be called "computational geometry" have begun to appear. Much of this work is due to M. I. Shamos [18,19], who has developed efficient algorithms for solving a great variety of geometric construction problems, and has pointed out the rich class of geometric problems that still remain open.

Some of these open problems can be thought of as special cases of well-studied graph problems. Whereas the general problems deal with abstract points joined by edges having arbitrarily specified lengths, the corresponding geometric problems deal with points in the plane or in 3-space, with the edge lengths being the actual interpoint distances under one of the standard metrics.

Three such problems are the Minimum Spanning Tree problem, the Steiner Tree problem, and the Traveling Salesman problem. Shamos and Hoey have shown [19] that a minimum spanning tree for n points in the plane, under the usual Euclidean metric, can be found using O(n log n) comparisons, whereas the best algorithm known for finding a minimum spanning tree in an n-vertex graph requires on the order of n^2 comparisons. This might seem to offer some hope that although the other two problems are NP-complete for arbitrary graphs [11,12], we might be able to find polynomial-time algorithms for the corresponding geometric problems dealing with points in the plane.

The two metrics under which such results would be most valuable are probably the L_1 (rectilinear or "Manhattan") metric and the L_2 (Euclidean) metric. For two points $x = (x_1,x_2)$ and $y = (y_1,y_2)$ in the plane, the L_1 distance $d_1(x,y)$ between them is

$$d_1(x,y) = |x_1-y_1| + |x_2-y_2|,$$

and the L_2 distance $d_2(x,y)$ between them is

$$d_2(x,y) = ((x_1-y_1)^2+(x_2-y_2)^2)^{1/2}.$$

The L_2 distance is, of course, the length of the straight line segment joining x to y, whereas the L_1 distance is the length of the shortest "path" joining x to y, which is composed solely of horizontal and vertical line segments. The L_1 distance is frequently of interest for circuit layout problems where conductor paths are made up of only horizontal and vertical line segments.

The main results of this paper say that the Steiner Tree and Traveling Salesman problems, for points in the plane under these two metrics, are both at least as hard as the corresponding problems for arbitrary graphs and distances. However, before we can provide a more precise statement of our results, we must first examine some technical difficulties involved with the L_2 metric.

Note that, even when we impose the standard restriction that only points with integer coordinates be allowed in the inputs, we can still have irrational interpoint distances under the L_2 metric. This in itself may not pose difficulties, since in the course of a computation it may be possible to deal with such distances merely as symbolic square roots, as is in fact done in the algorithm for finding minimum spanning trees under the L_2 metric. However, consider the Minimum Spanning Tree problem for points in the plane, phrased as a language recognition problem, i.e., "Does there exist a spanning tree with length L or less?" Generally one would expect such a recognition problem to be no harder than the corresponding optimization problem. However, it is not at all apparent that this recognition problem is even in NP, although we can find a minimum spanning tree in low order polynomial time. The symbolic expression for the length of a given spanning tree on n points may involve as many as n-1 square roots. An attempt to compare this to an integer L by repeated squaring to eliminate all the square roots can take exponential time. There is more hope for the alternate approach of evaluating all the square roots with sufficient accuracy that their sum can be compared to L. However, the best upper bound we can currently give on the number of places of accuracy required for the comparison is $O(m2^n)$, where m is the number of digits in the original symbolic expression [15]. To reach this amount of accuracy will clearly also take exponential time.

Since NP-completeness results must deal with language recognition problems, we encounter these same difficulties while treating the Steiner Tree and Traveling Salesman problems under the L_2 metric. However, it is not our intent to prove that these problems are hard merely because of the computational drawbacks resulting from the presence of irrational square roots. We shall avoid these drawbacks by replacing the L_2 metric by one that approximates it and reflects the manner in which distances must be computed in practice, i.e., by rounding. To be precise, we use the metric L_2', given by

$$d_2'(x,y) = \lceil d_2(x,y) \rceil,$$

(where $\lceil \alpha \rceil$ is the smallest integer not less than α). Arbitrary accuracy can still be obtained by appropriate scaling, so that using this modified metric does not change the practical problems in any essential way. Moreover, our NP-completeness proofs using this metric can be converted to NP-hardness proofs using L_2, thus eliminating any vagrant suspicion that it is perhaps the rounding involved in L_2' which makes the problems difficult.

Having disposed of the technical issues posed by the metrics, let us now turn to the problem with which our results are concerned. The Steiner Tree problem, stated as an optimization problem, is basically the following: "Given a set S of points in the plane, find a set $S' \supseteq S$ such that the minimum spanning tree for S' is as short as possible". A minimum spanning tree for such an S' is called a minimum Steiner tree for S, and the points in S'-S are called Steiner points. This problem has been studied extensively in recent years, both for the L_1 [1,9] and L_2 [3,4,7,13] metrics, although no general polynomial time algorithm has been found in either case.

The language recognition versions of this problem under the two metrics L_1 and L_2' can be combined as follows:

STEINER TREE PROBLEM:

> Given a set S of integer-coordinate points in the plane and an integer $L > 0$, does there exist a set $S' \supseteq S$ of integer-coordinate points such that the minimum spanning tree for S', with edge lengths measured by L_1 (L_2'), has total length at most L?

Note that we have not only restricted the points given in S to having integer coordinates, but also have put a similar restriction on the Steiner points. This is consonant with the practical necessities of rounding. Moreover, in the L_1 case it is actually no restriction at all, since a theorem of Hanan [9] tells us that there must exist a minimum Steiner tree, each of whose Steiner points has coordinate values chosen from those occuring in points of S. In the L_2' case, allowing Steiner points with non-integer coordinates can yield slightly shorter trees, but again the potential discrepancy can be made arbitrarily small by appropriate scaling.

Moreover, it is now easy to show that both of these problems belong to NP, which is one half of a proof of NP-completeness [2,11,12]. The key fact is that both metrics obey the triangle inequality, so that no Steiner point of degree 2 or less is necessary. From this one can conclude, using well known and straightforward arguments [7], that there need be no more than $|S| - 2$ Steiner points.

The second problem we consider is the Traveling Salesman problem: "Given a set S of points in the plane, find the shortest circuit that passes through all the points of S". This is a well-known and much-studied [10,17] problem, for which no polynomial time algorithm is known. The language recognition versions of this problem under our two metrics can be combined as follows (and are clearly in NP):

TRAVELING SALESMAN PROBLEM: Given a set S of integer coordinate points in the plane and an integer L, does there exist a circuit passing through all the points of S which, with edge lengths measured by L_1 (L_2'), has total length at most L?

Our main results are that the four problem versions described above are not only in NP, but are also NP-complete. To prove this, we must show that known NP-complete problems can be polynomially transformed into each of them. The known NP-complete problem we use in all four cases is the following:

> EXACT COVER BY 3-SETS (X3C): Given a family $\mathcal{F} = \{F_1, F_2, \ldots, F_t\}$ of 3-element subsets of a set U of 3n elements (without loss of generality taken to be $U = \{1,2,3,.,\ldots,3n\}$), does there exist a subfamily $\mathcal{F}' \subseteq \mathcal{F}$ of pairwise disjoint sets such that $\bigcup_{F \in \mathcal{F}'} F = U$?

This problem is known to be NP-complete as it contains the 3-DIMENSIONAL MATCHING problem of [11] as a subcase.

All our transformations involve the same basic scheme of construction. In Section 2 we present a fairly detailed view of this scheme and how it works, while proving NP-completeness for L_1-STEINER TREE. The construction of this section then serves as a model for the other proofs, which are given in less detail. In Section 3 an NP-completeness proof for L_2'-STEINER TREE is sketched, and Section 4 is devoted to the two TRAVELING SALESMAN results.

These four NP completeness results are the first we know to have been proved about geometric problems. An alternate NP-completeness proof for the L_2-TRAVELING SALESMAN problem (with distances rounded in a slightly different way) has, however, been obtained independently by Papidimitriou [16]. An alternate proof for L_1-STEINER TREE using a series of NODE COVER problems as intermediaries, will be presented by two of the current authors in [6].

