PACIFIC JOURNAL OF MATHEMATICS
Vol. 167, No. 2, 1995

EXISTENCE OF SHORTEST DIRECTED NETWORKS IN $\mathbb{R}^2$

MANUEL ALFARO

This paper establishes the existence of a shortest directed network connecting a given set of points. In such networks, up to six segments sometimes meet at a point.

1. Introduction. The standard Steiner problem considers shortest undirected networks, and at most three segments meet at a point ([**CR**, pp. 354-361], [**BG**], [**M1**], [**M2**]).

1.1. Definitions. A *directed network* is a finite system of one-way roads (oriented straight line segments) connecting all of a given set of starting points to all of a given set of ending points. (See Figure 1.1.) We refer to the given starting and ending points as *boundary points.* The *nodes* are any other points where the segments meet. We require that boundary points and nodes occur only at the endpoints of segments. Two segments meeting at a point count as one node. For $m \geq 3$, m segments meeting at a point count as $m-2$ nodes. When counting the number of edges meeting at a point, a *double edge* (an edge with both orientations) counts once, although its length counts twice. A *region* is the closure of a bounded component of the complement of the network.

1.2. Existence of length-minimizing directed networks in $\mathbb{R}^2$. The difficulty in demonstrating the existence of shortest directed networks lies primarily in obtaining an upper bound on the number of nodes in the network. In previous studies of networks in other settings, the number of nodes in the networks could be easily estimated since their minimizing networks obviously contained no cycles [**Ab**], [**A4**], [**L**]. Unfortunately, shortest directed networks may contain cycles. The number of cycles containing at least one boundary point may be estimated by the number of boundary points. Since shortest directed networks may have cycles containing no boundary

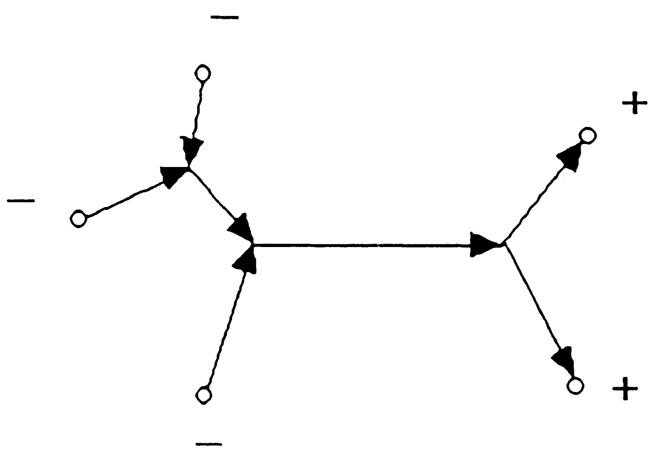

FIGURE 1.1. A directed network connecting three starting points (-) to two ending points (+) via three nodes.

points, we must first find an upper bound on the number of such cycles in order to bound the number of nodes.

We consider regions bounded by polygons. It turns out that if R is a region containing no boundary points, it is the only such region not containing any boundary points in the network. We prove this by first showing that R has interior angles of at most 120 degrees (Lemma 2.3). Then, by showing that R cannot share any edges or nodes with any other region (Lemma 2.5), we prove that it is the only region in the network containing no boundary points (Theorem 2.6) and thus obtain an upper bound on the number of regions in the network.

Given an upper bound on the number of regions in the network, we may use standard graph theory arguments to bound the number n of nodes in the network. Then standard compactness arguments yield the existence of a shortest directed network (Theorems 2.1, 2.8).

It is an open question whether our results generalize to $\mathbb{R}^n$.

