SPARSITY ANALYSIS OF THE QR FACTORIZATION

D. HARE, C.R. JOHNSON, D.D. OLESKY

AND

P. VAN DEN DRIESSCHE

DMS-536-IR

MARCH 1990
REVISED: NOVEMBER 1991

Sparsity Analysis of the QR Factorization

by

Donovan R. Hare
Department of Mathematics and Statistics
University of Victoria
Victoria, British Columbia
Canada V8W 3P4

Charles R. Johnson[1]
Department of Mathematics
College of William and Mary
Williamsburg, VA 23185
U.S.A.

D.D. Olesky[2]
Department of Computer Science
University of Victoria
Victoria, British Columbia
Canada V8W 3P6

and

P. van den Driessche[3]
Department of Mathematics and Statistics
University of Victoria
Victoria, British Columbia
Canada V8W 3P4

(1) The work of this author was supported in part by the National Science Foundation grant DMS90-00839 and Office of Naval Research Contract N00014-90-J-1739.

(2) The work of this author was supported in part by NSERC grant A-8214 and by the University of Victoria Committee on Faculty Research and Travel.

(3) The work of this author was supported in part by NSERC grant A-8965 and by the University of Victoria Committee on Faculty Research and Travel.

Sparsity Analysis of the QR Factorization

Abstract

Given only the zero–nonzero pattern of an m–by–n matrix A of full column rank, which entries of Q and which entries of R in its QR factorization must be zero and which entries may be nonzero? We give a complete answer to this question, which involves an interesting interplay between combinatorial structure and the algebra implicit in orthogonality. To this end some new sparse structural concepts are introduced, and an algorithm to determine the structure of Q is given. The structure of R then follows immediately from that of Q and A. The computable zero/nonzero structures for the matrices Q and R are proven to be tight, and the conditions on the pattern for A are the weakest possible (namely that it allows matrices A with full column rank). This complements existing work that focussed upon R and then only under an additional combinatorial assumption (the strong Hall property).

Key words: QR factorization, bipartite graph, Hall property, sparse matrix, combinatorial matrix theory, orthogonality.

1. Introduction.

For $m \geq n$, an m-by-n matrix $A = [a_{ij}]$ has a unique factorization $A = QR$, in which $Q = [q_{ij}]$ is m-by-n and has orthonormal columns and $R = [r_{ij}]$ is n-by-n and upper triangular with positive diagonal entries, exactly when A has full rank, n. In this event we say that A has a unique QR factorization, and we mean the one in which R has positive diagonal. There are several possible ways to compute such a factorization (with exact arithmetic), for example, using plane rotations (method of Givens), Householder transformations, or the Gram-Schmidt procedure. When A has a unique QR factorization, it does not matter, except for insight or computational considerations, which algorithm is used to obtain the factorization.

By a nonzero pattern $\mathcal{A}$ of size m-by-n we mean the set of positions in an m-by-n matrix in which the nonzero entries occur. We typically describe such a pattern via an m-by-n array with 0 and * entries, in which * denotes a nonzero. For example $A = \begin{bmatrix} -3 & 0 \\ 1 & 2 \\ 0 & \pi \end{bmatrix}$ is a matrix with pattern $\mathcal{A} = \begin{bmatrix} * & 0 \\ * & * \\ 0 & * \end{bmatrix}$. If matrix A has pattern $\mathcal{A}$, then we write $A \in \mathcal{A}$. By $\mathcal{A} + \mathcal{B}$ (or $\mathcal{A} \cup \mathcal{B}$) we mean the union of two patterns of the same size, that is, a pattern that has a zero in a given position only when both patterns are zero in that position. Multiplication of patterns is carried out under the following assumptions: $(* \cdot *) = *$ and $(* \cdot 0) = (0 \cdot *) = (0 \cdot 0) = 0$. Two column vectors (or vector patterns) are combinatorially orthogonal if each term in the inner product is zero.