2. THE L_1-STEINER TREE PROBLEM IS NP-COMPLETE

Let $\mathfrak{F} = \{F_1, F_2, \ldots, F_t\}$, $\bigcup_{i=1}^{t} F_i = \{1,2,\ldots,3n\}$, be an input to the X3C problem. We shall construct a set of points S and a bound L such that a minimum Steiner tree for S under the L_1 metric has length L or less if and only if $\mathfrak{F}$ has an exact cover. The construction will be clearly polynomially bounded, so this will prove the L_1-STEINER TREE problem to be NP-complete.

We build S in stages, starting with two basic units. Figure 1 shows a junction and a symbolic abbreviation for it. We follow the convention that a line segment stands for the set of all integer coordinate points it contains. The value of K is given by

$$K = 162n^2t^2 + 96n^2t + 10nt \qquad (2.1)$$

The area enclosed by the dotted line in Fig. 1 will be called the active region for the junction. It consists of all points within L_1-distance K of the central point (0,0) in the junction.

The second basic unit is the crossover, of which there are two forms: **standard** and warped. Figure 2 presents both forms and their abbreviations. They differ as to the value of α, and the coordinate of the topmost point. Each has two active regions. The upper active region consists of all points within distance K of (0,2K); the lower active region consists of all points within distance K of (0,-2K).

S is built up from these basic components as follows. A crossover stack of height k is a vertical sequence of k crossovers, each crossover having its top point coincide with the bottom point of the one above it. The topmost crossover is a warped one and all others are standard ones. A crossover stack of height 3 is illustrated symbolically in Figure 3.

Each set $F_i = \{a_i, b_i, c_i\} \in \mathfrak{F}$ will be represented by a set structure consisting of one junction and three crossover stacks, of heights a_i, b_i, and c_i respectively. These are joined by making the three top points of the junction coincide with the bottom points of the crossover stacks, as in Figure 4.

The set representations are then put together to form the set S as follows. The backbone of the construction is pictured in Figure 5, with t+1 prongs in sequence as shown. The representation for each set F_i, $1 \leq i \leq t$, is placed so that its bottom point coincides with the top point of prong i of the backbone. We complete the construction by adding additional points as follows.

All crossovers which are bottom crossovers in their stack will be called level 0 crossovers. In general, if a crossover is above j crossovers in its stack, it will be called a level j crossover, $0 \leq j \leq 3n-1$. Let y_j be the y-coordinate of the leftmost points of the crossovers at level j and let y_{3n} be the coordinate of the top point in all level 3n-1 crossovers (that is, $y_j = 10K+8jK$). Observe that prong 0 of the backbone has y-coordinate y_0, and that in a warped crossover at level j the top point has y-coordinate y_{j+1}. To complete our construction of S, add all integer-coordinate points whose y-coordinate is y_j for some j, $0 \leq j \leq 3n$, whose x-coordinate is the same as that for some point in the backbone ($0 \leq x \leq 30K(t+1)$), and which is not "inside" any crossover. Figure 6 represents the final construction for $\mathfrak{F} = \{\{1,2,3,\},\{2,4,5\},\{1,2,6\}\}$.

We now construct the bound L on the size of the desired minimum L_1-Steiner tree. Let T_0 be the set of all edges between pairs of points in S whose L_1-distance is 1 (our representation of sequences of integer-coordinate points by straight lines corresponds to drawing in all the T_0 edges between them). Let q be the number of crossovers in S. (Note that $q < 3nt$.) Then

$$L = |T_0| + 54qnt + 96n^2t - 9n \qquad (2.2)$$

We claim that a minimum L_1-Steiner tree for S has length L or less if and only if $\mathfrak{F}$ has an exact cover.

(Remark: S has been designed for ease of description, rather than for minimality of |S|.)

Let T be a minimum L_1-Steiner tree which has length L or less and contains a maximum number of edges from T_0. We shall see that T must be of a rather restricted form.

Claim 2.1 T contains all the edges T_0.

Proof. Suppose it did not. Let {u,v} be an edge of T_0 not in T. By definition of T_0, {u,v} has length 1. Adding {u,v} to T must create a cycle, as T is already a spanning tree. This cycle must contain at least one edge not in T_0, as by construction T_0 contains no cycles. This edge must have L_1-length at least 1 since it is between integer-coordinate points. Thus deleting this edge and adding {u,v} gives us a new spanning tree of no greater length than T which contains one more edge of T_0, a contradiction. ■

Thus T is made up of T_0 plus some additional edges of total length less than K. Ignoring these additional edges for a moment, we can see that the graph made up of just the edges from T_0 is made up of 3n+2q+1 connected components. A T_0-component which contains a point with y-coordinate y_j will be called a level j component. There are 3n+1+q level components all told. The remaining q components are those that run vertically between levels, each made up of the top part of one crossover or junction joined to the bottom part of the crossover above it. In Figure 6, the T_0-components

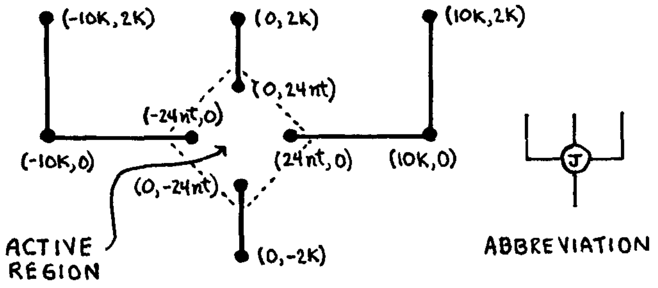


FIGURE 1. L_1-JUNCTION

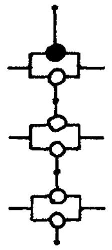


FIGURE 3. CROSSOVER STACK OF HEIGHT 3

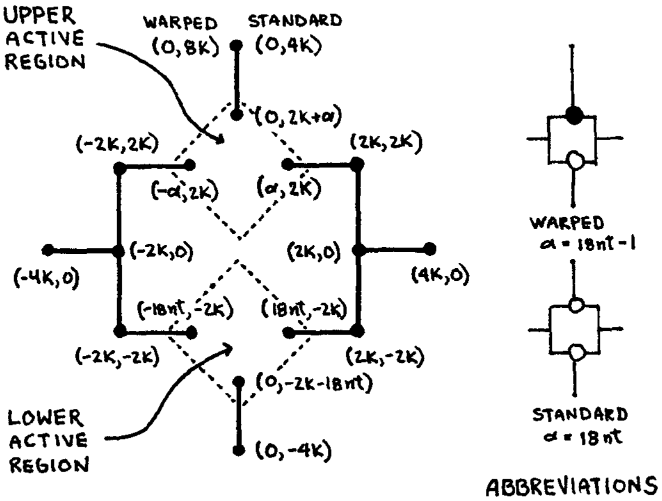


FIGURE 2. STANDARD AND WARPED L_1-CROSSOVERS

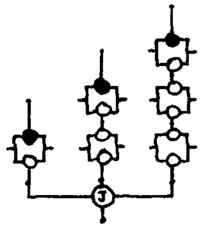


FIGURE 4. SET STRUCTURE FOR $F_i = \{1,2,3\}$

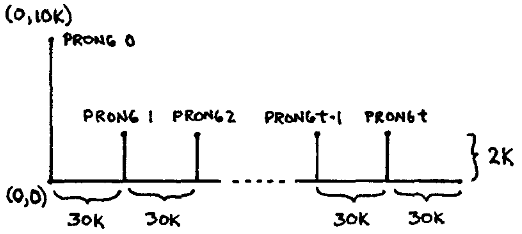


FIGURE 5. THE BACKBONE

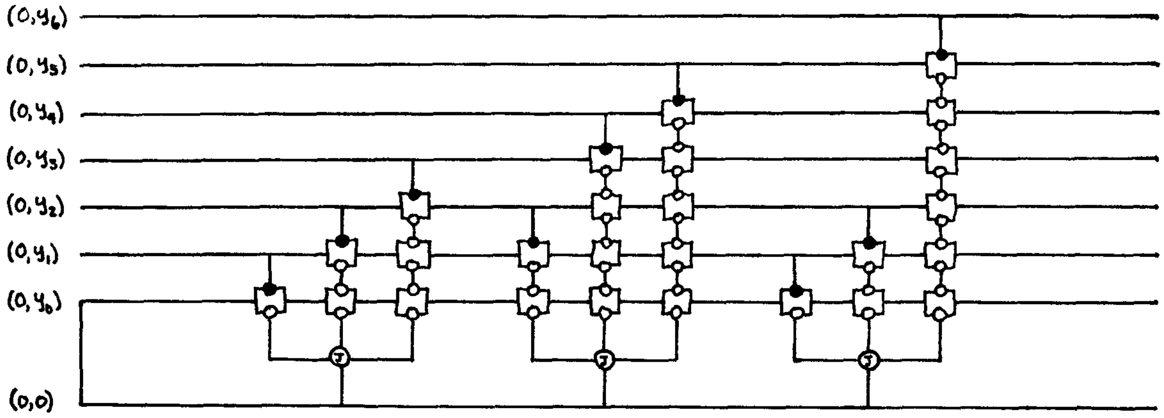


FIGURE 6. FINAL CONSTRUCTION FOR $\mathfrak{F} = \{\{1,2,3\},\{2,4,5\},\{1,2,6\}\}$

