SIAM J. NUMER. ANAL.
Vol. 10, No. 2, April 1973

NESTED DISSECTION OF A REGULAR FINITE ELEMENT MESH*

ALAN GEORGE†

Abstract. Let M be a mesh consisting of n^2 squares called *elements*, formed by subdividing the unit square $(0, 1) \times (0, 1)$ into n^2 small squares of side $1/h$, and having a *node* at each of the $(n + 1)^2$ grid points. With M we associate the $N \times N$ symmetric positive definite system $Ax = b$, where $N = (n + 1)^2$, each x_i is associated with a node of M, and $A_{ij} \neq 0$ if and only if x_i and x_j are associated with nodes of the same element. If we solve the equations via the standard symmetric factorization of A, then $O(n^4)$ arithmetic operations are required if the usual row by row (banded) numbering scheme is used, and the storage required is $O(n^3)$. In this paper we present an unusual numbering of the mesh (unknowns) and show that if we avoid operating on zeros, the LDL^T factorization of A can be computed using the same standard algorithm in $O(n^3)$ arithmetic operations. Furthermore, the storage required is only $O(n^2 \log_2 n)$. Finally, we prove that all orderings of the mesh must yield an operation count of at least $O(n^3)$, provided we use the standard factorization algorithm.

1. Introduction. It is well known that if we avoid operating on and storing zeros, the way we number or order the unknowns of a sparse system of equations can drastically affect the amount of computation and storage required for their direct solution. In this paper we consider this ordering problem for a finite element system of equations associated with a regular $n \times n$ mesh or grid. This sparse system has a very definite structure which we show can be exploited to considerable advantage. The problem we consider is special, but as we indicate in our concluding remarks, the techniques we use can be applied in more general situations.

We first define a finite element system of equations. Let M be any mesh formed by subdividing a planar region R with boundary ∂R by a number of arcs, all of which terminate on an arc or on ∂R. The mesh so formed consists of the union of subregions which we call *elements*. We require that M have a *node* at each vertex in the mesh, and it may have nodes on edges and in the interior of some or all of the elements. (A vertex of M is any point in $R \cup \partial R$ having more than two arcs emanating from it.) We refer to these nodes as *vertex*, *edge*, and *interior* nodes respectively. We call such a mesh a *finite element mesh.*

Let M have N nodes, numbered in some way from 1 to N. Associating an unknown x_i with the ith node leads us to the following.

DEFINITION. A *finite element system of equations* associated with the finite element mesh M is any $N \times N$ symmetric positive definite system $Ax = b$ having the property that $A_{ij} \neq 0$ if and only if x_i and x_j are associated with nodes of the same element of M.

The reader familiar with the use of finite element techniques may wonder why we allow M to be more general than the meshes usually employed in finite element methods. Usually M is the union of triangular and/or quadrilateral elements with adjacent elements having a common side or vertex. Our intention

* Received by the editors March 17, 1972, and in revised form July 25, 1972.

† This work was supported in part by a University of Waterloo research grant, and in part by Canadian National Research Council Grant A8111. Alan George received his Ph.D. in Computer Science in 1971 from Stanford University under the joint direction of Professor Forsythe and Dr. Fred Dorr. He is now Assistant Professor of Computer Science at the University of Waterloo, Waterloo, Ontario, Canada.

is to introduce a sequence of meshes which correspond (in the sense above) to a sequence of matrices which arise during the computation. These derived meshes have a less restricted topology than the original M.

Since there is a 1–1 correspondence between nodes and unknowns, we do not distinguish between them in the sequel.

In this paper we consider the ordering problem for the finite element system

$$Ax = b, \tag{1.1}$$

associated with the mesh M formed by subdividing the unit square $R = (0, 1) \times (0, 1)$ into n^2 small squares (elements) having side length $1/h$. The mesh has a node at each of the $N = (n + 1)^2$ vertices. The method we use to solve (1.1) is direct; we compute the symmetric factorization LDL^T of A, where L is unit lower triangular and D is a positive diagonal matrix. We then obtain x by solving $Ly = b$, $Dz = y$, and finally $L^Tx = z$.

Our measure of computational difficulty for the solution of (1.1) is θ, the number of multiplicative operations (multiplications and divisions) required to factor A. We regard this as a reasonable measure, since the required number of additions and subtractions is about the same, and the factorization is typically the major portion of the computation. In the following, "operations" will mean multiplicative operations. We measure storage requirements by η, the number of nonzero off-diagonal components in L.

Consider any matrix $A^{(5)}$ obtained from a matrix A in our class by setting A_{ij} to zero unless x_i and x_j are associated with nodes on the extremity of the same element edge. Such a matrix arises when we apply the usual 5-point difference operator in connection with solving self-adjoint elliptic boundary problems with rectangular domains [3]. Hoffman, Martin and Rose [6] show that symmetric positive definite matrices having the structure of $A^{(5)}$ require at least $O(n^3)$ operations for their factorization, and the corresponding lower triangular factor must have at least $O(n^2 \log_2 n)$ nonzero components. Since A is obtained from $A^{(5)}$ by adding nonzero components, it follows that these results also hold for A. It is important to appreciate that these results apply for the particular algorithm described in § 2, and, in this context, we conclude that the ordering we present here is optimal, in the order of magnitude sense.

Since we do not make explicit use of the actual numerical values of A, our upper bounds for θ and η hold regardless of whether or not unknowns associated with the same element are indeed connected. However, to avoid tedious qualifications in the discourse that follows, we assume that if unknowns x_i and x_j are associated with the same element, then $A_{ij} \neq 0$. If this is not the case, our upper bounds will simply not be sharp.

An ordering similar but somewhat inferior to the one we present here appears in [4]. The bound on θ we obtain here halves the one obtained in that article.

In § 2 and § 3 we introduce quantities and a model which allow us to conveniently determine θ and η for a given ordering (numbering) of M. In § 4 we present an ordering of the mesh which results in $\theta = O(n^3)$ and $\eta = O(n^2 \log_2 n)$, and in § 5 we show that if we apply the symmetric decomposition algorithm, *any* ordering of the mesh must result in $\theta = O(n^3)$.