Our primary purpose is to give a complete solution to the following problem. For $m \geq n$, given an m-by-n nonzero pattern $\mathcal{A}$ that allows full rank, determine the union, over all full column rank matrices $A \in \mathcal{A}$, of all

patterns occurring in the matrix Q and in the matrix R such that A = QR is the QR factorization of A. The resulting unions are denoted by $\mathcal{Q}$ and $\mathcal{R}$, respectively. Our solution to this problem continues earlier work [GH, GLN, CEG, and a small part of GN] that focussed mainly upon the upper triangular pattern $\mathcal{R}$ (rather than $\mathcal{Q}$) and gave a description only under an additional combinatorial assumption (the strong Hall property) on the pattern $\mathcal{A}$. The present work is based on the Gram-Schmidt procedure and is in the spirit of an analogous combinatorial analysis of the LU factorization of a square matrix that was given in [JOD]. However, there are subtleties that are quite different from the LU case.

We note that, for a given pattern $\mathcal{A}$, different zero patterns may occur among the Q's, or among the R's, in the QR factorization of different full rank matrices A with pattern $\mathcal{A}$. This is due to the possibility of chance numeric relations among the entries of a specific matrix A which cause "accidental zeros". Our interest is in ignoring chance cancellation and in obtaining the union of all patterns for both Q and R. Knowledge of $\mathcal{Q}$ and $\mathcal{R}$ is useful for computational purposes as they are the smallest patterns that are guaranteed to contain the nonzero entries of the QR factorization of an arbitrary $A \in \mathcal{A}$. Thus $\mathcal{Q}$ and $\mathcal{R}$ could be used to define a suitable data structure for Q and R, respectively.

Some particularly simple cases occur when $\mathcal{A}$ is square. If $\mathcal{A}$ is full upper triangular, then $\mathcal{R}$ has the same pattern and $\mathcal{Q}$ is diagonal. If $\mathcal{A}$ is full upper Hessenberg, then $\mathcal{Q}$ has this same pattern and $\mathcal{R}$ is full. However, if $\mathcal{A}$ is full lower triangular, then both $\mathcal{R}$ and $\mathcal{Q}$ are full.

A combinatorial necessary condition for an m-by-n matrix A to have rank n is the following (see, e.g. [CEG]).

Definition. For $m \geq n$, an m–by–n matrix A (or pattern $\mathcal{A}$) satisfies the Hall property if every k columns, $1 \leq k \leq n$, collectively have nonzero entries in at least k rows.

It is clear that there are full rank matrices with pattern $\mathcal{A}$ if and only if $\mathcal{A}$ satisfies the Hall property. So that our problem be well–defined, we wish to assume that each matrix A to be factored has a unique QR factorization. Thus, the problem statement assumed that A has full column rank, and therefore, we restrict our attention to patterns satisfying the Hall property. If A has a unique QR factorization $A = QR$, then, for an m–by–m permutation matrix P, PA also has a unique QR factorization $PA = (PQ)R$. Since Q is modified in this simple way, and R not at all, we may permute the rows of A for convenience. For example, under the Hall property assumption, we may permute the rows so that the diagonal entries are all nonzero. We note, however, that permutation of the columns of A, in general, radically changes both Q and R.

In [CEG] the possible nonzeros for $\mathcal{R}$ are identified under an additional assumption. (The following definition, a slight variant from what appears in [CEG], now seems to be standard.)

Definition. For $m \geq n$, an m–by–n matrix A (or pattern $\mathcal{A}$) satisfies the strong Hall property if

(i) $m = n > 1$ and every k columns, $1 \leq k < n$, collectively have nonzero entries in more than k rows, or

(ii) $m > n$ and every k columns, $1 \leq k \leq n$, collectively have nonzero entries in more than k rows.