are separated from each other by the circles representing active regions. The additional edges in T must serve to link up these T_0-components into one overall connected tree structure.

We first observe that, since we are operating in the L_1-metric, each of these additional edges can be drawn as a path made up of horizontal and vertical length 1 line segments, whose total L_1-length is the same as the length of the edge. Let us assume that all additional edges are so drawn. The segments making up the paths will be called supplementary segments, and will form the set T_1. Observe that we must have

$$|T_1| \le 54qnt + 96n^2t - 9n < K \qquad (2.3)$$

Claim 2.2 All supplementary segments in T_1 are contained in active regions of crossovers and junctions of S.

Proof. By our overall construction of S and the specification of the active regions, a point not in an active region cannot lie on a path of length K or less between two different T_0-components. Yet, since T is a *minimum* L_1-Steiner tree, all points on a supplementary segment *must* lie on such a path. ■

We thus know that all connections between T_0-components occur in active regions. This greatly reduces the possibilities we need consider, as it is easy to determine a minimum length way of achieving any given connection of the T_0-components entering an active region. Figures 7, 8, and 9 show a minimum length connecting configuration for each of the possibilities, with symmetric cases combined, and the case of zero connections omitted. Note that the *number* of connections for a given configuration is the difference between the number of T_0-components entering the region, and the number of connected components present in the configuration.

We may assume without loss of generality that in T, all connections between T_0-components are made by one of the configurations listed in the figures. Since there are $3n + 2q + 1$ T_0-components, exactly $3n + 2q$ connections must be made. For each type x of configuration, let N(x) be the number of times that connecting configuration is used in T.

Claim 2.3 $N(\alpha_1) + N(\beta_1) = q$, and each crossover has exactly one of its active regions connected by a type α_1 or type β_1 configuration.

Proof. If any crossover had both its active regions so connected, it would contain a cycle, which is impossible since T is a tree. Thus $N(\alpha_1) + N(\beta_1) \le q$. If $N(\alpha_1) + N(\beta_1) \le q\text{-}1$, then at most 2q-2 connections of T_0-components are made at average cost 27nt or less. The remaining 3n + 2 connections must have average cost at least 32nt. Thus

$$|T_1| \ge (2q-2)(27nt-3/2) + (3n+2)(32nt)$$
$$= 54qnt + 96n^2t + 10nt - 3q + 3,$$

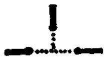
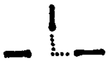
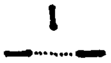
CONFIGURATION	TYPE	LENGTH OF SUPPLEMENTARY EDGES	NUMBER OF CONNECTIONS	LENGTH PER CONNECTION
	α_1	54nt	2	27nt
	α_2	36nt	1	36nt
	α_3	36nt	1	36nt

FIGURE 7. POSSIBLE CONNECTING CONFIGURATIONS IN UPPER AND LOWER ACTIVE REGIONS OF STANDARD CROSSOVERS, AND LOWER ACTIVE REGION OF WARPED CROSSOVERS

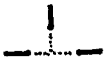
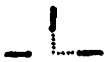
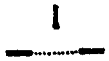
	β_1	54nt - 3	2	$27nt - \frac{3}{2}$
	β_2	36nt - 2	1	36nt - 2
	β_3	36nt - 2	1	36nt - 2

FIGURE 8. POSSIBLE CONNECTING CONFIGURATIONS IN UPPER ACTIVE REGION OF WARPED CROSSOVERS

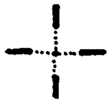
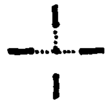
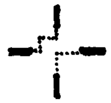
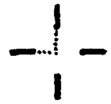
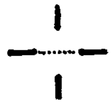
	γ_1	96nt	3	32nt
	γ_2	72nt	2	36nt
	γ_3	96nt	2	48nt
	γ_4	48nt	1	48nt
	γ_5	48nt	1	48nt

FIGURE 9. POSSIBLE CONNECTING CONFIGURATIONS IN JUNCTION ACTIVE REGIONS

which violates (2.3) since $q < 3nt$. ■

By Claim 2.3, the type α_1 and type β_1 connections insure that all level j components of T_0 are connected into a single level j supercomponent in T, $0 \leq j \leq 3n$. From this we can conclude the following.

Claim 2.4 $N(\beta_1) = 3n$, and exactly one warped crossover is connected by a type β_1 configuration at each level j, $0 \leq j \leq 3n-1$.

Proof. Suppose two warped crossovers at the same level j were connected by type β_1 configurations. These must occur in the top active regions of the crossovers, each of which by our construction contains a T_0-component of level j + 1. Thus T would contain two distinct paths from level j to level j + 1, and hence a cycle involving the two paths and parts of the level j and level j + 1 supercomponents. Since this is impossible, and since only warped crossovers can contain type β_1 configurations, $N(\beta_1) \leq 3n$. A lower bound argument similar to the one for Claim 2.3 can then be used to force equality. ■

Observe that a type β_1 configuration at level j will connect up the level j and level j + 1 supercomponents. Thus Claim 2.4 tells us that all the supercomponents are connected into one overall skeleton component, which includes the backbone as that is connected to the level 0 supercomponent at prong 0. Observe that the skeleton has been connected up without the use of any of the "between-level" T_0-components. Hence each of these can be directly connected to at most one of the two levels it lies between. Otherwise it would create a cycle.

Claim 2.5 $N(\gamma_1) = n$

This follows from another lower bound argument similar to the one used for Claim 2.3. We are now ready, using Claim 2.5, to show that the existence of T implies that $\mathfrak{F}$ has an exact cover.

Claim 2.6 $\mathfrak{F}' = \{F_i$: the junction active region for the set structure representing F_i contains a type γ_1 configuration in T$\}$

is an exact cover for $\mathfrak{F}$.

Proof. From Claim 2.5, we know that $|\mathfrak{F}'| = n$, as desired. All that remains to be shown is that the sets in $\mathfrak{F}'$ are all pairwise disjoint, in which case their union must be all 3n elements. Let us consider the set structure for an $F_i \in \mathfrak{F}'$. We show that the top crossovers in its three crossover stacks must all contain type β_1 configurations. Since there cannot be two β_1 configurations at the same level in T, $\mathfrak{F}'$ cannot then contain two F_i whose set structures have crossover stacks of the same height, and hence all the $F_i \in \mathfrak{F}'$ are pairwise disjoint.