2. Symmetric elimination. Using the outer product formulation employed by Rose [8], we describe the factorization of A into LDL^T by the following equations. Setting $A = A_0 = B_0$, we have

$$\begin{aligned} A_0 &= \begin{pmatrix} d_1 & v_1^T \\ v_1 & B_1' \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \frac{v_1}{d_1} & I_{N-1} \end{pmatrix} \begin{pmatrix} d_1 & 0 \\ 0 & B_1' - \frac{v_1 v_1^T}{d_1} \end{pmatrix} \begin{pmatrix} 1 & v_1^T/d_1 \\ 0 & I_{N-1} \end{pmatrix} \\ &= L_1 \begin{pmatrix} d_1 & 0 \\ 0 & B_1 \end{pmatrix} L_1^T = L_1 A_1 L_1^T, \end{aligned}$$

(2.1)

$$\begin{aligned} A_1 &= \begin{pmatrix} d_1 & & 0 \\ 0 & d_2 & v_2^T \\ & v_2 & B_2' \end{pmatrix} \\ &= \begin{pmatrix} 1 & & 0 \\ & 1 & \\ 0 & \frac{v_2}{d_2} & I_{N-2} \end{pmatrix} \begin{pmatrix} d_1 & & 0 \\ & d_2 & \\ 0 & & B_2' - \frac{v_2 v_2^T}{d_2} \end{pmatrix} \begin{pmatrix} 1 & & 0 \\ & 1 & v_2^T/d_2 \\ 0 & & I_{N-2} \end{pmatrix} \\ &= L_2 A_2 L_2^T, \\ &\ \vdots \\ A_{N-1} &= D. \end{aligned}$$

Here d_k is a positive scalar, v_k is a vector of length $N - k$, and B_k' is an $(N - k) \times (N - k)$ symmetric positive definite matrix. In the sequel $B_k = B_k' - v_k v_k^T/d_k$ is referred to as "the part of A remaining to be factored" after the first k steps of the factorization have been performed. Following Rose [8], we refer to performing the kth step of the factorization as "eliminating variable x_k."

Since A is sparse, the vectors v_k will usually also have some zeros. Define ν_k to be the number of nonzero components in v_k. Then we have the following lemma.

LEMMA 2.1 (Rose [8]). *Provided we avoid operating on zeros, the number of multiplicative operations required to factor A into LDL^T is*

$$\theta = \sum_{k=1}^{N-1} \frac{\nu_k(\nu_k + 3)}{2}, \tag{2.2}$$

and the number of nonzero off-diagonal components in L is

$$\eta = \sum_{i=1}^{N-1} \nu_k. \tag{2.3}$$

Proof. It is straightforward to verify from (2.1) that the cost θ_k of performing the kth step of the factorization is $\nu_k(\nu_k + 3)/2$. Summing over k and noting that $\nu_N \equiv 0$ yields (2.2).

Equation (2.3) can be derived by observing first that (2.1) implies that $A = L_1 L_2 \cdots L_{N-1} D L_{N-1}^T L_{N-2}^T \cdots L_1^T$, and it is easy to show that

$$L = \sum_{k=1}^{N-1} L_k - (N - 2)I.$$

Thus, the kth column of L is precisely the kth column of L_k, which immediately implies (2.3). This concludes the proof.

Un-eliminated variables x_j and x_k are referred to as being *connected* if their corresponding off-diagonal components in B_i $(i < j, k)$ are nonzero. Obviously, unconnected variables can become connected as the factorization proceeds. Specifically, if we assume that we do not create zeros through cancellation (which is a reasonable assumption in the presence of roundoff error), then the following holds.

LEMMA 2.2 (Parter [7]). *The elimination of variable x_k pairwise connects all variables x_i, $i > k$, to which x_k was connected at the point of its elimination.*

Proof. Referring to equations (2.1), we note that eliminating x_k modifies B_k' by subtracting the rank-one matrix $v_k v_k^T$ from it, forming B_k. The matrix $v_k v_k^T$ has nonzeros in position (i, j) for all i and j corresponding to nonzero components in v_k. Assuming no cancellation in the subtraction, B_k must have nonzeros in the same positions, proving the lemma.

Rose [8] refers to sets of unknowns which are pairwise connected (every unknown in the set is connected to every other unknown in the set) as *cliques.* Thus, eliminating an unknown x_k which is connected to a set of variables S renders that set a clique.

The following lemma is a direct application of Lemmas 2.1 and 2.2.

LEMMA 2.3. *Let $Q = \{x_{i_1}, x_{i_2}, \cdots, x_{i_q}\}$ and $R = \{x_{j_1}, x_{j_2}, \cdots, x_{j_r}\}$ be two sets of unknowns, with unknowns $x_k \in Q$ connected to every unknown $x_l \in Q \cup R$ and no others. Then if $i_k < j_l$ for $1 \leqq k \leqq q$ and $1 \leqq l \leqq r$, the contribution to θ from elimination of the unknowns in Q is*

$$m(q, r) = m_1(q) + m_2(q, r), \tag{2.4}$$

where

$$m_1(q) = \frac{q^3}{6} + \frac{q^2}{2} - \frac{2}{3}q \tag{2.5}$$

and

$$m_2(q, r) = \frac{qr}{2}(q + r + 2). \tag{2.6}$$

The storage required for columns $i_1, i_2, \cdots, i_q$ of L is

$$s(q, r) = \tfrac{1}{2}q(q + 2r - 1). \tag{2.7}$$

Proof. Since we assume variables in Q are connected only among themselves and to those in R, we can without loss of generality assume $i_k = k$ and $j_k = q + k$. We then need only consider the cost of performing the first q steps of the factorization of a $(q + r) \times (q + r)$ matrix whose first q columns are *dense.* This implies $v_k = q + r - k$, $k = 1, 2, \cdots, q$, and using (2.2), we have

$$\begin{aligned} m(q, r) &= \sum_{k=1}^{q} \tfrac{1}{2}(q + r - k)(q + r - k + 3) \\ &= \sum_{k=1}^{q} \tfrac{1}{2}(q - k)(q - k + 3) + \sum_{k=1}^{q} \tfrac{1}{2}r(r + 2q - 2k + 3) \\ &= m_1(q) + m_2(q, r). \end{aligned}$$

Similarly, using (2.3) along with the above formula for v_k, we have

$$s(q,r) = \sum_{k=1}^{q} (q + r - k) = \tfrac{1}{2}q(q + 2r - 1).$$

3. A mesh model for the analysis of the factorization. In the Introduction we *defined* the zero-nonzero structure of finite element systems of equations in terms of a planar mesh. We now establish a correspondence between elimination of certain sets of unknowns in (1.1) and corresponding changes in M. Denoting M by M_0, we arrange that after the kth step of the factorization, we have a mesh M_k which corresponds to B_k in the sense that unknowns still to be eliminated are connected only if they are associated with the same element of M_k. In other words, *B_k is a finite element matrix corresponding to M_k.*

For example, consider the mesh M_0 depicted in Fig. 3.1, consisting of rectangular elements, and numbered as indicated. Unnumbered nodes are assumed to have numbers greater than 1.

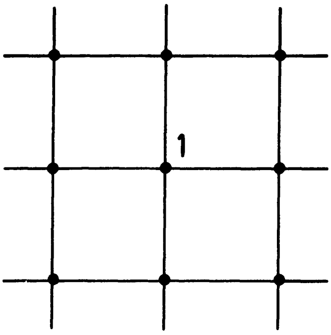


FIG. 3.1. *The mesh M_0*

By our definition of finite element matrices, x_1 is connected to all of the 8 other unknowns which are associated with the same elements as x_1. Keeping in mind that *only* unknowns associated with the same element may be connected, and recalling Lemma 2.2, the mesh M_1 which reflects the structure of B_1 must have the 8 unknowns mentioned above all associated with the same element. Thus, M_1 must be as shown in Fig. 3.2, where a new element has been formed by merging four elements of M.