1.3. Structure of singularities of length-minimizing directed networks in $\mathbb{R}^n$. It was known that segments in shortest directed networks in the plane may meet in threes and fours but never in sevens or more [**A2**, §3]. The argument of Lemma 2.7 actually generalizes that result to $\mathbb{R}^n$. Moreover, one can show that shortest directed networks can meet in fives and sixes, already in the

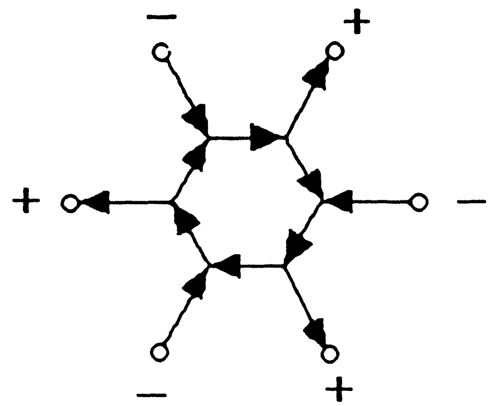

FIGURE 1.3. A shortest network can contain an interior cycle.

plane [**A1**]. This gives a complete characterization of singularities in shortest directed networks.

Indeed, the network consisting of six rays from the origin to the six vertices of a regular hexagon, alternately labeled $+$ or $-$, is a shortest directed network. Replacing the portion inside a disc about the origin by a regular hexagon yields an equally short network and demonstrates that a shortest directed network can contain a cycle which does not pass through any of the given boundary points (see Figure 1.3). I do not know of any different example of such cycles.

2. Existence. This chapter establishes the existence of length-minimizing directed networks. Theorem 2.1 shows that it suffices to bound the number of nodes (counting multiplicities as in 1.1).

THEOREM 2.1. *Given a set of boundary points, if there exist networks with at most n nodes, then there is a shortest one among those with at most n nodes.*

Proof. Let $\{N_k\}$ be a sequence of networks with at most n nodes and lengths approaching the infimum. We may assume that the networks are connected and that they are contained in some large ball B (since their length is bounded by some large positive number). We may assume (by taking a subsequence) that the networks all have exactly $s \leq n$ nodes. Consider the sequence of s-tuples of nodes in B^n. Since B^n is compact, we may assume the sequence converges to nodes $a_1, \cdots, a_s$. Since there are only finitely many ways to connect s nodes, we may assume (by taking a subsequence) that all s-tuples of nodes are connected the same way. Hence, if we connect the limit in the same way, the length of the limit is less than or equal to the limit of the length, which equals the infimum.

To show that in the limit the number of nodes $M(N_\infty)$ in N_∞ is less than or equal to s we let A be a node in the limit. Let U be a small disc about A. (See Figures 2.1.1, 2.1.2.) In the limit, $N_\infty \cap U$ looks like a wheel with s spokes. For k large, $N_k \cap U$ resembles $N_\infty \cap U$ except possibly in a small disc about A.

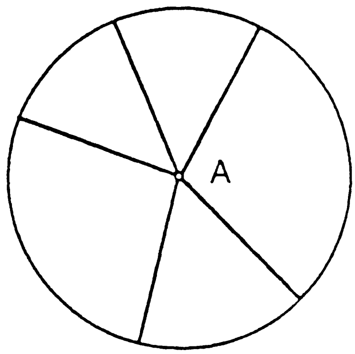

FIGURE 2.1.1. In the limit, $N_\infty \cap U$ looks like a wheel with s spokes. $N_k \cap U$ resembles $N_\infty \cap U$ except possibly in a small disc about A.

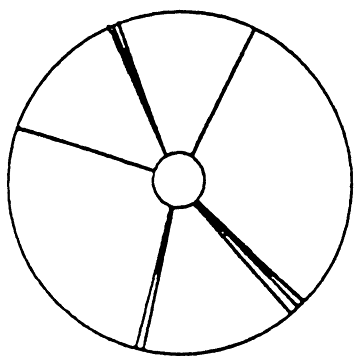

FIGURE 2.1.2. We do not know what is happening inside the smaller disk.

If any connected component of $N_k \cap U$ spans less then 180° in the limit, or spans exactly 180° and has another segment in between, pushing this component out a bit yields a contradiction. (See Figure 2.1.3.)

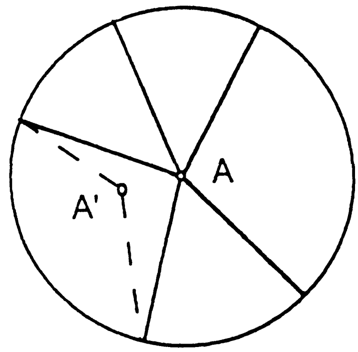

FIGURE 2.1.3. An angle less than 180° could be shortened.