So suppose the top crossover in a crossover stack of height j for F_i does not contain a type β_1 configuration in its upper active region. Then by Claim 2.3 it must contain a type α_1 configuration in its lower active region. Thus the between-level T_0-component just below the crossover is directly connected to level j. By our discussion after Claim 2.4, that between-level component cannot also be directly connected to the level below. Thus if $j > 0$, the crossover at level j-1 in the stack cannot have a type α_1 configuration in its upper active region, and so must have one in its lower active region. By induction, the level 0 crossover in the stack has a type α_1 configuration in its lower active region. But this means that the between-level T_0-component between that crossover and the junction active region for the set structure is joined directly to level 0, and cannot be joined directly to the backbone without creating a circuit. This contradicts the fact that the junction active region contains a type γ_1 configuration. ■

Now we complete the proof that the L_1-STEINER TREE problem is NP-complete by showing that if there is an exact cover for $\mathfrak{F}$, there is an L_1-Steiner tree for S of total length L or less. Suppose $I \subseteq \{1,2,\ldots,t\}$ satisfies $|I| = n$, and $\mathfrak{F}' = \{F_i : i \in I\}$ is an exact cover for $\mathfrak{F}$. We construct an L_1-Steiner tree T* as follows.

a) Include all edges of T_0;

b) For all $i \in I$, include a type γ_1 configuration in the junction active region of the set structure representing F_i;

c) For all $i \in I$, include a type β_1 configuration in the upper active regions of the warped crossovers topping the three crossover stacks of the F_i set structure, and a type α_1 configuration in the upper active region of each standard crossover in the F_i set structure;

d) For all $i \in \{1,2,\ldots,t\} - I$, include a type α_1 configuration in the lower active region of every crossover in the F_i set structure.

The total length of the edges in T* is clearly

$$|T_0| + n\cdot(96nt) + 3n\cdot(54nt-3) + (q-3n)(54nt) = L.$$

Moreover, T obeys Claims 2.3 and 2.4, so that all T_0-components except the between-level components must be connected together into a single, connected "skeleton", as argued above. The reader should easily be able to verify that (b), (c), and (d) insure that all between-level T_0-components are connected to the skeleton and no cycles are created. Thus T connects all the points of S and is the desired tree.

3. THE L_2'-STEINER TREE PROBLEM IS NP-COMPLETE

In this section we sketch a proof that the STEINER TREE problem, with distance measured by our discretized L_2' metric, is NP-complete (full details can be found in [5]). Given $\mathfrak{F}$, we construct as before a set S′ of points organized into crossovers, junctions, etc., and a constant L′ such that a minimum L_2'-Steiner tree for S′ has length L′ or less if and only if $\mathfrak{F}$ has an exact cover.

For heuristic purposes, however, we shall first describe our construction as if it were taking place in ordinary L_2-space, obtaining a set S of points, some of which may possibly have irrational coordinates. The set S′ will be obtained from S by a process of scaling and rounding. The reason for working with L_2 as an intermediary is that a number of useful lemmas about minimum Steiner trees are easier to prove under that metric. Using these lemmas, we shall prove a theorem of the following form, for a specific L and $\delta > 0$.

Theorem 3.1 (a) If $\mathcal{J}$ has an exact cover, then S has a minimum L_2-Steiner tree of length L or less.
(b) If $\mathcal{J}$ does not have an exact cover, then a minimum L_2-Steiner tree for S has length at least $L + \delta$.

(A careful examination of the proof in the preceding section for the L_1 metric will show that our construction there satisfied a theorem analogous to the one above, with $\delta = 3$.)

The important thing about a theorem of the above form for L_2 is the gap $\delta > 0$ it provides between the length of a minimum Steiner tree when an exact cover for $\mathcal{J}$ does or does not exist. Rounding the coordinates of the points in S to integers to obtain S′ and converting from the L_2 to the L_2' metric will affect the length of a minimum Steiner tree, but with appropriate scaling beforehand, the cumulative effect can be kept less than $\delta/2$. Thus a residual gap will be left in the L_2' case, and NP-completeness for that case will follow.

The lemmas we shall use are presented without proof. (Missing details here and elsewhere can be found in [5,7,8,13].) Let T* be an L_2-minimum Steiner tree for S containing the least possible number of Steiner points.

Lemma 3.1 If two edges of T* meet at a common endpoint, the angle between them is 120° or more [7,13].

Lemma 3.2 Every Steiner point of T* has degree 3 and each of the three edges meeting at it makes angles of 120° with the other two [7,13].

In light of the above two lemmas, our construction will be arranged so that edges we wish to be present in T* do not meet at angles less than 120°. However, this alone will not insure that the analogues of the T_0-edges of Section 2 will all be present in T*. The situation under L_2 is more complicated than under L_1. Here we no longer have Hanan's result to restrict the locations of possible Steiner points, and must proceed by a more indirect route, using two additional lemmas. The first is true of minimum Steiner trees in general, but for convenience we state it in terms of the L_2 metric. If T is a spanning tree for S and $u,v \in S$, let $P_T(u,v)$ denote the path in T between points u and v. For any path P, let $m(P)$ be the L_2-length of the longest edge in P. Let $m(T)$ be the longest edge-length in the whole tree T.

Lemma 3.3 Suppose T is a spanning tree for S and T* is an L_2-minimum Steiner tree for S. Then $m(T^*) \leq m(T)$ and for any $u,v \in S$, $m(P_{T^*}(u,v)) \leq m(P_T(u,v))$. [5,8]

Our final lemma will be very useful in restricting Steiner points to a very narrow range of possibilities, and can be proved using Lemmas 3.2 and 3.3. Let T* be, as before, an L_2-minimum Steiner tree for S with the least possible number of Steiner points.

Lemma 3.4 Consider the region shown in Figure 10, which we shall call a probe. If $m(T^*) \leq 1$, then for each way of positioning the probe in the plane so that no points of S are in the probe or on its boundary, the point where the "tip" of the probe is located cannot be a Steiner point of T*. [5,8]

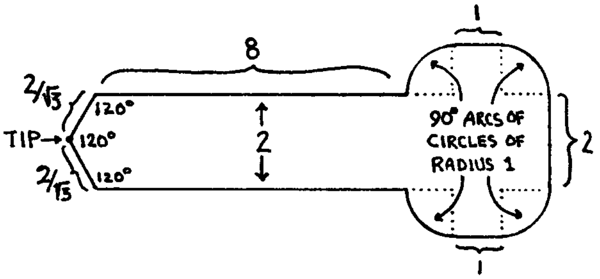


FIGURE 10. THE PROBE

We now describe the L_2 construction. Figure 11 shows the junction and Figure 12 shows the standard and warped crossovers. The value of K used here is given by

$$K = 3nt\sqrt{3} + n(1 + \sqrt{3}) \qquad (3.1)$$

Each line segment ℓ in our L_2 figures represents a set $S(\ell)$ of points as follows. Starting at one end of ℓ, divide ℓ into a sequence of subsegments: The first $\lceil 5K \rceil$ subsegments and the last $\lceil 5K \rceil$ subsegments all have length exactly 1/10. The subsegments in the middle all have lengths in the range [1/11, 1/10]. Such a subdivision can be made since all the line segments in our figures have length exceeding K + 1. We then let $S(\ell)$ be the set of all the endpoints of the subsegments. This rather involved definition is required so that $S(\ell)$ will be defined for line segments of non-integral and even irrational length. (Note that $S(\ell)$ can itself contain points with irrational coordinates, although it will be possible to choose these coordinates so that they can be represented symbolically.)

The junctions and crossovers are put together to form the overall structure S representing $\mathcal{J}$ in a fashion analogous to that used for the L_1 metric. The major difference is that junctions are joined to the backbone, and warped crossovers to the level above themselves, in such a way as to avoid angles of less than 120° and line segments of length less than K + 1. See Figure 13 for a schematic of the L_2 construction corresponding to Figure 5.