Note that, if $m = n$ and each main diagonal entry of A is nonzero, then A satisfies the strong Hall property if and only if A is irreducible. Under the strong Hall assumption on $\mathcal{A}$ it is shown in [CEG] that $\mathcal{R}$ is accurately predicted by applying "symbolic Gaussian elimination" (i.e. "symbolic LU factorization") to the symbolic product $\mathcal{A}^T\mathcal{A}$, or by doing "symbolic Givens rotations" on $\mathcal{A}$. In this case, $\mathcal{R}$ can also be bounded by the pattern of U in the symbolic LU factorization (with partial pivoting) of $\mathcal{A}$, see [GN].

Without the strong Hall assumption, the complete analysis is considerably more delicate. There can be interplay between the combinatorics of the arrangement of zeros and nonzeros and the algebra (associated with orthogonality required in Q) of the QR calculation that forces zeros that are missed by any naive analysis. This is illustrated in the following example from [CEG], in which $\mathcal{A}$ satisfies the Hall property but not the strong Hall property.

Example (1.1). Let

$$\mathcal{A} = \begin{bmatrix} * & 0 & 0 & 0 & 0 & 0 \\ * & * & * & 0 & 0 & 0 \\ 0 & 0 & * & * & * & * \\ 0 & 0 & 0 & * & 0 & 0 \\ 0 & 0 & 0 & 0 & * & 0 \\ * & 0 & 0 & 0 & 0 & * \end{bmatrix}. \tag{1.1}$$

As pointed out in [Ge], when $\mathcal{Q}$ and $\mathcal{R}$ are predicted by a symbolic analog of the (numeric) method of Givens (where nonzero entries in the jth column are zeroed out in the order $a_{j+1,j}$ to $a_{n,j}$), too much fill is predicted. Specifically this approach gives

$$\mathcal{Q} = \begin{bmatrix} * & * & \otimes & \otimes & \otimes & * \\ * & * & \otimes & \otimes & \otimes & \otimes \\ 0 & 0 & * & \otimes & \otimes & \otimes \\ 0 & 0 & 0 & * & \otimes & \otimes \\ 0 & 0 & 0 & 0 & * & \otimes \\ * & * & \otimes & \otimes & \otimes & * \end{bmatrix}, \quad \mathcal{R} = \begin{bmatrix} * & * & * & 0 & 0 & * \\ & * & * & 0 & 0 & * \\ & & * & * & * & * \\ & & & * & \otimes & \otimes \\ & O & & & * & \otimes \\ & & & & & * \end{bmatrix}, \quad (1.2)$$

where $\otimes = *$, whereas $\otimes = 0$ gives the true patterns. The entries in $\mathcal{R}$ designated by $\otimes$ are called "bogus fill" in [CEG]. The pattern for $\mathcal{R}$ in (1.2) with $\otimes = *$ is also obtained when symbolic LU factorization of the symbolic product $\mathcal{A}^T\mathcal{A}$ is performed [GH] or when symbolic LU factorization (with partial pivoting) of $\mathcal{A}$ is performed [GN]. ■

In the next section, we begin by defining the notion of a Hall set. This enables us to define a sequence of bipartite graphs that lead to the identification of all zero entries in $\mathcal{Q}$. The orthogonality of Q can require zero entries in $\mathcal{Q}$ in somewhat subtle ways. Some graph-theoretic results concerning these bipartite graphs are given in section 3. The purpose of section 4 is to show that the remaining entries of $\mathcal{Q}$ are, in fact, nonzeros. It begins with explicit consideration of a special form for the pattern of $\mathcal{A}$ and then uses embedding and continuity to prove our result for general patterns. An algorithm for determining $\mathcal{Q}$ is given. In section 5 it is shown how to determine the pattern $\mathcal{R}$ in a simple way, once the pattern $\mathcal{Q}$ is known.

2. Zero Entries in $\mathcal{Q}$.

Related to the Hall properties, we introduce the concept of a Hall set. Given any m-by-n matrix A (or pattern $\mathcal{A}$) we denote the jth column of A (or $\mathcal{A}$) by A_j (or $\mathcal{A}_j$), and the restriction of A_j to the row index set α by $A_j[\alpha]$.