Consider a second mesh example (Fig. 3.3), which has some edge nodes as well as vertex nodes. Here we assume unnumbered nodes have numbers greater than 2. Which mesh M_1 correctly reflects the structure of B_1, the part of the matrix remaining after the first step of the factorization is complete? By the same arguments as before, the 8 unknowns x_k, $k > 2$, associated with the two elements sharing nodes 1 and 2 must be associated with the same element in M_1. Thus, these two adjacent elements must coalesce. Node (unknown) 2 is already connected to all unknowns of the two elements, and eliminating x_1 does not connect it to any others. Thus, B_1's structure is correctly described by the mesh in Fig. 3.4, where node 2 is now interior to the newly formed element. Since elimination of

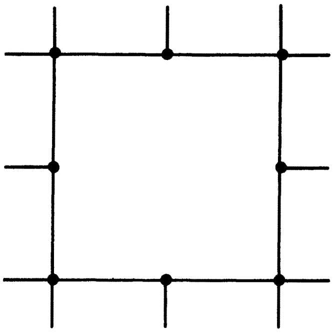


FIG. 3.2. *Derived mesh M_1*

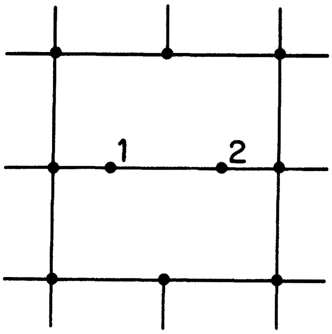


FIG. 3.3. *Mesh M_0*

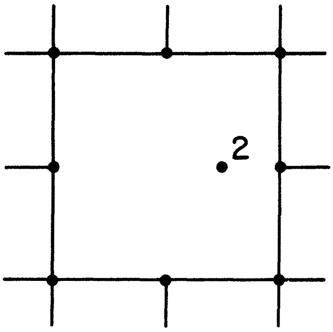


FIG. 3.4. *Mesh M_1 derived from the mesh M_0 of Fig.* 3.3

variable x_2 only involves variables to which x_2 is already connected, the mesh M_2 derived from M_1 and corresponding to B_2 would simply be the mesh of Fig. 3.4 with node 2 removed.

Finally, to completely fix these ideas in mind, consider the mesh M_0 depicted in Fig. 3.5. Unnumbered nodes are assumed to have numbers greater than 9. Using our examples developed above, the reader should now be able to verify

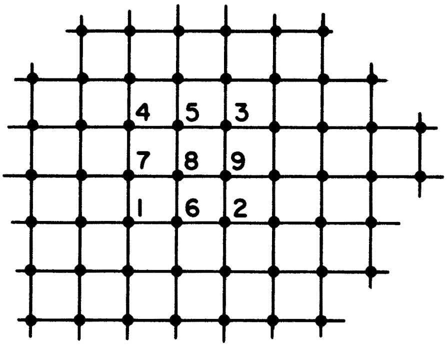
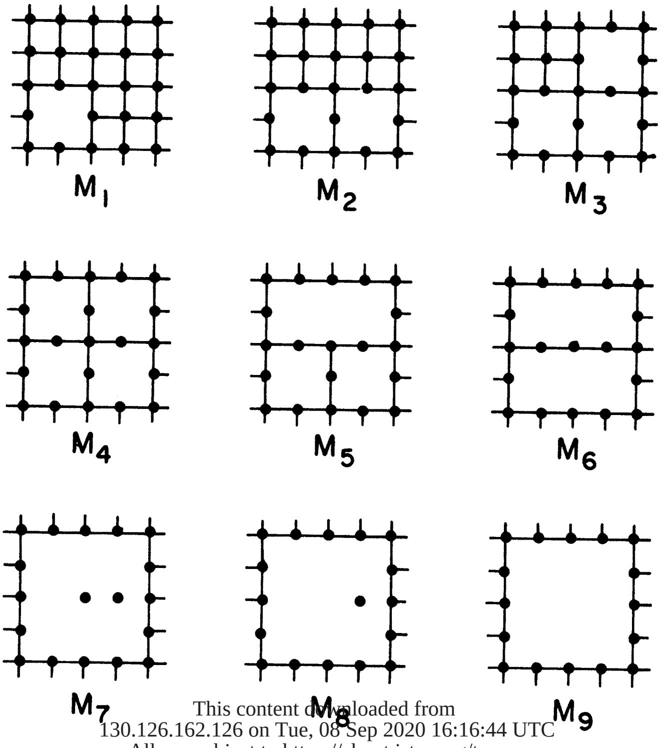


FIG. 3.5. *Mesh* M_0

that the meshes $M_1, M_2, \cdots, M_9$ shown in Fig. 3.6 correspond to $B_1, B_2, \cdots, B_9$ respectively.

Summarizing, we have developed the following *mesh transformation rules.*

Rule 1. The elimination of a variable associated with an interior node corresponds to removal of that node from the mesh; for example, the transformation from M_7 to M_8 in Fig. 3.6.



Rule 2. The elimination of a variable associated with an edge node on an edge e corresponds to removal of the node and edge from the mesh, but other nodes on e remain, and become interior nodes of a new element; for example, the transformation of M_6 to M_7 in Fig. 3.6.

Rule 3. The elimination of a variable associated with a vertex node corresponds to removal of the vertex node and all its incident element edges. Other nodes lying on these edges remain, and become interior nodes of a newly formed element.

In order to preserve our element structure when eliminating nodes on the boundary, we must interpret our rules somewhat carefully. First observe that as far as the matrix structure is concerned, the two meshes in Fig. 3.7 are equivalent.

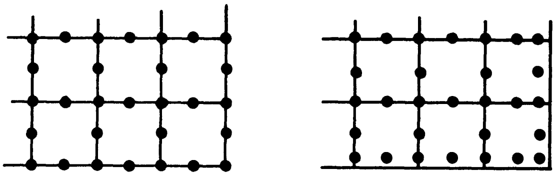


FIG. 3.7. *Two meshes having equivalent matrix structure*

When eliminating unknowns associated with nodes on the boundary, we shall assume that edge nodes have been moved to the interior of an element and vertex nodes have been moved onto an edge as indicated in Fig. 3.7. We then apply our rules as before to this *perturbed mesh*, which has no nodes on ∂R. An example of the applications of the rules in this situation is given in Fig. 3.8.

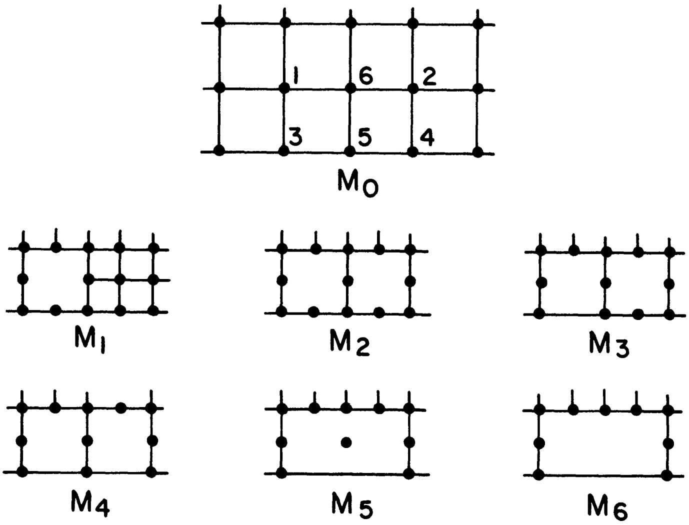


FIG. 3.8. *Modification of the mesh when variables associated with boundary nodes are eliminated*

Thus, using this model, every numbering of the mesh determines a sequence of finite element meshes which reflects the changing structure of the part of the matrix remaining to be factored. The last member in the sequence $M_0, M_1, \cdots, M_N$ consists of a single boundary element whose boundary is ∂R, with all nodes removed.