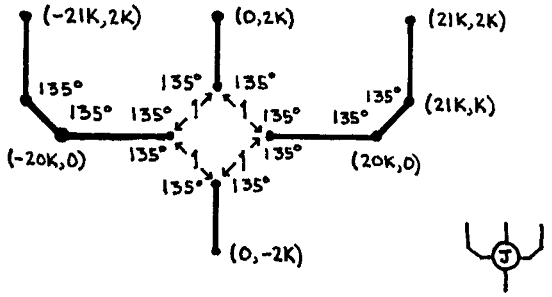


FIGURE 11. L_2-JUNCTION

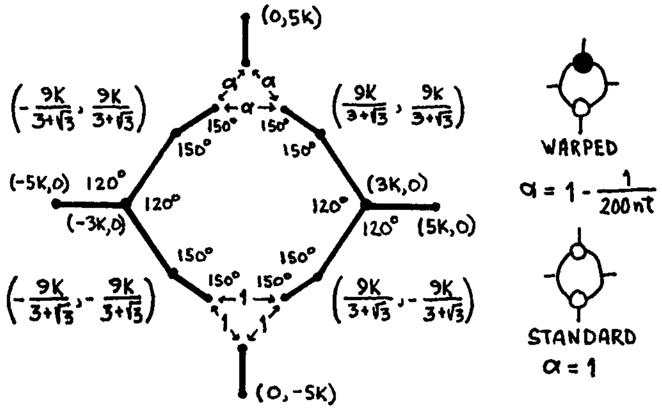

(0,5K)
α
α
←α→
150°
150°
150°
150°
$\left(-\frac{9K}{3+\sqrt{3}}, \frac{9K}{3+\sqrt{3}}\right)$
$\left(\frac{9K}{3+\sqrt{3}}, \frac{9K}{3+\sqrt{3}}\right)$
(-5K,0)
120°
120°
(-3K,0)
120°
(3K,0)
120°
(5K,0)
150°
150°
150°
150°
←1→
1
1
$\left(-\frac{9K}{3+\sqrt{3}}, -\frac{9K}{3+\sqrt{3}}\right)$
$\left(\frac{9K}{3+\sqrt{3}}, -\frac{9K}{3+\sqrt{3}}\right)$
(0,-5K)
WARPED
$\alpha = 1 - \frac{1}{200nt}$
STANDARD
α = 1

FIGURE 12. STANDARD AND WARPED L_2-CROSSOVERS

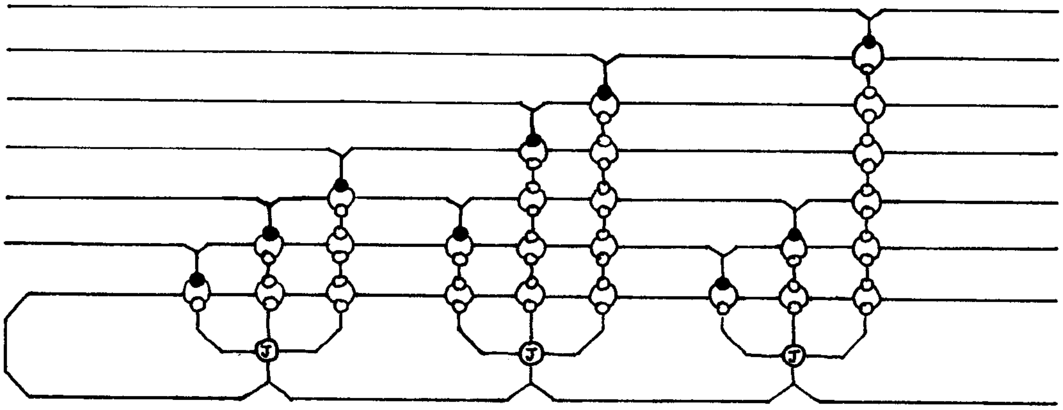


FIGURE 13. FINAL L_2 CONSTRUCTION FOR $\mathcal{F} = \{\{1,2,3\},\{2,4,5\},\{1,2,6\}\}$

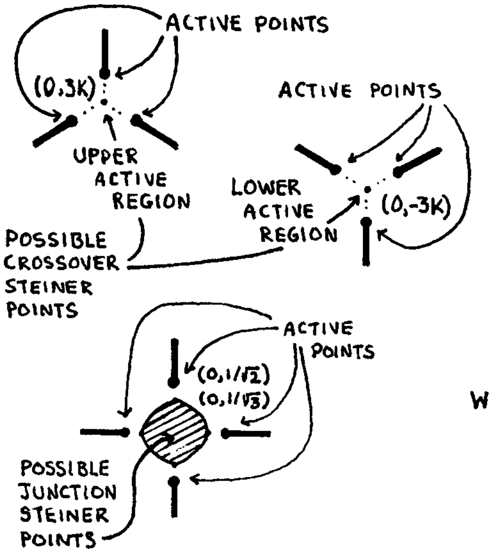
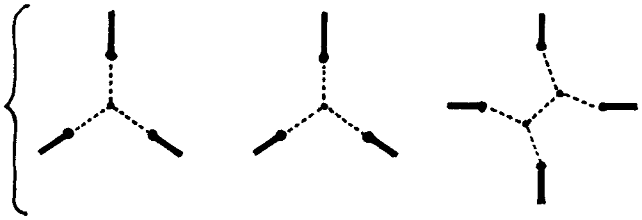


FIGURE 14. POSSIBLE LOCATIONS OF L_2 STEINER POINTS



TYPE	α_1	β_1	γ_1
ACTIVE REGIONS WHERE IT OCCURS	STANDARD CROSSOVER: UPPER & LOWER WARPED CROSSOVER: LOWER	WARPED CROSSOVER: UPPER	JUNCTION
LENGTH PER CONNECTION	$\frac{\sqrt{3}}{2} = .866...$	$(1-\varepsilon)\frac{\sqrt{3}}{2} = (1-\varepsilon)(.866..)$	$\frac{1+\sqrt{3}}{3} = .910...$

FIGURE 15. POSSIBLE L_2 CONNECTING CONFIGURATIONS $\left(\varepsilon = \frac{1}{200nt}\right)$

We now begin a proof that Theorem 3.1 holds for S, with L and δ to be specified later. Let T* be an L_2-minimum Steiner tree for S containing the least possible number of Steiner points. Since there is a spanning tree for S with maximum edge length 1, we know by Lemma 3.3 that every edge of T* has length 1 or less. Thus Lemma 3.4 applies, and by using its "probe" we can substantially limit the possible locations of Steiner points.

Claim 3.1 A Steiner point in T* can occur only in a location corresponding to one of the following (see Figure 14)

(a) the points (0, 3K) and (0, -3K) in a crossover,

(b) the region defined by $\{(x,y):|x^2+xy\sqrt{3}+y^2| \leq 1/3\}$ in a junction.

In light of Claim 3.1, no Steiner point is closer than $1/\sqrt{2} - 1/\sqrt{3} = 0.1297... > 1/10$ to any point in S. Thus, if we let T_0 be the set of all edges between pairs of points which are no further than 1/10 apart, Lemma 3.3 implies the following.

Claim 3.2 T* contains all edges of T_0

Thus we know, as in Section 2, that T* is made up of T_0-components which are interlinked in some manner by non-T_0 edges. From now on, we shall assume that T* has length less than $D_0 + K$, where D_0 is the total length of the T_0 edges, and K is as in (3.1). Given this, the non-T_0 edges have total length less than K. Moreover, it is a fairly straightforward task to prove the following claim, using the nature of our construction and the distances involved, along with Lemma 3.1.

Claim 3.3 If $\langle u,v \rangle$ is an edge of T*, but is not in T_0, and if $u \epsilon S$, then u corresponds to one of the points labeled as "active points" in Figure 14.

Thus the T_0-components can be interconnected only in the vicinity of the possible Steiner points, areas which we again call "active regions". As before, we provide a list of minimum length connecting configurations for each of the possible ways of connecting the T_0-components within a given active region (see Figure 15). Symmetric cases have been combined, and all cases with average length per connection exceeding $(1 + \sqrt{3})/3$ have been omitted, because it will turn out that they are too costly. Note that this leaves only three relevant configurations, which correspond in a natural way to the three configurations used in the L_1 case. In a continuing analogy with the L_1 case, we now can prove Theorem 3.1 for S, with L and δ specified as follows. Let $\varepsilon = 1/(200nt)$ and $q < 3nt$ be the number of crossovers in S. We then set

$$L = D_0 + q\sqrt{3} + n(1 + \sqrt{3}) - 3n\varepsilon\sqrt{3} < D_0 + K \tag{3.1}$$

$$\delta = \varepsilon \tag{3.2}$$

We construct an input to the L_2'-STEINER TREE problem from S and L in two steps. The first step is to scale the problem up. Let M = |S| and let

$$S'' = \left\{\left(\frac{12M}{\varepsilon} x, \frac{12M}{\varepsilon} y\right) : (x,y) \epsilon S\right\} \tag{3.3}$$

Theorem 3.1 now holds for S″ with $L'' = \frac{12M}{\varepsilon} L$ and $\delta'' = 12M$.