This leaves two cases:

(i) $N_k \cap U$ is connected.

(ii) $N_\infty \cap U$ is two spokes at a 180° angle (see Figure 2.1.4).

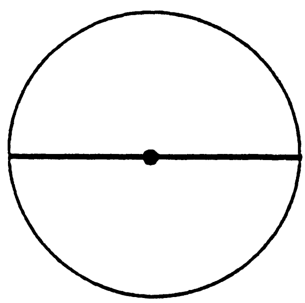

FIGURE 2.1.4. Two spokes meeting at 180°.

In case (ii), the network only has one node here, so we are done. For case (i), denote the nodal points in $N_\infty \cap U$ as $P_1, \ldots, P_t$. Let their multiplicities be m_j. Each P_j has at most $m_j + 2$ edges emanating from it. Since $N_k \cap U$ is connected, at least $t-1$ edges collapse in the limit. Hence, at most

$$\begin{aligned}\sum_{j=1}^{t}(m_j+2) - 2(t-1) &= \sum_{j=1}^{t} m_j + 2t - 2t + 2 \\ &= \sum_{j=1}^{t} m_j + 2\end{aligned}$$

segments emanate from A in $N_\infty \cap U$. Hence, the number of nodes at A

$$M(A) \leq \sum_{j=1}^{t} m_j.$$

Thus, the number of nodes does not increase in the limit. □

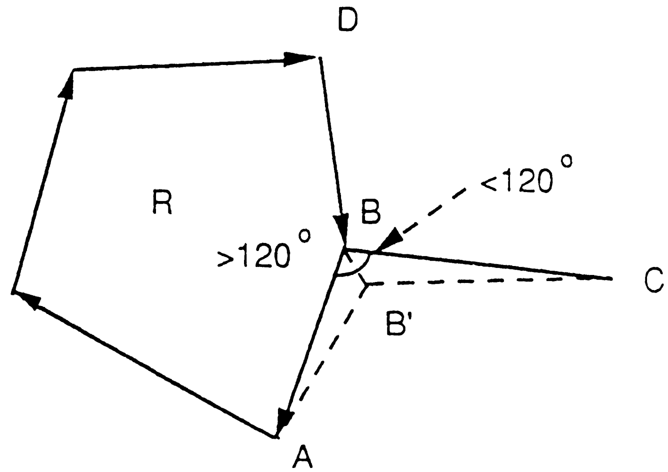

FIGURE 2.3.1. An angle greater than 120° would yield a shorter network.

LEMMA 2.2. *Suppose two segments AP and BP meet at a point P at an angle of less than* 120 *degrees. Let O be a point a small distance from P along the angle bisector. Then the length of $Y = AO \cup BO \cup OP$ is less than the length of $V = AP \cup BP$.*

Proof. Calculation, see [**A4**, §2]. □

LEMMA 2.3. *Let N be a directed network, shortest among those with at most n nodes. If R is a region in N containing no boundary points, then N does not enter the interior of R and R is a polygonal region with interior angles of at most* 120 *degrees.*

Proof. First, the network does not enter the interior of R. If it did, removing the edges in the interior of R would shorten the total length of N (the edges would be superfluous since R contains no boundary points). The boundary β of R, a polygon, can be oriented to form a cycle (since starting points connected to β can still get to all destinations by simply going around the cycle). If the hypothesis of the lemma fails, at least one interior angle of β is greater than 120°, with at least one additional edge BC emanating from this vertex B of β (see Figure 2.3.1). (Of course two segments could meet only at 180°, but we are ignoring those angles.) So, one edge AB of β and BC must form an angle less than 120°. By Lemma 2.2, replacing $AB \cup BC$ by $AB' \cup BB' \cup B'C$ where B' is a point a small distance from B along the bisector of ABC (and replacing $DB \cup BB'$ by DB' if BC is the only edge emanating from B), and orienting each of the segments such that N remains connected would decrease the total length of N without increasing the number of nodes. At worst, the number of nodes remains constant. Therefore, a region R in N containing no boundary points is a polygonal region with interior angles of at most 120°. □

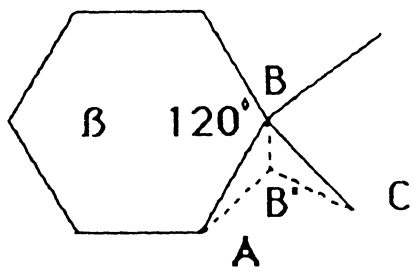

FIGURE 2.4.1. If two additional edges emanated from a vertex, the network could be shortened.