Definition. For an m-by-n matrix A (or pattern $\mathcal{A}$) with $m \geq n$, a set of k columns in A (or $\mathcal{A}$), $1 \leq k \leq n$, is a Hall set if these columns collectively have nonzero entries in exactly k rows.

Under the assumption that A (or $\mathcal{A}$) has the Hall property, the union of two Hall sets is a Hall set, so there exists a unique Hall set of maximum cardinality (≥ 0) in any given set of columns. Let S_j be the Hall set of maximum cardinality in the first j columns; we define $S_0 = \phi$. For example, in the pattern (1.1), $S_1 = \phi$ and $S_j = \left\{\mathcal{A}_2, \ldots, \mathcal{A}_j\right\}$ for $2 \leq j \leq 5$. Note that if an m-by-n matrix has the strong Hall property, then $S_j = \phi$, $1 \leq j \leq n-1$.

The Gram-Schmidt orthonormalization process (see, e.g. [HJ, 0.6.4]) can be used to construct the matrix Q of the QR factorization of a matrix A. We show below how to adapt the Gram-Schmidt procedure to determine $\mathcal{Q}$ for a pattern $\mathcal{A}$. Clearly $\mathcal{Q}_1 = \mathcal{A}_1$ as Q_1 is a (nonzero) multiple of A_1. If $\mathcal{A}_2$ is combinatorially orthogonal to $\mathcal{A}_1$, then $\mathcal{Q}_2 = \mathcal{A}_2$. However, if $\mathcal{A}_2$ is not combinatorially orthogonal to $\mathcal{A}_1$ and $S_1 = \phi$, then $\mathcal{Q}_2 = \mathcal{A}_1 + \mathcal{A}_2 = \mathcal{Q}_1 + \mathcal{A}_2$. The effect of Hall sets on this columnwise "naive accumulation" of nonzero entries is now investigated.

Let s_j be the set of all row indices covered by the columns of S_j, thus $|S_j| = |s_j|$. (Note that if we assume all diagonal entries of $\mathcal{A}$ are

*, then $s_j = \{i: 1 \le i \le j,\ \mathcal{A}_i \in S_j\}$; however, we do not make this assumption in this section.) For a given $\mathcal{A}$, we define a bipartite graph that leads to a partition of the row indices of $\mathcal{A}$ and identifies the zero entries in a column of $\mathcal{A}$. For a fixed j, $1 \le j \le n$, the bipartite graph $B_j(\mathcal{A}) = (R_j(\mathcal{A}), C_j(\mathcal{A}); E_j(\mathcal{A}))$, is defined as follows:

$C_j(\mathcal{A}) = \{c_k = k: 1 \le k \le j$ and $\mathcal{A}_k \notin S_{j-1}\}$;

$R_j(\mathcal{A}) = \{r_i = i: 1 \le i \le m$ and there exists $c_k \in C_j(\mathcal{A})$ with a * in row r_i and $r_i \notin s_{j-1}\}$;

$\{r_i, c_k\} \in E_j(\mathcal{A})$, the undirected edge set, if and only if $r_i \in R_j(\mathcal{A})$ and $c_k \in C_j(\mathcal{A})$ and the (i, k) position of $\mathcal{A}$ is *.

Thus, to construct $B_j(\mathcal{A})$, delete columns in S_{j-1}, rows s_{j-1} and zero rows from $[\mathcal{A}_1, \ldots, \mathcal{A}_j]$ and take the bipartite graph of the remaining pattern (with the labelling inherited from the original pattern).

A <u>path</u> (undirected) in $B_j(\mathcal{A})$ is an alternating sequence of row and column vertices of the form $r_{k_1}, c_{j_1}, r_{k_2}, \ldots, r_{k_i}, c_{j_i}$ with each edge $\{r_{k_1}, c_{j_1}\}, \{c_{j_1}, r_{k_2}\}, \ldots, \{r_{k_i}, c_{j_i}\} \in E_j(\mathcal{A})$. Such a path is associated with the sequence $(k_1, j_1), (k_2, j_1), \ldots, (k_i, j_i)$ of entries that are * in $\mathcal{A}$. A path may begin or end with either a row or column vertex, may have just one vertex, and is always assumed to be simple.