The next step is to round the coordinates up to integers. For $x = (x_1, x_2)$, define f(x) to be $(\lceil x_1 \rceil, \lceil x_2 \rceil)$. Then set

$$S' = \{f(x) : x \epsilon S''\} \tag{3.4}$$

Observe that there is a natural correspondence between Steiner trees for S″ and ones for S′. This correspondence may not preserve minimality, but the length of an individual edge cannot change by much, even as we go from the L_2 to the L_2' metric. The change is made up of a contribution of less than $\sqrt{2}$ due to the translation of the edge's endpoints, and a contribution of less than 1 due to the change in measure, for a total change of less than 3. Recalling from Section 1 that a minimum Steiner tree for a set with M points need have at most M-2 Steiner points and hence at most 2M-3 edges, we can thus conclude the following.

Theorem 3.2 S′ has an L_2'-minimum Steiner tree of length less than $\lceil L'' \rceil + 6M$ if and only if $\mathfrak{J}$ has an exact cover.

Since it is clearly possible (although admittedly a complicated process) to construct S′ in time bounded by a polynomial in n and t, Theorem 3.2 leads to the desired conclusion that the L_2'-STEINER TREE problem is NP-complete. Moreover, note that the change in Steiner tree edge length as we go from S″ to S′ is still less than 3 if we use the L_2 metric for both. Thus Theorem 3.2 also holds if L_2' is replaced by L_2, and consequently the L_2-STEINER TREE problem, even when restricted to integer coordinate inputs, is NP-hard.

4. THE L_1 and L_2' TRAVELING SALESMAN PROBLEM ARE NP-COMPLETE

Our TRAVELING SALESMAN constructions will follow the same general scheme as did our STEINER TREE constructions. A set S will be built up out of junctions, crossovers, etc. However, instead of using single rows of closely spaced points to build the junctions and crossovers and to link them together, we shall use pairs of parallel rows. These will in effect form "tubes", whose interiors will be forced to be "inside" the TRAVELING SALESMAN circuit.

To explain more clearly what we mean by "inside" a circuit, we must first set up a correspondence between a circuit of S (thought of as a sequence of edges in a graph) and the representation of such a circuit by line segments in the plane. For the two distance metrics L_1 and L_2', it is not true that a straight line is the unique shortest path between two points in the plane, as is the case under the L_2 metric. For the L_1 metric, there can be infinitely many paths of length $d_1(x,y)$ between x and y made up of horizontal and vertical line segments, so long as x and y do not agree in either coordinate. Similarly, if $d_2(x,y)$ is not an integer, there can be infinitely many paths made up of straight line segments that go from x to y with total length $d_2'(x,y) = \lceil d_2(x,y) \rceil$.

Thus, we shall say that a representation of a given edge $\langle x,y \rangle$ under distance measure d is any path from x to y made up of line segments whose total length under d is $d(x,y)$. In the L_1 case we make the further restriction that all the line segments be either horizontal or vertical. A representation of a circuit C under d is a collection of edge representations, one for each edge in C.

These definitions allow us to prove the following lemma, using the triangle inequality.

Lemma 4.1. If S is a set of points in the plane which are not all collinear and $L \epsilon \{L_1, L_2'\}$, then under L there is a minimum length circuit of S which visits each point exactly once and which has a representation in the plane in which no two edge representations intersect or overlap except at a common endpoint.

Observe that by the Jordan Curve Theorem [14], such a representation must divide the plane into two connected regions, one inside the circuit and one outside. Our "tube" construction will force all the tube interiors to be inside the circuit. Since the tube interiors will only be able to connect up with each other in "active regions" of junctions and crossovers, we can see the analogy with the STEINER TREE case becoming more apparent. To complete the analogy, we observe that just as we could not make a connection in the STEINER TREE case if it would create a cycle, here we cannot make a connection if it will make a "hole" in the inside region, as this would mean that the plane was divided into at least three regions by the circuit representation.

We now begin the actual construction. In contrast to the case of STEINER TREE, no L_2 intermediary is needed for the L_2' construction. In fact the L_2' construction is so similar to that for L_1 (they differ only in the fine structure of their junctions) that we shall present the two in parallel. Given $\mathfrak{J}$, the corresponding junctions and crossovers are shown in Figures 16 and 17, where

$$K = 108nt^2 + 1008n^2t^2 + 108n^2t. \qquad (4.1)$$

Each line segment once more stands for the set of integer coordinate points it contains. Note that in the crossovers, the central point (0,0) is included in the set of points the crossover represents. (The point (0,0) is not included in the junctions.) For both junctions and crossovers, the active region is defined to be the set of all points within distance 3K of (0,0), under the appropriate metric.

These basic units are put together to form an overall structure representing $\mathfrak{J}$ in a fashion analogous to that for the STEINER TREE constructions, with what previously were connected components now being connected "tube" systems. See Figure 18 for a schematic of the construction corresponding to Figures 5 and 13.

Let T_0 be the set of all edges of length 1 between points of S (under the relevant metric), and let $q < 3nt$ be the number of crossovers in S. We shall show that $\mathfrak{J}$ has an exact cover if and only if there is a circuit passing through all the points of S with total length not exceeding

$$L = \begin{cases} |T_0|+108nt^2+336ntq+108n^2t-6n, & \text{under } L_1 \\ |T_0|+72nt^2+312ntq+108n^2t-6n, & \text{under } L_2' \end{cases} \qquad (4.2)$$

We shall argue in parallel for both metrics, distinguishing between them only when necessary.

Let C* be a minimum length circuit of S, with length $\ell^* \leq L$. By Lemma 4.1, we can assume that every vertex $s \epsilon S$ has degree 2 in C*, and that there is a representation R(C*) of C* in the plane which does not intersect or overlap itself.

We first note that $|S| \geq |T_0|$. Since all edges of C* must have length at least 1 by our construction, we thus have $\ell^* \geq |T_0|$. In fact, if we let $\langle x_1, x_2, \ldots, x_{|S|} \rangle$ be the cyclic permutation of S induced by C*, then we have

$$\sum_{i=1}^{|S|} [d(x_i,x_{i+1})-1] + d(x_{|S|},x_1) - 1 \leq L-|T_0| < K \qquad (4.3)$$

From this and the fact that R(C*) does not intersect or overlap itself, we can derive the following.

Claim 4.1 Suppose $\langle a,b \rangle \epsilon T_0$ and all points of S within distance K of a or b are on the same horizontal or vertical line as a and b. Then $\langle a,b \rangle \epsilon C^*$.

From Claim 4.1 we can conclude that C* contains all edges of T_0 which are not in an active region or within distance K of a place where line segments meet at 90° angles. A simple argument using (4.3) and the fact that all points of S have degree 2 in C* then suffices to prove the following.

Claim 4.2 C* contains all edges of T_0 that are not in active regions.

This means that all points of S outside the active regions have their two edges in C* supplied by T_0. Since all active regions are at least 2K apart, the remaining edges of C* must each be between points in the same active region, and hence must have their representations entirely contained within single active regions. Thus Figure 18 indicates how R(C*) must look outside the active regions.