COROLLARY 2.4. *Let N be a shortest directed network among those with at most n nodes. Let β be the boundary of a region R, in N, containing no boundary points. Then if all the interior angles of β are equal to* 120° *degrees, each of the vertices of β has precisely one edge not in β emanating from it.*

Proof. First, β has no double edges emanating from its vertices (since replacing an edge DE from the double edge and an edge from β adjacent to DE by the third side of the triangle they determine would decrease the total length of N). Now, suppose a vertex B of β has two or more additional edges emanating from it (see Figure 2.4.1). Since all the interior angles of β are precisely 120°, at least one of these edges, BC (say), must form an angle of less than 120° with a side BA of β adjacent to it. By Lemma 2.2, we can replace ABC with $AB'C \cup BB'$, where B' is a small distance from B along the bisector of ABC, without increasing the number of nodes or disturbing connectedness. Orienting the new ∂R coherently keeps N connected. The number of nodes remains constant (one fewer at B, one more at B'). Therefore each vertex of β has precisely one additional edge emanating from it. □

LEMMA 2.5. *Let N be a shortest directed network among those with at most n nodes. If R_1 and R_2 are regions in N, and if R_1 contains no boundary points, then R_1 and R_2 are disjoint.*

Proof. Suppose R_1 and R_2 are not disjoint. Then one of the following holds:

(i) R_1 and R_2 share only vertices,

(ii) R_1 and R_2 share precisely one edge,

(iii) R_1 and R_2 share precisely two edges, or

(iv) R_1 and R_2 share three or more edges.

(i) R_1 and R_2 share only vertices (see Figure 2.5.1). Let A be a vertex shared by the boundary of R_1 and R_2. Let β_1 and β_2 be the boundaries of R_1 and R_2. Since the interior angles of R_1 and R_2 are less than 180, but greater than 0°, an edge BA of β_1 and an edge CA of β_2 meeting at A must form an angle less than 180°. Projecting A to A', where A' is a small distance along the bisector of BAC, and connecting B to A', C to A', and all other segments meeting at A falling in the region spanned by BAC to A' (and reorienting the boundary of the new region, without loss of generality, clockwise) would decrease the total length of N without disturbing connectedness. Further, the number of nodes (counting multiplicities) would not increase.

(ii) If R_1 and R_2 share precisely one edge (Figure 2.5.2), removing the shared edge (and reorienting what is left of the boundaries of R_1 and R_2, without loss of generality, clockwise) would decrease the total length of N without disconnecting it or increasing the number

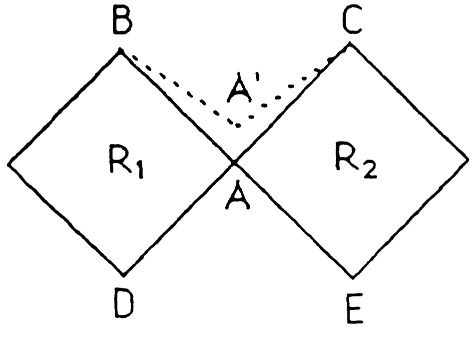

FIGURE 2.5.1. If two regions intersect in a point, the network could be shortened.

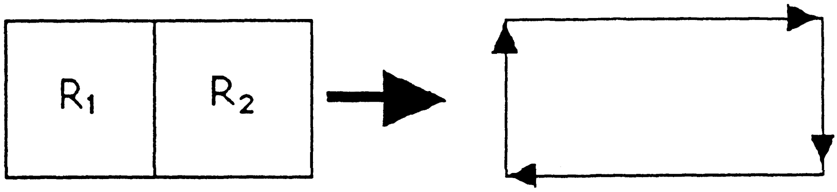

FIGURE 2.5.2. If the regions share one edge, the network can be shortened.

of nodes, a contradiction.