Let p_j be the set of all row indices occurring in $R_j(\mathcal{A})$ that can be reached by a path from c_j. Let u_j be the set of all row indices occurring in $R_j(\mathcal{A})$ that cannot be reached by a path from c_j. Finally, let t_j be

the set of all row indices of $\mathcal{A}$ not in $R_j(\mathcal{A})$ or s_{j-1}. We then have that s_{j-1}, p_j, u_j and t_j form a partition of all m row indices.

To illustrate this partition, consider again the pattern $\mathcal{A}$ in example (1.1); the bipartite graphs for $j = 3$ and for $j = 6$ are shown in Figure 1.

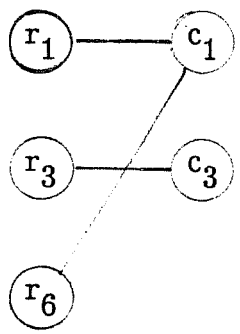
The bipartite graph $B_3(\mathcal{A})$

$s_2 = \{r_2\}$

$p_3 = \{r_3\}$

$u_3 = \{r_1, r_6\}$

$t_3 = \{r_4, r_5\}$



The bipartite graph $B_6(\mathcal{A})$

$s_5 = \{r_2, r_3, r_4, r_5\}$

$p_6 = \{r_1, r_6\}$

$u_6 = \phi$

$t_6 = \phi$

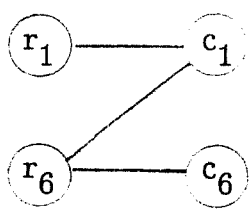


Figure 1.

We use the partition defined above to identify the zero entries in a column of $\mathcal{Q}$.

Theorem 2.1: For $m \geq n$, let $\mathcal{A}$ be an m-by-n pattern with the Hall property. For any fixed j, $1 \leq j \leq n$, the entries of $\mathcal{Q}_j$ in positions s_{j-1}, t_j and u_j are zero.

Proof: We rely on two facts that follow from the Gram-Schmidt procedure:
(i) $Q_j \in \text{span}\{A_1,\ldots,A_j\}$, and
(ii) Q_j is orthogonal to $A_1, \ldots, A_{j-1}$.
Consider first the positions t_j. In these rows the entries in columns $A_1, \ldots, A_j$ of any $A \in \mathcal{A}$ are all zero; so by (i), the positions t_j are all zero for any $Q_j \in \mathcal{Q}_j$.

For the positions s_{j-1}, fact (ii) above implies that Q_j is orthogonal to each column in S_{j-1}. But the columns of S_{j-1}, restricted to rows s_{j-1}, span a subspace of dimension $|s_{j-1}|$ and are zero outside these rows (by the definition of a Hall set). Thus $Q_j[s_{j-1}]$ is orthogonal to every vector in the above subspace, and so must be zero. Thus the positions s_{j-1} of any $Q_j \in \mathcal{Q}_j$ are zero; these are due to the presence of a Hall set.

Finally consider the positions u_j, and let $\left\{A_{i_1},\ldots,A_{i_k}\right\}$ be the set of columns on a path in $B_j(\mathcal{A})$ from indices in u_j. From (i) and because $u_j \cap (p_j \cup s_{j-1}) = \phi$, $Q_j[u_j] \in \text{span}\left\{A_{i_1}[u_j],\ldots,A_{i_k}[u_j]\right\}$. Because of (ii), and because $A_{i_g}[p_j] = 0$, $g = 1, \ldots, k$, and $Q_j[s_{j-1}] = 0$ (and $A_{i_g}[t_j] = 0$, $g = 1, \ldots, k$, and $Q_j[t_j] = 0$), $Q_j[u_j]$ is orthogonal to each $A_{i_g}[u_j]$. However, as the intersection of a subspace and its orthogonal complement is zero, $Q_j[u_j] = 0$, and the positions u_j of any $Q_j \in \mathcal{Q}_j$ are zero. ■