The reader familiar with the graph theory model of elimination extensively employed by Rose [8] for analyzing sparse matrix computations will recognize the close relationship between our model and the graph model. By simply joining every pair of nodes associated with the same element of M, we obtain the (undirected) graph corresponding to A. Doing the same for our $M_k, k = 1, 2, \cdots, N$, yields a sequence of *elimination graphs*. Elements correspond to *cliques*. Thus, our model can be viewed as a graph theory model where many of the edges of the graph are not explicitly drawn, but are made implicit through the introduction of elements. For our particular application, our model has the advantage of being somewhat easier to draw and interpret. For less structured problems and in other contexts, the graph model is more appropriate.

4. A nested dissection ordering of *M* and some crude bounds. In this section we describe an ordering of our $n \times n$ mesh M which yields $\theta = O(n^3)$ and $\eta = O(n^2 \log_2 n)$. To simplify the presentation, we do not attempt to find the constants involved; the interested reader will find a careful analysis in the Appendix.

We begin by defining some sets of unknowns (nodes), where x_{ij} denotes the unknown associated with node (ih, jh). We assume initially that $n = 2^l$. For $i = 1, \cdots, n - 1$ define

$$\pi(i) = p + 1 \quad \text{if } i = 2^p(2q + 1).$$

Furthermore, define $\pi(0) = 1$ and $\pi(n) = 1$. For example, when $n = 16$, the values of $\pi(i), i = 0, \cdots, 16$, are given in the first row of Fig. 4.1. For $k = 1, \cdots, l$ define the set of nodes P_k by

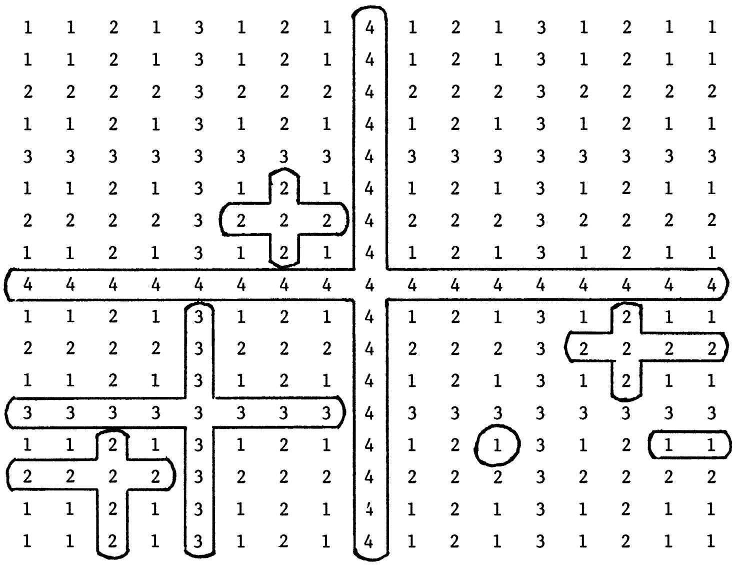
$$P_k = \{x_{ij} | \max(\pi(i), \pi(j)) = k\}. \tag{4.1}$$

Denoting membership in P_k by the number k, these node sets are depicted in Fig. 4.1 for $n = 2^4 = 16$. To aid the description we have put lines around P_4 and a few subsets of P_1, P_2 and P_3.

Notice that P_4 subdivides the nodes (unknowns) into 4 subsets which are mutually independent in the sense that if x_i and x_j are in different subsets, then $A_{ij} = 0$. In the same way, P_3 subdivides each of these subsets into 4 mutually independent subsets, and so on. Hence the name "nested dissection."[1] Each of the P_i themselves consists of independent sets of nodes which increase in size with i. For $i > 1$ these subsets are "+" shaped. The subsets of each P_i might be appropriately named "separating sets."

Now our overall strategy is to number the unknowns in P_1, followed by those in P_2 and so on, finally numbering the unknowns in P_l. The results we establish in this section are independent of the way each P_k is numbered.

[1] The author is indebted to Professor Garret Birkhoff of Harvard University for coining this very appropriate term.



1	1	2	1	3	1	2	1	4	1	2	1	3	1	2	1	1
1	1	2	1	3	1	2	1	4	1	2	1	3	1	2	1	1
2	2	2	2	3	2	2	2	4	2	2	2	3	2	2	2	2
1	1	2	1	3	1	2	1	4	1	2	1	3	1	2	1	1
3	3	3	3	3	3	3	3	4	3	3	3	3	3	3	3	3
1	1	2	1	3	1	2	1	4	1	2	1	3	1	2	1	1
2	2	2	2	3	2	2	2	4	2	2	2	3	2	2	2	2
1	1	2	1	3	1	2	1	4	1	2	1	3	1	2	1	1
4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
1	1	2	1	3	1	2	1	4	1	2	1	3	1	2	1	1
2	2	2	2	3	2	2	2	4	2	2	2	3	2	2	2	2
1	1	2	1	3	1	2	1	4	1	2	1	3	1	2	1	1
3	3	3	3	3	3	3	3	4	3	3	3	3	3	3	3	3
1	1	2	1	3	1	2	1	4	1	2	1	3	1	2	1	1
2	2	2	2	3	2	2	2	4	2	2	2	3	2	2	2	2
1	1	2	1	3	1	2	1	4	1	2	1	3	1	2	1	1
1	1	2	1	3	1	2	1	4	1	2	1	3	1	2	1	1

FIG. 4.1. *The node sets P_k, $k = 1, 2, 3, 4$*

Using our example above ($n = 16$), and the model we introduced in § 3, it is not difficult to see that the meshes M_{100}, M_{196}, and M_{256} are as indicated in Figs. 4.2, 4.3 and 4.4 respectively. They correspond to the structure of the matrix remaining to be factored after unknowns in P_1, $P_1 \cup P_2$, and $P_1 \cup P_2 \cup P_3$ respectively, have been eliminated.