(iii) Suppose R_1 and R_2 share precisely two edges (Figure 2.5.3). First, if one of the shared edges is shorter than the other, make the shortest of the two a double edge and remove the other. Second, replace a side of R_1 (adjacent to the double edge) and an edge from the double edge by the third side of the triangle they determine. This decreases the total length of N. To insure that N is connected, reorient the boundary of the new region, clockwise (say).
The number of nodes does not increase because the number of segments meeting at A and C remains constant, while the number of segments meeting at B and D decreases by one (respectively), a contradiction.

(iv) If R_1 and R_2 share three or more edges and R_1 is a polygon with five or fewer sides, then using an argument analogous to that of cases (ii) and (iii) on the two edges of R_1 not shared with R_2 yields a contradiction. By Lemma 2.3, R_1 cannot have seven or more sides. Finally, if R_1 has six sides, then each of its vertices has precisely one edge emanating from it (by 2.3 and 2.4). So, if R_1 and R_2 share precisely three edges and R_1 is a hexagon (see

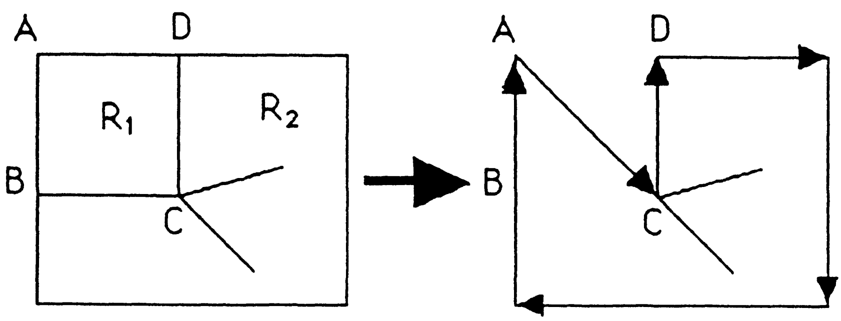

FIGURE 2.5.3. If the regions share two edges, the network can be shortened.

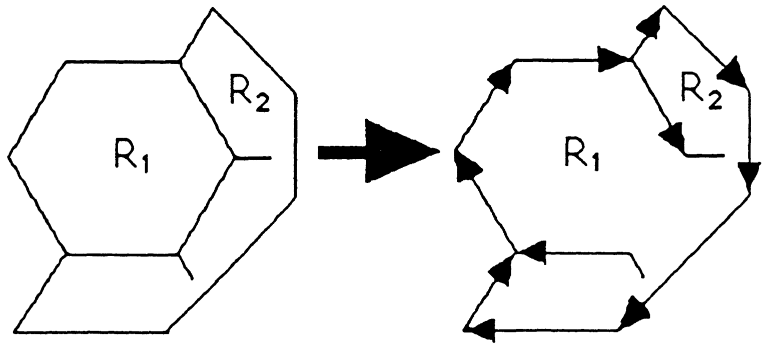

FIGURE 2.5.4. If the regions share three or more edges, the network can be shortened.

Figure 2.5.4), removing the middle shared edge (and reorienting one segment if necessary) would decrease the total length of N without disconnecting it or increasing the number of nodes. The number of nodes remains constant since the removal of the edge leaves two nodal points with multiplicity one.

Therefore, as cases (i)-(iv) yield contradictions, R_1 and R_2 must be disjoint. □

THEOREM 2.6. *If N is a shortest directed network among those with at most n nodes, then N has at most one region containing no boundary points.*

Proof. Suppose N has two regions R_1 and R_2 containing no boundary points. Let $S_1, S_2, \ldots, S_m$ be the segments emanating from the vertices of R_1. By Lemma 2.5, R_1 and R_2 are disjoint from each other and from all other regions in N. Hence we may divide the space around R_1 into regions $\rho_1, \ldots, \rho_m$ by drawing dotted curves emanating from the boundary of R_1 between S_1 and S_2, S_2 and S_3, $\ldots, S_m$ and S_1, as far as we like, without ever intersecting a region having nodes or edges with R_1. (See Figure 2.6.1.) Similarly, the same holds for R_2. Now at least two of the regions around R_1 contain a starting point. Otherwise we could remove an edge from R_1 without disconnecting the network. Similarly, at least two regions contain a destination point. The same holds for R_2. We may assume, without loss of generality, that starting points O_1 and O_2 and destination points D_1 and D_2 lie in the regions shown in Figure 2.6.1.