We have thus identified three sets of row indices in $\mathcal{Q}_j$ that are zero for any full rank matrix $A \in \mathcal{A}$ with $A = QR$. In particular, the presence of a Hall set in columns $\left\{\mathcal{A}_1,\ldots,\mathcal{A}_{j-1}\right\}$ means that $q_{i,j} = 0$ for $i \in s_{j-1}$ (in

fact, $q_{i,j+k} = 0$ for $i \in s_{j-1}$ and $k = 0, \ldots, n-j)$. If $i \in t_j$, that is $a_{ik} = 0$ for $k = 1, \ldots, j$, then $q_{ij} = 0$. If $i \in u_j$, then $q_{ij} = 0$ because of combinatorial orthogonality in the columns of $\mathcal{A}$ which occurs in a subtle way (see corollary 4.9).

3. Strong Hall Bipartite Graphs.

In this section we analyze the strong Hall property from a graph–theoretic point of view. The objective is to obtain a result (theorem 3.2) concerning the bipartite graphs $B_j(\mathcal{A})$ defined in section 2. We first prove a preliminary result concerning strong Hall bipartite graphs, which we now define.

Definition. Let $G(X,Y;E)$ denote a bipartite graph with vertex sets X, Y and undirected edge set E and if $S \subseteq Y$, let $N_G(S)$ denote the set of all neighbors in X of vertices in S. Then $G(X,Y;E)$ is strong Hall with respect to Y if

(i) $|X| = |Y| > 1$ and $|S| < |N_G(S)|$, for all proper nonempty subsets S of Y, or

(ii) $|X| > |Y|$ and $|S| < |N_G(S)|$, for all nonempty subsets S of Y.

There is an obvious equivalence between patterns $\mathcal{A}$ which satisfy the strong Hall property and bipartite graphs $G(X,Y;E)$ that are strong Hall with respect to Y when the vertex sets X, Y are associated with the row and column sets of $\mathcal{A}$, respectively, and edges correspond to nonzero entries in $\mathcal{A}$. Note that the bipartite graphs $B_j(\mathcal{A})$ defined in section 2 are strong Hall with respect to $C_j(\mathcal{A})\backslash\{c_j\}$.

The following theorem is used to prove theorem 3.2, but is of independent interest as it gives a characterization of strong Hall critical graphs $G = (X,Y;E)$ with $|X| > |Y|$ (i.e., strong Hall graphs that are no longer strong Hall if any one edge is removed). Let G be strong Hall with respect

to Y. If for all $y \in Y$, the degree $d_G(y) = 2$, then it is clear that G is strong Hall critical with respect to Y. We now show that the converse is also true.

Theorem 3.1. Let $G = (X,Y;E)$ be strong Hall with respect to Y where $|X| > |Y|$. If $y \in Y$ has degree $d_G(y) \geq 3$, then there is at most one edge $e = \{x,y\} \in E$ such that $G-e$ is not strong Hall with respect to Y.

Proof: Suppose there exists $y \in Y$ with $d_G(y) \geq 3$, and there is an edge $e = \{x,y\} \in E$ such that $H = G - e$ is not strong Hall with respect to Y. Let $S \subseteq Y$, $S \neq \emptyset$, be such that $|N_H(S)| \leq |S|$. If $y \notin S$, then $N_H(S) = N_G(S)$ and hence $|N_H(S)| = |N_G(S)| > |S|$ since G is strong Hall with respect to Y. Therefore $y \in S$. If $x \in N_G(S\backslash\{y\})$, then again $N_H(S) = N_G(S)$, a contradiction, and so $x \notin N_G(S\backslash\{y\})$. Thus $|N_G(S)| = |N_H(S)| + 1 \leq |S| + 1$ and since G is strong Hall with respect to Y, $|N_G(S)| = |S| + 1$.