THEOREM 4.1. *Let $n = 2^l$ and define the node sets P_k, $k = 1, 2, \cdots, l$, by* (4.1). *Number the nodes in increasing order beginning with those in P_1, followed by those in P_2, etc., finally numbering those in P_l. Then there exist constants C_1 and C_2*

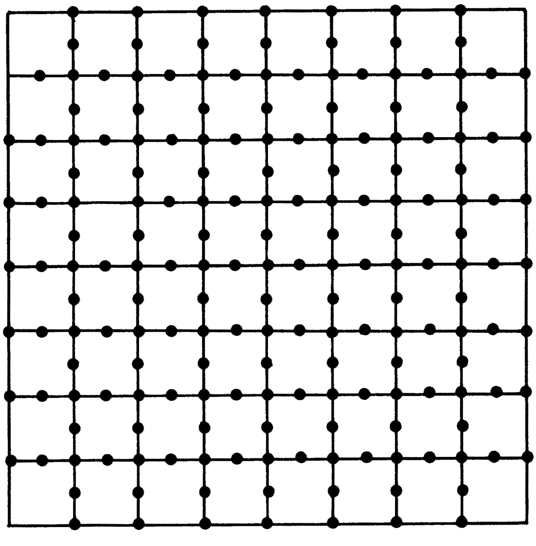


FIG. 4.2. *The mesh M_{100} formed by elimination of unknowns in P_1*

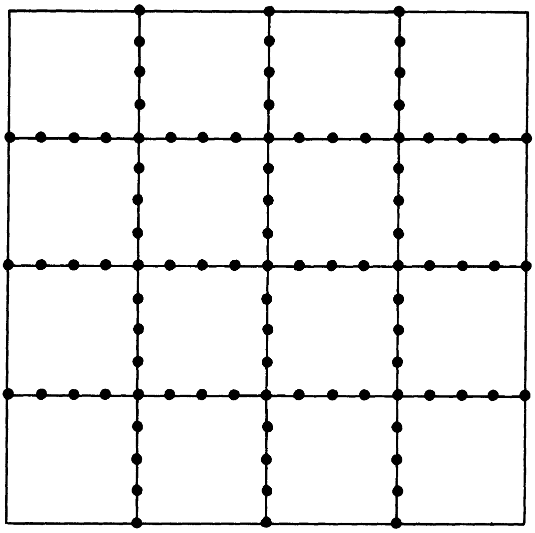


FIG. 4.3. *The mesh M_{196} formed by eliminating unknowns in P_1 and P_2*

such that

$$\theta < C_1 n^3 \tag{4.2}$$

and

$$\eta < C_2 n^2 \log_2 n. \tag{4.3}$$

Proof. First observe that P_k consists of $n^2/2^{2k}$ independent sets of unknowns, and note that they *remain independent during the elimination.* That is, unknowns in different subsets of P_k never become connected during the elimination. Each independent set has no more than 2^{k+1} unknowns in it, and each unknown in the

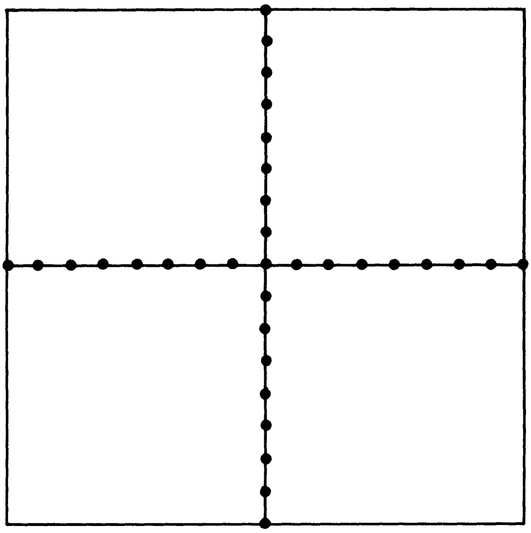


FIG. 4.4. *The mesh M_{256} formed by eliminating unknowns in P_1, P_2 and P_3*

set is connected to no more than $6 \cdot 2^k - 3$ unknowns at the point of its elimination. For example, unknowns in P_2 are connected to at most 20 unknowns before they are eliminated. Thus, using (2.2), we have

$$\begin{aligned}\theta &< \sum_{k=1}^{l} \left(\frac{n^2}{2^{2k}}\right) 2^{k+1}(6 \cdot 2^k)^2 \\ &= C_1' n^2 \sum_{k=1}^{l} 2^k \leqq C_1 n^3.\end{aligned}$$

Similarly, using (2.3), we have

$$\begin{aligned}\eta &< \sum_{k=1}^{l} \frac{n^2}{2^{2k}} \cdot 2^{k+1} \cdot 6 \cdot 2^k = C_2 n^2 \sum_{k=1}^{l} 1 \\ &= C_2 n^2 \log_2 n.\end{aligned} \tag{4.4}$$

COROLLARY 4.2. *For any $n > 2$, there exist constants C_3 and C_4 such that*

$$\theta < C_3 n^3 \tag{4.5}$$

and

$$\eta < C_4 n^2 \log_2 n. \tag{4.6}$$

Proof. Let $2^{l-1} < n < 2^l = \bar{n}$, let M and $\bar{M}$ be meshes corresponding to n and $\bar{n}$, and let A correspond to M. Now augment our system of equations (1.1) by adding trivial equations of the form $1 \cdot x_i = 1$, $i > (n + 1)^2$ corresponding to nodes of $\bar{M}$ that are not nodes of M, so that the dimension of our expanded coefficient matrix is $(\bar{n} + 1)^2$. Now number $\bar{M}$ as in Theorem 4.1 and solve our expanded system. A subvector of the solution will be the solution of the original problem. Let $\alpha = \bar{n}/n < 2$. Then by Theorem 4.1,

$$\theta < C_1 \bar{n}^3 = C_1 \alpha^3 n^3 < 8 C_1 n^3 = C_3 n^3$$

and

$$\begin{aligned}\eta &< C_2 \bar{n}^2 \log_2 \bar{n} < C_2 \alpha^2 n^2 (\log_2 n + \log_2 \alpha) \\ &< C_2 \alpha^2 n^2 (\log_2 n + 1) < 8 C_2 n^2 \log_2 n \\ &= C_4 n^2 \log_2 n.\end{aligned}$$

As one might expect, using this crude analysis yields constants C_1, C_2, C_3 and C_4 which are, from a practical viewpoint, discouragingly large. However, in the Appendix we show that if $n = 2^l$, then $\theta < 10 n^3$ and $\eta < 8 ln^2$.

5. A lower bound for θ. In this section we show that for the algorithm described in § 2, any order of M must lead to $\theta = O(n^3)$ or greater. Following a suggestion of D. J. Rose, our strategy is to show that some member M_i of the mesh sequence M_k, $k = 0, 1, 2, \cdots, N$, must contain an element T having $n + 1$ nodes associated with it. This implies that B_i must have a dense $n + 1$ submatrix in it.

LEMMA 5.1. *Let $M = M_0$ be the regular $n \times n$ finite element mesh described in the Introduction, and let M_k, $k = 1, 2, \cdots, N = (n + 1)^2$, be the mesh sequence generated by an arbitrary ordering of M.*

Then at least one element T having $n + 1$ or more unknowns associated with it appears in the mesh sequence.

Proof. Let $Q = (x_i, y_i)$ be the first node of M to be removed which completely vacates a row or column of the mesh. Since all nodes are eventually removed, this situation must arise.