Since R(C*) does not overlap or intersect itself, it divides the plane into two connected regions. A simple coloring argument now suffices to show that the interiors of the "tubes" are all part of the same region. Let an inactive segment

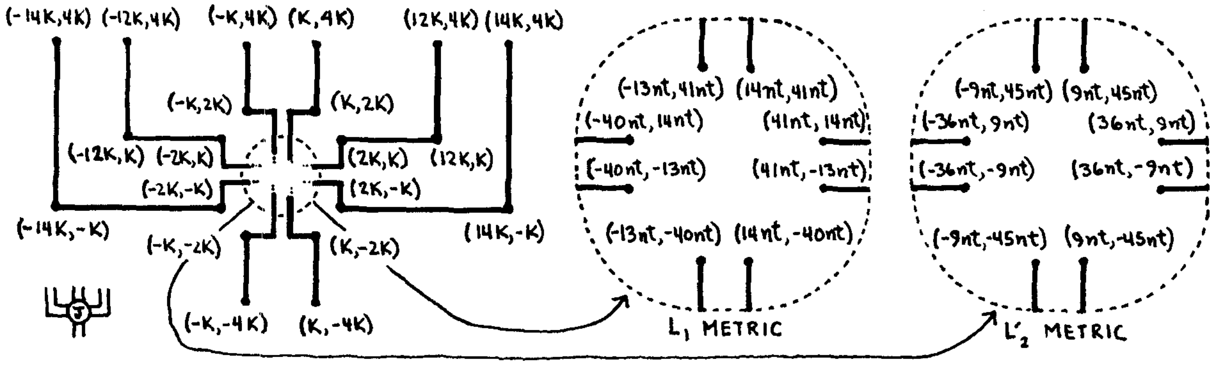


FIGURE 16. TRAVELING SALESMAN JUNCTIONS UNDER L_1 AND L'_2 METRICS

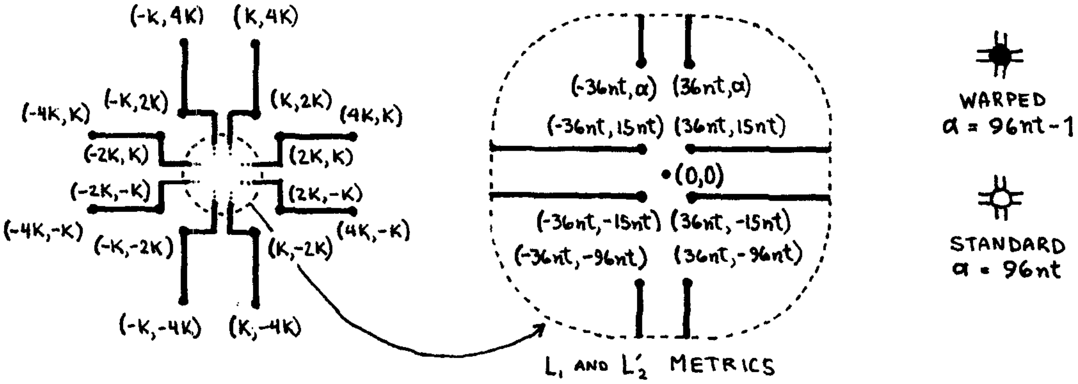


FIGURE 17. STANDARD AND WARPED TRAVELING SALESMAN CROSSOVERS UNDER BOTH L_1 AND L'_2 METRICS

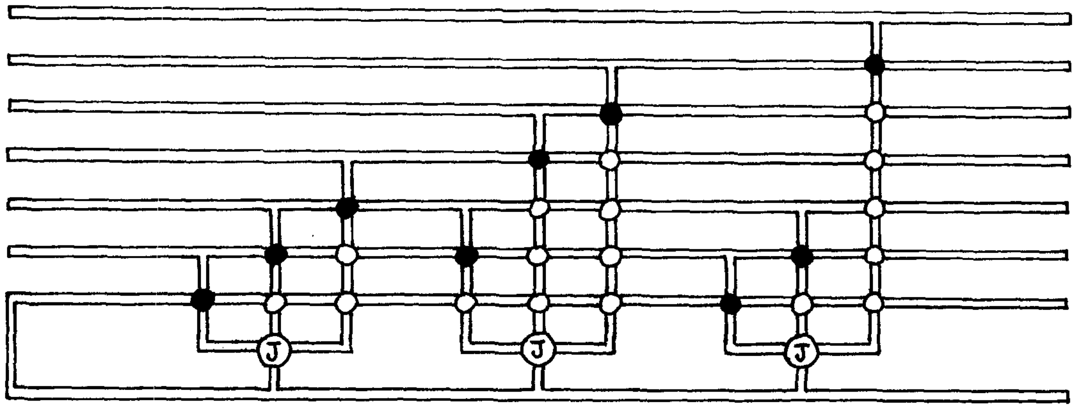


FIGURE 18. FINAL TRAVELING SALESMAN CONSTRUCTION FOR $\mathfrak{F}$ = {{1,2,3},{2,4,5},{1,2,6}} UNDER L_1 AND L'_2

of R(C*) be a path made up of T_0 edge representations which are not in active regions. That part of the plane which is not contained in active regions can then be thought of as made up of inactive regions, which are separated from each other by inactive segments. Clearly each inactive region must either be entirely "inside" R(C*) or entirely "outside", and each inactive segment separates an inside region from an outside one. We color the inactive regions as follows. Start with any tube interior region, and color it red for "inside". Now pick an inactive region which borders our red region, and color it blue for "outside". Continue in this way, always choosing an uncolored inactive region which borders a colored one, and giving it the opposite color. It should be easy for the reader to verify that this will yield a unique coloring of the inactive regions, in which no two adjacent regions get the same color, all tube interiors are red, and all other inactive regions are blue. No colors have been assigned to the active regions, and indeed these regions will each be part "inside" and part "outside". However, the edge representations in R(C*) in the active regions must be such that all the red inactive regions belong to the connected region "inside" R(C*), and all the blue inactive regions belong to the single "outside" connected region.

Thus once again we can think of the active regions of S as performing "connections" - this time of tube interiors rather than of T_0-components, as in the STEINER TREE case. Moreover, there are 3n + 2q + 1 inactive regions which are tube interiors, and hence there are, as before, 3n + 2q connections to be made.

However, unlike the STEINER TREE case, the cost of making "no connections" in an active region is not zero. In addition to the T_0 edges, non-T_0 edges must be included to insure that each time R(C*) enters the region it continues along an unbroken path until it leaves the active region (i.e., when no connections are made the tube interior regions that enter the active region must be "closed off"). Moreover, in the crossovers non-T_0 edges will be needed to insure that the "central point" of the crossover is included in C*. Figure 19 gives canonical ways of achieving "no connections" for both junctions and crossovers. Observe that all T_0 edges in the active regions are used. The "base length" quoted in the figure is the total length of the non-T_0 edges used.

For configurations that perform one or more connections, we shall compute "excess length" as the difference between the total length needed to make the connections and the total length needed to make no connections. Canonical ways of achieving one or more connections using minimum excess length are shown in Figure 20. We omit all connection possibilities that require average excess length exceeding 36nt, as they will prove too expensive. We can assume without loss of generality that each active region of R(C*) contains one of our canonical configurations.

We thus have 3n + 2q connections to be performed, and no connection can be made which creates a "hole" in the inside region of R(C*), just as no connection could be made which created a cycle in the STEINER TREE case. The reader should now be able to complete the proof using Section 2 as a guide.

We thus conclude that the desired circuit exists if and only if the desired cover exists. Since the construction is clearly polynomially bounded, this means that the L_1 and L_2' TRAVELING SALESMAN problems are NP-complete.

We conclude by remarking that the L_2' construction works equally well for the L_2 TRAVELING SALESMAN problem restricted to integer coordinate inputs. The crossovers and junctions were designed so that any edge $\langle x,y \rangle$ usable in C* would have integral length under L_2, and hence $d_2(x,y) = d_2'(x,y)$. The reader may verify that the lemmas and claims continue to hold when L_2' is replaced by L_2. Thus we can conclude that this L_2 problem is NP-hard, although the technical problems mentioned in the introduction leave the question of NP-completeness open.

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, Mass., (1974) Chapter 10.

[3] W. M. Boyce and J. B. Seery, "STEINER 72, An Improved Version of Cockayne and Schiller's Program STEINER for the Minimal Network Program", Bell Laboratories Computing Science Technical Report #35 (1975).

[4] E. J. Cockayne and D. G. Schiller, "Computation of Steiner Minimal Trees", in Combinatorics, D. J. A. Welsh and D. R. Woodall (eds.), Inst. Math. Appl., (1972) 53-71.

[5] M. R. Garey, R. L. Graham, and D. S. Johnson, "The Complexity of Computing Steiner Minimal Trees", (to appear).

[6] M. R. Garey and D. S. Johnson "The Rectilinear Steiner Tree Problem is NP-Complete", (to appear).

[7] E. N. Gilbert and H. O. Pollak, "Steiner Minimal Trees", SIAM J. Appl. Math., 16 (1968) 1-29.