Now suppose there exists $y \in Y$ with $d_G(y) \geq 3$ and there exist distinct edges $e_1 = \{x_1,y\}$ and $e_2 = \{x_2,y\}$ in E such that for $i = 1, 2$, $H_i = G - e_i$ is not strong Hall with respect to Y. Then for $i = 1, 2$ there exists $S_i \subseteq Y$, $S_i \neq \emptyset$, such that $|N_{H_i}(S_i)| \leq |S_i|$. From above, $y \in S_1 \cap S_2$ and for $i = 1, 2$

$$|N_G(S_i)| = |S_i| + 1 \tag{3.1}$$

and

$$x_i \notin N_G(S_i\backslash\{y\}). \tag{3.2}$$

First suppose $(S_1 \cap S_2)\backslash\{y\} = S \neq \emptyset$. For each $z \in S$ and $i = 1, 2$, $\{z,x_i\}$ is not an edge in G by (3.2). Thus

$$|N_G(S_1 \cap S_2)| = |N_G(\{y\}) \setminus N_G(S)| + |N_G(S)|$$

$$\geq 2 + |N_G(S)| \quad \text{since} \quad x_1, x_2 \notin N_G(S)$$

$$> 2 + |S| \quad \text{since} \quad S \neq \emptyset \quad \text{and} \quad G \quad \text{is strong Hall}$$

$$= 2 + (|S_1 \cap S_2| - 1)$$

$$= |S_1 \cap S_2| + 1.$$

If, on the other hand, $S_1 \cap S_2 = \{y\}$, then

$$|N_G(S_1 \cap S_2)| = |N_G(\{y\})| \geq 3 > |S_1 \cap S_2| + 1.$$

Therefore in any case

$$|N_G(S_1) \cap N_G(S_2)| \geq |N_G(S_1 \cap S_2)| \geq |S_1 \cap S_2| + 2. \tag{3.3}$$

Thus

$$|S_1 \cup S_2| = (|S_1|+1)+(|S_2|+1)-(|S_1 \cap S_2|+2)$$

$$\geq |N_G(S_1)| + |N_G(S_2)| - |N_G(S_1) \cap N_G(S_2)| \quad \text{by (3.1) and (3.3)}$$

$$= |N_G(S_1) \cup N_G(S_2)|$$

$$= |N_G(S_1 \cup S_2)|.$$

This contradicts that G is strong Hall with respect to Y (note that $|X| > |Y|$). Hence, at most one such edge e_i can exist. ■

The next theorem provides a matching property of strong Hall bipartite graphs that is used to prove the main theorem in section 4. (Recall that a matching M in a bipartite graph G is a subset of its edges no two of which are adjacent. We say that M saturates a subset S of the vertices of G if every vertex in S is incident with one of the edges in M.) We will use the following notation in the next theorem. If γ_1 and γ_2 are paths in a graph G, and γ_1 ends in a vertex adjacent in G to the first vertex in γ_2, then $\gamma_1\gamma_2$ denotes the juxtaposition of γ_1 and γ_2 as alternating sequences of vertices and edges so that the edge from the end of γ_1 to the beginning of γ_2 is added giving a single path.

Theorem 3.2. Let $G = (X,Y;E)$ be a bipartite graph with $|X| \geq |Y|$, let $y \in Y$, and suppose that $G-y$ is strong Hall with respect to $Y\backslash\{y\}$. If there is a path γ from y to x for some $x \in X$, then there exists a path τ from y to x and a matching M in $G-\tau$ such that M saturates all vertices in Y that are not on the path τ.