Suppose the removal of Q vacates exactly one row $\{\text{column}\}$ L. Then L is contained in an element T, since by Rule 2, § 3, removal of the nodes on L removes their incident edges. Moreover, the $n + 1$ columns $\{\text{rows}\}$ orthogonal to L necessarily have one or more nodes remaining on them, so proceeding along them in one direction or the other from L we must collide with a node lying on the boundary of T, proving the theorem.

Suppose the removal of Q simultaneously vacates both row R and column C of M. Since all other nodes on R and C have been removed, Q must necessarily be an interior node of an element T containing both R and C. By the same argument as above, T is an element having one or more nodes in each of the $n + 1$ rows and columns, again proving the theorem.

THEOREM 5.2. *The number of multiplicative operations required to complete the symmetric factorization of A is greater than $n^3/6$, regardless of the way the equations are numbered.*

Proof. By Lemma 5.1, during the decomposition we necessarily create one or more elements T_i having at least $n + 1$ unknowns associated with it, and they are all connected by virtue of belonging to T_i. Thus, B_i has a dense $(n + 1) \times (n + 1)$ submatrix whose symmetric factorization alone using the algorithm of § 2 requires $n^3/6 + n^2/2 - 2n/3$ multiplicative operations.

COROLLARY 5.3. *Every ordering of M results in a bandwidth $m = \max_{A_{ij} \neq 0} |i - j|$ satisfying $m \geqq n$.*

Proof. The B_i of Theorem 5.2 has a dense $(n + 1) \times (n + 1)$ submatrix, which means the bandwidth of B_i is at least n, which in turn implies the bandwidth of A is at least n.

Corollary 5.3 indicates how unreliable bandwidth can be as a measure of computational complexity, since computation estimates based on bandwidth would suggest θ must be at least $O(n^4)$. It is interesting in this connection to observe that for our ordering, $m \simeq n^2$!

6. Concluding remarks. We have presented an ordering of a system of $N = (n + 1)^2$ equations $Ax = b$ derived from a regular $n \times n$ mesh and have shown that using the given ordering, A can always be factored in $O(n^3)$ multiplicative operations, and the number of nonzero components in L is $O(n^2 \log_2 n)$. These results, combined with lower bounds of the same form due to Hoffman, Martin and Rose [6], leads us to conclude that our ordering is optimal in the order of magnitude sense (provided we use the algorithm described in § 2) with respect to both θ and η.

The class of matrices we have studied includes some well-known special cases, notably those which arise when the usual five-point or nine-point difference operator [3] is applied in connection with solving self-adjoint elliptic boundary value problems with rectangular domains. In certain special circumstances it is well known that systems with coefficient matrices having the structure of A can be solved in $O(n^3)$ or even $O(n^2 \log_2 n)$ operations [1], but these algorithms are special in the sense that they exploit the actual values of the components of A.

However, if the matrix has no special characteristics other than being symmetric, positive definite, and having the zero-nonzero structure corresponding to M, then it has been assumed that factorization of A required $O(n^4)$ operations and $O(n^3)$ storage locations. (That is, the row by row ordering was used.) Since it can be shown that some iterative schemes, under suitable circumstances, will reduce the error in the solution of (1.1) below a specified tolerance in $O(n^3)$ operations, it is often argued that iterative schemes are more efficient than direct methods for solving these sparse systems [2], [9]. We feel that Theorem 4.1 indicates that a re-comparison of direct and iterative methods is due. Of course, iterative schemes in general require much less storage than direct methods, even when we use our ordering, but the development of large memories and fast peripheral storage devices makes storage a less important factor than it has been in the past.

The extension of our upper bound results to the regular unit cube mesh containing n^3 cuboid elements is straightforward, although a little tedious. One now numbers sets of independent unknowns enclosed in increasingly large cubes rather than squares, finally numbering three intersecting planes of nodes, each pair having a common line of nodes, and the three having one common vertex at the centroid of the mesh. For this ordering of the mesh, $\theta = O(n^6)$ and $\eta = O(n^4)$. We have been unable to extend our lower bound proof to the three-dimensional problem.

Another straightforward generalization of our results is the case where M has nodes on edges and in the interior of each element, and more than one unknown is associated with each node. The results are essentially unchanged, except for some adjustments in constants. See [4] for some results in this direction, for a slightly different ordering than the one presented here.

A closely related matrix problem, arising in connection with the use of splines, has the property that grid points (unknowns) p and q are connected provided that the maximum difference in their x- and y-cordinates is bounded by some number d (which depends on the degree of the spline). We must now define our P_k's to consist of strips of horizontal and vertical grid lines, each strip consisting of d parallel grid lines. In this way nodes on opposite sides of a strip cannot be connected. We then proceed as before, obtaining $\theta = O(n^3)$ and $\eta = O(n^2 \log_2 n)$.

Finally, it should be obvious that our results apply even if A is unsymmetric, provided that its zero-nonzero structure is symmetric and we do not have to pivot during the decomposition to maintain numerical stability. Such matrix problems arise in connection with the use of finite element methods for solving non-self-adjoint elliptic boundary value problems. Depending on the way the problems are formulated, the matrix may or may not be symmetric, but the matrices always have symmetric structure. Furthermore, they are often diagonally dominant, which implies that pivoting for numerical stability is not required [10].

Appendix. We now decribe in detail a particular ordering using the general strategy described in § 4. We assume $n = 2^l$, and obtain sharp bounds for θ and η.

Our numbering strategy is most easily described in terms of the mesh sequence M_i, $i = 1, 2, \cdots, N$. Recall that numbering the first k nodes corresponds to carrying out the first k steps of the elimination, which leaves us with a matrix B_k remaining to be factored which has structure represented by M_k. Within each set P_i, our numbering (elimination) strategy is *to number* (*eliminate*) *the node* (*unknown*) *in* M_k (B_k) *which is connected to the fewest nodes* (*unknowns*). When this strategy is used solely, it is known as the *minimum degree algorithm*. Here its application is restricted to the sequence of sets $P_1, P_2, \cdots, P_l$. Since the subsets of $P_i, 0 < i \leqq l$, are independent, their relative numbering is immaterial. Therefore, we can apply the above strategy to each subset of P_i in turn, or apply it globally to the entire set P_i. In the example in Fig. A.1, we take the former course.