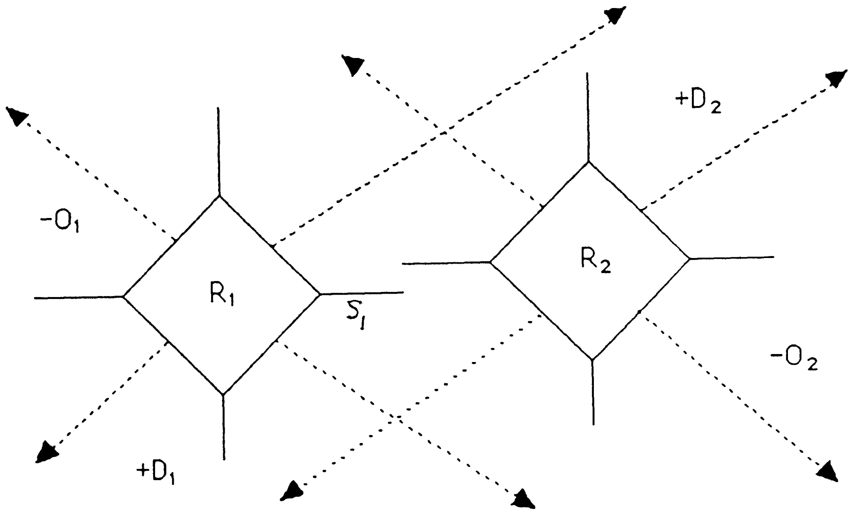

FIGURE 2.6.1. There cannot be two regions disjoint from the given boundary points.

Now, there exist paths P_1 from O_1 to D_2 and P_2 from O_2 to D_1. But R_2 lies inside, say, ρ_1. Hence P_1 must go out of S_1. Similarly, P_2 enters R_1 through S_1. This implies that S_1 is, for at least a bit, a double edge (an edge with both orientations). This yields a contradiction since we can now improve N by moving the nodal point out a bit (Lemma 2.2). This does not increase the number of nodes in N or disturb connectedness (since we do not remove any edges). Therefore shortest directed networks with at most n nodes have at most one region containing no boundary points. □

LEMMA 2.7. *Let N be a shortest directed network among those with at most n nodes. Then at most six edges can meet at a point in N.*

Proof. Suppose four or more edges enter a point B. (See Figure 2.7.1.) Then two of the edges must form a "V" with an angle less than 120 degrees. By Lemma 2.2, replacing this "V" with a "Y" (orienting all segments of "Y" with the same orientation as the segments of the "V" would decrease the length without increasing the number of nodes (since removing the "V" reduces the number of nodes by one, and adding the "Y" adds a node). (When counting the nodes use the node counting convention defined in the introduction

1.1.) It follows that at most three segments can enter a point. Similarly, at most three segments can leave a point. Therefore at most six segments can meet at a point in shortest directed networks. (See Figure 2.7.2.) □

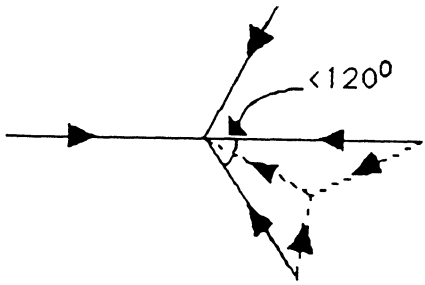

FIGURE 2.7.1. If four edges come into a point, the network can be shortened.

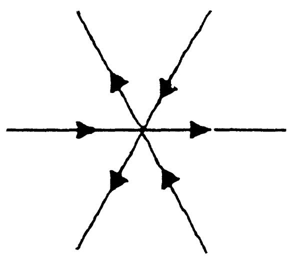

FIGURE 2.7.2. At most six edges meet at a point.