[8] R. L. Graham, "Some Results on Steiner Minimal Trees", Bell Laboratories Technical Memorandum, (1967).

[9] M. Hanan, "On Steiner's Problem with Rectilinear Distance", SIAM J. Appl. Math., 14 (1966), 255-265.

[10] S. Lin and B. W. Kernighan, "An Effective Heuristic Algorithm for the Traveling-Salesman Problem", BSTJ 21 (1973), 498-516.

[11] R. M. Karp, "Reducibility Among Combinatorial Problems", in Complexity of Computer Computations, R. E. Miller and J. W. Thatcher (eds.), Plenum Press, New York, 1972, 85-104.

[12] R. M. Karp, "On the Computational Complexity of Combinatorial Problems", Networks, 5 (1975), 45-68.

[13] Z. A. Melzak, "On the Problem of Steiner", Canad. Math. Bull., 4 (1961), 143-148.

[14] M. H. A. Newman, Elements of the Topology of Plane Sets of Points, Cambridge University Press, Cambridge, (1964) 115-116.

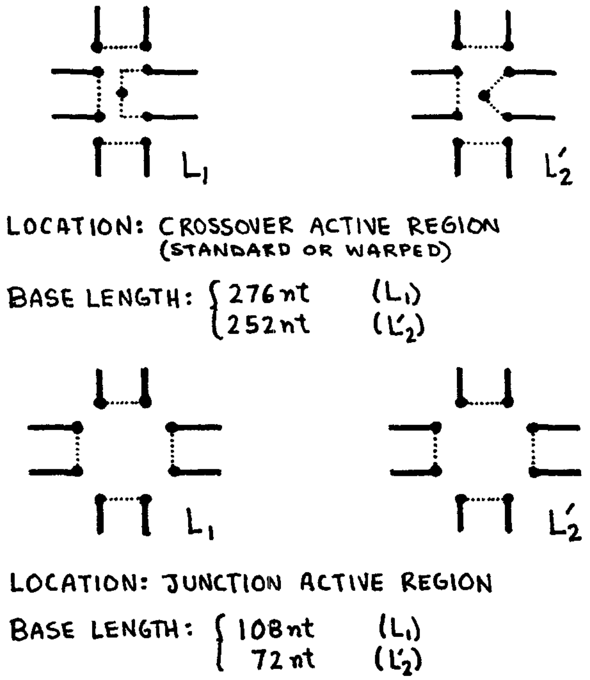
[15] A. M. Odlyzko, personal communication.

[16] C. H. Papadimitriou, "The Euclidean Traveling Salesman Problem is NP-Complete", Princeton University Computer Science Technical Report No. 191 (1975).

[17] D. J. Rosenkrantz, R. E. Stearns, and P. M. Lewis, "Approximate Algorithms for the Traveling Salesperson Problem", 15th IEEE Annl. Symp. on Switching and Automata Theory, (1974), 33-42.

[18] M. I. Shamos, "Geometric Complexity", 7th Annl. ACM Symp. on Theory of Computing, (1975), 224-233.

[19] M. I. Shamos and D. Hoey, "Closest Point Problems", 16th Annl. Symp. on Foundations of Computer Science, IEEE, 1975, 151-162.



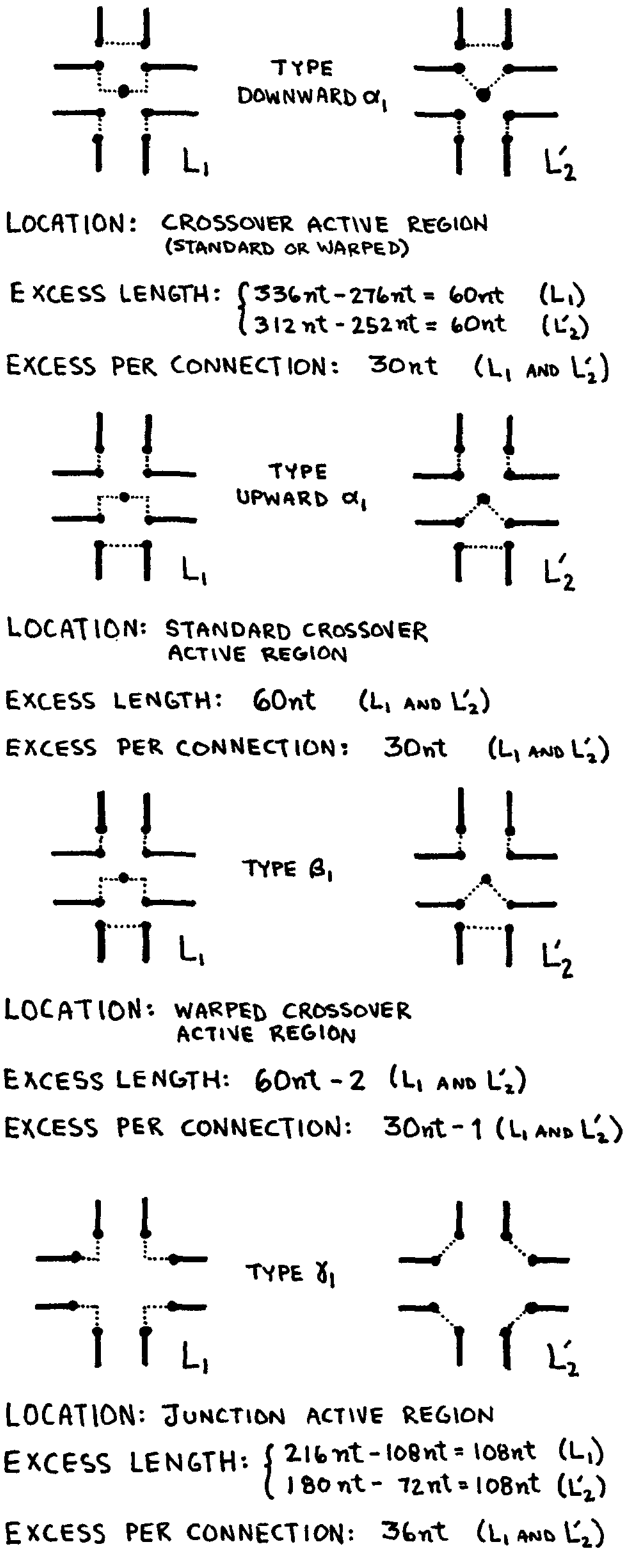
LOCATION: CROSSOVER ACTIVE REGION (STANDARD OR WARPED)

BASE LENGTH: $\begin{cases} 276nt & (L_1) \\ 252nt & (L_2') \end{cases}$

LOCATION: JUNCTION ACTIVE REGION

BASE LENGTH: $\begin{cases} 108nt & (L_1) \\ 72nt & (L_2') \end{cases}$

FIGURE 19. BASE CONFIGURATIONS IN TRAVELING SALESMAN ACTIVE REGIONS (NO CONNECTIONS)



LOCATION: CROSSOVER ACTIVE REGION (STANDARD OR WARPED)

EXCESS LENGTH: $\begin{cases} 336nt - 276nt = 60nt & (L_1) \\ 312nt - 252nt = 60nt & (L_2') \end{cases}$

EXCESS PER CONNECTION: $30nt$ (L_1 AND L_2')

LOCATION: STANDARD CROSSOVER ACTIVE REGION

EXCESS LENGTH: $60nt$ (L_1 AND L_2')

EXCESS PER CONNECTION: $30nt$ (L_1 AND L_2')

LOCATION: WARPED CROSSOVER ACTIVE REGION

EXCESS LENGTH: $60nt - 2$ (L_1 AND L_2')

EXCESS PER CONNECTION: $30nt - 1$ (L_1 AND L_2')

LOCATION: JUNCTION ACTIVE REGION

EXCESS LENGTH: $\begin{cases} 216nt - 108nt = 108nt & (L_1) \\ 180nt - 72nt = 108nt & (L_2') \end{cases}$

EXCESS PER CONNECTION: $36nt$ (L_1 AND L_2')

FIGURE 20. POSSIBLE CONNECTIONS IN TRAVELING SALESMAN ACTIVE REGIONS