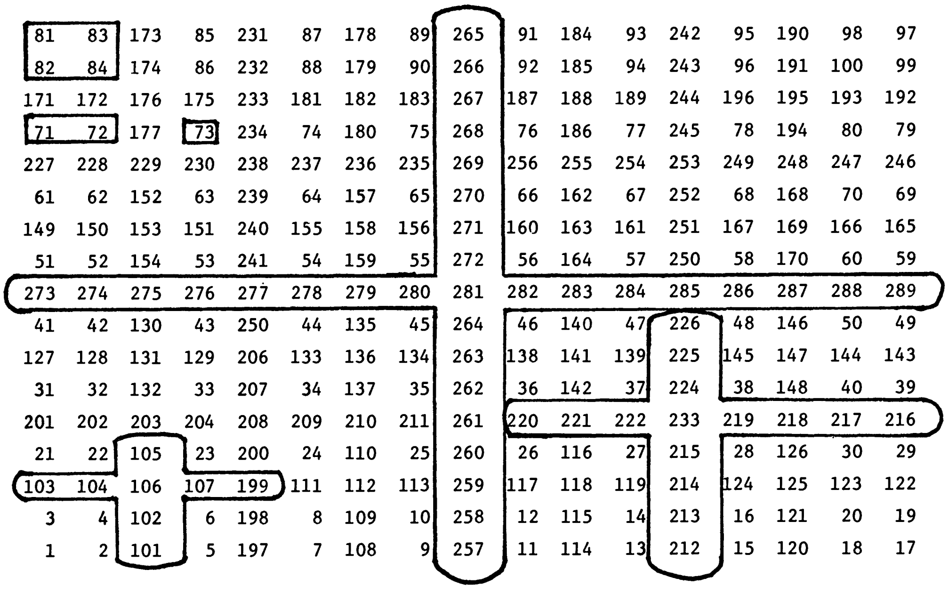
It is convenient at this point to define a set $P_0 = \{x_{ij} | i, j = 0, n\}$, and redefine P_1 to be $P_1 - P_0$. Then each subset of P_k, $0 < k \leqq l$, consists of a "+" shaped set of vertices, which we refer to as a *hub* with four *spokes*. (For $k = 1$, most of the spokes are null.)

Now the application of the minimum degree strategy described above to each subset leads to a numbering best described by the diagrams in Figs. A.2, A.3 and A.4, where the hashed lines indicate the boundary ∂R. The actual relative numbering on each edge is immaterial. The value of p is undetermined and not important here.

Obviously, the number of nodes on the edges in Figs. A.2–A.4 depend on k. To distinguish between the three types of subsets, we refer to those depicted in Figs. A.2–A.4 as *interior*, *boundary*, and *corner* subsets respectively.

We now make repeated use of Lemma 2.3 to compute θ and η.

LEMMA A.1. *The contribution to* θ *from the elimination of the unknowns in* P_0 *is* 36, *and the number of nonzeros in the first four columns of L is* 12.



81	83	173	85	231	87	178	89	265	91	184	93	242	95	190	98	97
82	84	174	86	232	88	179	90	266	92	185	94	243	96	191	100	99
171	172	176	175	233	181	182	183	267	187	188	189	244	196	195	193	192
71	72	177	73	234	74	180	75	268	76	186	77	245	78	194	80	79
227	228	229	230	238	237	236	235	269	256	255	254	253	249	248	247	246
61	62	152	63	239	64	157	65	270	66	162	67	252	68	168	70	69
149	150	153	151	240	155	158	156	271	160	163	161	251	167	169	166	165
51	52	154	53	241	54	159	55	272	56	164	57	250	58	170	60	59
273	274	275	276	277	278	279	280	281	282	283	284	285	286	287	288	289
41	42	130	43	250	44	135	45	264	46	140	47	226	48	146	50	49
127	128	131	129	206	133	136	134	263	138	141	139	225	145	147	144	143
31	32	132	33	207	34	137	35	262	36	142	37	224	38	148	40	39
201	202	203	204	208	209	210	211	261	220	221	222	233	219	218	217	216
21	22	105	23	200	24	110	25	260	26	116	27	215	28	126	30	29
103	104	106	107	199	111	112	113	259	117	118	119	214	124	125	123	122
3	4	102	6	198	8	109	10	258	12	115	14	213	16	121	20	19
1	2	101	5	197	7	108	9	257	11	114	13	212	15	120	18	17

FIG. A.1. *Detailed numbering of M for n* = 16. *The set* P_4 *and some subsets of* P_1, P_2 *and* P_3 *have been outlined.*

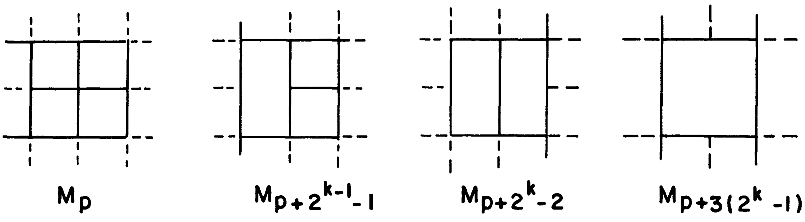


FIG. A.2. *Mesh sequence corresponding to elimination of unknowns in a subset of P_k which has no nodes on ∂R*

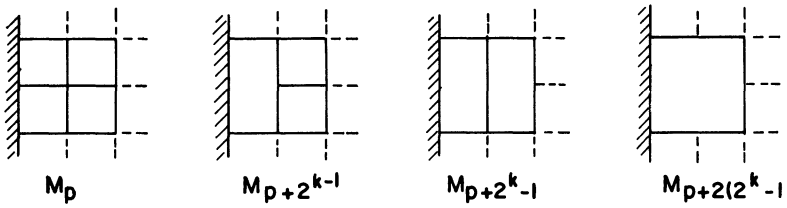


FIG. A.3. *Mesh sequence corresponding to elimination of unknowns of a subset of P_k which has one node on ∂R*

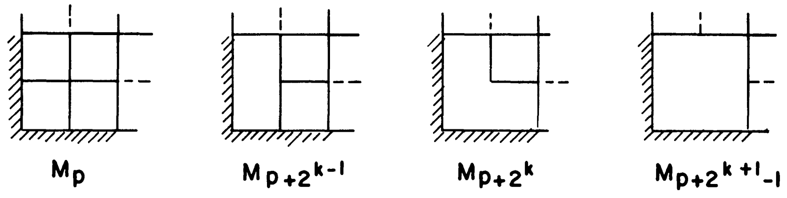


FIG. A.4. *Mesh sequence corresponding to elimination of unknowns of a subset of P_k having two nodes on ∂R*

Proof. Obviously each corner node is connected to only 3 other unknowns, which means $q = 1$ and $r = 3$ in Lemma 2.3. Thus, the contribution to θ is $4 \cdot m(1, 3) = 36$, and the contribution to η is $4 \cdot s(1, 3) = 12$.

LEMMA A.2. *The contribution to θ from the elimination of interior subsets in P_k, $0 < k < l - 1$, is*

$$\theta_I^k = \left(\frac{n}{2^k} - 2\right)^2 \left(\frac{371}{24} \cdot 2^{3k} - 17 \cdot 2^{2k} - \frac{22}{3} \cdot 2^k + 3\right) \tag{A.1}$$

and the number of nonzeros in the corresponding columns of L is

$$\eta_I^k = \left(\frac{n}{2^k} - 2\right)^2 \left(\frac{31 \cdot 2^{2k}}{4} - 13 \cdot 2^k + 3\right). \tag{A.2}$$

Proof. First observe that there are $(n/2^k - 2)^2$ such interior subsets in P_k, $0 < k < l - 1$. The sets P_0, P_{l-1} and P_l have none. The first two spokes eliminated contribute $m(2^{k-1}, 3\cdot 2^k)$, since each spoke has 2^{k-1} nodes on it, and the number of nodes not on the spoke but connected to unknowns on the spoke is $3\cdot 2^k$. Elimination of the final two spokes and hub contributes $m(2^k - 1, 2^{k+2})$.