THEOREM 2.8. *Given b boundary points, there exists a shortest directed network connecting all origins to all destinations.*

Proof. Fix $n \geq 52b$ large enough to ensure there exists a network with at most n nodes connecting the given boundary points. By Theorem 2.1, there is a connected directed network N, shortest among those with at most n nodes, connecting the boundary points. We may assume there are no nodes where just two segments meet, since they would have to meet at 180 degrees and the node could be removed. By Theorem 2.5, N has at most one region containing no boundary points. By Lemma 2.7, each boundary point is in at most six regions. Hence, N has at most $6b+1$ regions. Removing one edge from each region yields a tree. A tree with b leaves has at most $b-2$ nodal points where 3 or more edges meet. In addition, where the $6b+1$ edges were removed there are at most $2(6b+1)$ nodal points where 2 edges meet. Therefore N has at most $b-2+2(6b+1) = 13b$

nodal points. Since by Lemma 2.7 at most six segments can meet at a point, each nodal point counts as at most four nodes (see 1.1). Therefore N has at most $52b$ nodes, a bound independent of the initial n. It follows that N is the desired shortest network. $\square$

Acknowledgements. This paper represents my honors thesis [**A1**], a continuation of work of the 1989 Williams SMALL undergraduate research project Geometry Group, consisting of Josh Sher (group leader), Tessa Campbell, Andres Soto, and me. It was edited by the 1990 Geometry Group, consisting of Joel Foisy, Nickelous Hodges, Jason Zimba, and Jeffrey Brock and me (group leaders). All of this research has been advised by Professor Frank Morgan.

In the Williams SMALL Undergraduate Research Project, for a period of nine or ten summer weeks, each of some twenty students works in two of the six groups comprising the Project. Support has been provided by the National Science Foundation (including site and add-on funds from the Research Experiences for Undergraduates Program), the Ford Foundation, the New England Consortium for Undergraduate Science Education, G.T.E., Shell, and Williams College (NSF REU site awarded to Morgan et al., NSF grants to C. Adams, D. Bergstrand, F. Morgan). [**Ad**] gives more information.

References

[Ab] J. Abrahamson, *Curves length-minimizing modulo ν in $\mathbb{R}^n$*, Michigan Mathematics Journal, **35** (1988), 285-290.

[Ad] C. Adams, D. Bergstrand and F. Morgan, *The Williams SMALL undergraduate research project*, UME Trends, January, 1991.

[A1] M. Alfaro, *Existence of shortest directed networks in $\mathbb{R}^2$*, Honors thesis, Williams College, 1990.

[A2] M. Alfaro, Tessa Campbell, Josh Sher, and Andres Soto, *Length-minimizing directed networks can meet in fours*, Williams College SMALL undergraduate research project, Geometry Group, 1989.

[A3] M. Alfaro, Mark Conger, Kenneth Hodges, Rajiv Kochar, Lisa Kuklinski, Adam Levy, Zia Mahmood, and Karen von Haam, *Segments can meet in fours in energy-minimizing networks*, Journal of Undergraduate Mathematics, **22** (1990), 9-20.

[A4] ———, *The structure of singularities in Φ-minimizing networks in $\mathbb{R}^2$*, Pacific Jour. Math., **149** (1991), 201-210.

[BG] M. W. Bern and R.L. Graham, *The shortest-network problem*, Scientific American, January, 1989, 84-89.

[CR] R. Courant and H. Robbins, *What is Mathematics?*, Oxford Univ. Press, 1941.

[L] A. Levy, *Energy-minimizing networks meet only in threes*, Journal of Undergraduate Mathematics, **22** (1990), 53-59.

[M1] F. Morgan, *Minimal surfaces, crystals, shortest networks, and undergraduate research*, Math. Intelligencer, Vol.14, Summer, 1992, 37-44.

[M2] ______, *Riemannian Geometry: a Beginner's Guide*, A.K. Peters, Wellesley, 1993.

Received 1991 and in revised form March 7, 1993.

c/o Professor Frank Morgan
Williams College
Williamstown, MA 01267
E-mail address: Frank.Morgan@williams.edu