Similarly, the storage required for the corresponding columns of L is $2\cdot s(2^{k-1}, 3\cdot 2^k) + s(2^k - 1, 2^{k+2})$. Expanding and simplifying these functions yields (A.1) and (A.2).

LEMMA A.3. *The contribution to θ from the elimination of unknowns in boundary subsets of P_k, $0 < k < l - 1$, is*

$$\theta_B^k = 4\left(\frac{n}{2^k} - 2\right)\left(\frac{121}{12}\cdot 2^{3k} - \frac{13}{14}\cdot 2^{2k} - \frac{41}{6}\cdot 2^k + 1\right), \tag{A.3}$$

and the number of nonzeros in the corresponding columns of L is

$$\eta_B^k = 4\left(\frac{n}{2^k} - 2\right)\left(\frac{25}{4}\cdot 2^{2k} - 7\cdot 2^k + 1\right). \tag{A.4}$$

Proof. Elimination of variables on the boundary spoke and the ones on the opposite spoke requires $m(2^{k-1}, 2^{k+1} + 1)$ and $m(2^{k-1} - 1, 3\cdot 2^k)$ operations respectively. The remaining variables require $m(2^k - 1, 3\cdot 2^k + 1)$ operations. Using this, along with the fact that there are $4(n/2^k - 2)$ boundary subsets in P_k, yields (A.3).

Similarly, the number of nonzeros in the corresponding columns of L for each subset is

$$s(2^{k-1}, 2^{k+1} + 1) + s(2^{k-1} - 1, 3\cdot 2^k) + s(2^k - 1, 3\cdot 2^k + 1).$$

Expanding, simplifying and multiplying by $4(n/2^k - 2)$ yields (A.4). Note that P_0, P_{l-1} and P_l have no boundary subsets.

LEMMA A.4. *The contribution to θ from the elimination of variables in the four corner subsets of P_k, $0 < k < l$, is given by*

$$\theta_C^k = \frac{125}{6}\cdot 2^{3k} + 18\cdot 2^{2k} - \frac{34}{3}\cdot 2^k, \tag{A.5}$$

and the number of nonzero components in the corresponding columns of L is

$$\eta_C^k = 18\cdot 2^{2k} - 8\cdot 2^k. \tag{A.6}$$

Proof. Elimination of variables of the first spoke contribute

$$m(2^{k-1}, 3\cdot 2^{k-1} + 1)$$

and $s(2^{k-1}, 3\cdot 2^{k-1} + 1)$ to θ and η respectively. The second spoke contributes $m(2^{k-1}, 2^{k+1} + 1)$ and $s(2^{k-1}, 2^{k+1} + 1)$ to θ and η respectively. Finally, the remaining 2^{k-1} unknowns contribute $m(2^k - 1, 2^{k+1} + 1)$ operations and $s(2^k - 1, 2^{k+1} + 1)$ nonzeros respectively. The set P_l has no corner subsets.

LEMMA A.5. *The number of arithmetic operations required to eliminate the unknowns in P_l is*

$$\theta^l = \tfrac{23}{24}n^3 + \tfrac{7}{2}n^2 + \tfrac{5}{8}n, \tag{A.7}$$

and the number of nonzeros in the corresponding columns of L is

$$\eta^l = \tfrac{7}{4}n^2 + n. \tag{A.8}$$

Proof. Elimination of the first two spokes requires $2 \cdot m(n/2, n + 1)$ operations, and the final $n + 1$ unknowns require $m(n + 1, 0)$ operations. Similarly, the number of nonzeros is given by $2 \cdot s(n/2, n + 1) + s(n + 1, 0)$.

THEOREM A.6. *The number of multiplicative operations required to factor the finite element matrix A associated with a regular $n \times n$ mesh, numbered as described above, and with $n = 2^l$, is given by*

$$\theta = \tfrac{267}{28}n^3 - 17n^2 \log_2 n + \tfrac{847}{28}n^2 + O(n \log_2 n), \tag{A.9}$$

and the number of nonzero components in L is

$$\eta = \tfrac{93}{12}n^2 \log_2 n - \tfrac{73}{3}n^2 + 24n \log_2 n + O(n). \tag{A.10}$$

Proof. The proof consists of merely summing the quantities given in Lemmas A.1–A.5 over the appropriate ranges. Normally, the algebra would be extremely tedious, but fortunately the author had access to the symbolic algebra system ALTRAN [5], so the quantities were summed by machine. The expansions of the m's and s's in the above lemmas were also checked by machine.

Briefly, we compute

$$\sum_{k=1}^{l-1} \theta_C^k + \sum_{k=1}^{l-2} (\theta_I^k + \theta_B^k),$$

and then add θ^l for P_l and 36 for P_0 yielding (A.9).

Similarly, to obtain η we compute

$$\sum_{k=1}^{l-1} \eta_C^k + \sum_{k=1}^{l-2} (\eta_I^k - \eta_B^k)$$

and then add η^l for P_l and 12 for P_0.

Table A.1 compares θ and η for this ordering with the corresponding values ($\bar{\theta}$ and $\bar{\eta}$) which result when the natural row by row numbering scheme is used.

TABLE A.1

n	N	θ	η	$\bar{\theta}$	$\bar{\eta}$
4	25	376	100	504	120
8	81	3,172	572	4,496	720
16	289	28,664	3,340	50,336	4,896
32	1,089	257,036	18,828	669,792	36,459

Acknowledgment. The author thanks Professors Donald J. Rose, Cleve B. Moler and R. B. Simpson for many helpful suggestions. The referees' comments are also gratefully acknowledged.

REFERENCES

[1] F. W. DORR, *The direct solution of the discrete Poisson equation on a rectangle*, SIAM Rev., 12 (1970), pp. 248–263.

[2] GEORGE FIX AND KATE LARSEN, *Iterative methods for finite element approximations to elliptic boundary value problems*, this Journal, 8 (1971), pp. 536–547.

[3] G. E. FORSYTHE AND W. R. WASOW, *Finite-Difference Methods for Partial Differential Equations*, John Wiley, New York, 1960.

[4] J. A. GEORGE, *Block elimination on finite element systems of equations*, Sparse Matrices and their Applications, D. J. Rose and R. A. Willoughby, eds., Plenum Press, New York, 1972.

[5] A. D. HALL, *The Altran system for rational function manipulation—a survey*, Comm. ACM, 14 (1971), pp. 517–522.

[6] A. J. HOFFMAN, M. S. MARTIN AND D. J. ROSE, *Complexity bounds for regular finite difference and finite element grids*, this Journal, 10 (1973), pp. 364–369.

[7] S. V. PARTER, *The use of linear graphs in Gauss elimination*, SIAM Rev., 3 (1961), pp. 119–130.

[8] D. J. ROSE, *A graph-theoretic study of the numerical solution of sparse positive definitive systems of linear equations*, Graph Theory and Computing, R. Read, ed., Academic Press, New York, 1972.

[9] R. S. VARGA, *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, N.J., 1962.

[10] J. H. WILKINSON, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, 1965.