Proof: Suppose the statement of the theorem is false and let $G = (X,Y;E)$ be a counterexample with the fewest edges. Then there exists $x \in X$ and a path γ from y to x. Let $\hat{Y} = Y\backslash\{y\}$ and $\hat{G} = G - y$. Since G is a counterexample there must exist a vertex $\hat{y} \in Y$ not on γ.

If $d_{\hat{G}}(\hat{y}) \geq 3$, then by theorem 3.1, there exists $\hat{e} = \{\hat{x},\hat{y}\}$ such that if $H = G - \hat{e}$, then $H-y$ is strong Hall with respect to $\hat{Y}$. Now γ is a path in H also, and since G is a counterexample with the fewest edges, there

exists some path τ from y to x and a matching M in $H-\tau$ such that M saturates all the vertices in Y that are not on the path τ. But this matching and path are also in G, a contradiction.

Therefore $d_{\hat{G}}(\hat{y}) = 2$ (note that $d_{\hat{G}}(\hat{y}) > 1$ since $\hat{G}$ is strong Hall with respect to $\hat{Y}$). Let x_1 and x_2 be the neighbors of $\hat{y}$ and let $H = (U,V;F)$ be the graph obtained from G by identifying x_1 with x_2 and by removing $\hat{y}$. That is, $U = (X\backslash\{x_1,x_2\})\cup\{x^*\}$ where $x^* \notin X \cup Y$ and $V = Y\backslash\{\hat{y}\}$. Moreover, for all $u \in U$ and $v \in V$, $\{u,v\} \in F$ if and only if

(i) $u \in X\backslash\{x_1,x_2\}$, $v \in V$ and $\{u,v\} \in E$; or
(ii) $u = x^*$, $v \in V$, and $\{x_1,v\} \in E$ or $\{x_2,v\} \in E$.

Let $\hat{V} = V\backslash\{y\}$, $\hat{H} = H - y$ and note that since $|X| > |\hat{Y}|$, we have $|U| > |\hat{V}|$.

Claim: $\hat{H}$ is strong Hall with respect to $\hat{V}$.

Proof of Claim: Let $S \subseteq \hat{V}$, $S \neq \emptyset$. If $x^* \notin N_{\hat{H}}(S)$, then $x_1 \notin N_{\hat{G}}(S)$ and $x_2 \notin N_{\hat{G}}(S)$. Thus $N_{\hat{H}}(S) = N_{\hat{G}}(S)$ and since $\hat{G}$ is strong Hall with respect to $\hat{Y}$, $|S| < |N_{\hat{G}}(S)| = |N_{\hat{H}}(S)|$.

Suppose there exists $S \subseteq \hat{V}$ such that $|S| \geq |N_{\hat{H}}(S)|$. Then $x^* \in N_{\hat{H}}(S)$, by the above, and hence at least one of x_1 or x_2 is contained in $N_{\hat{G}}(S)$. If only one is contained in $N_{\hat{G}}(S)$, then since it is only replaced with x^* in H, we have $|N_{\hat{G}}(S)| = |N_{\hat{H}}(S)|$, a contradiction to $|S| < |N_{\hat{G}}(S)|$. Thus both x_1 and x_2 are contained in $N_{\hat{G}}(S)$. Since x_1 and x_2 are replaced with x^* in H, $|N_{\hat{G}}(S)| = |N_{\hat{H}}(S)| + 1$ and $N_{\hat{G}}(S\cup\{\hat{y}\}) = N_{\hat{G}}(S)$. But then

$$|S\cup\{\hat{y}\}| = |S| + 1 \geq |N_{\hat{H}}(S)| + 1 = |N_{\hat{G}}(S)| = |N_{\hat{G}}(S\cup\{\hat{y}\})|$$

contradicts the fact that $\hat{G}$ is strong Hall with respect to $\hat{Y}$ (note that $S\cup\{\hat{y}\}\subseteq\hat{Y}$). This completes the proof of the claim.

We now identify a path η in H that is dependent upon the path γ in G. The following table describes how η is defined in H given the structure of γ. In the table, π, ρ and σ are nonempty paths (i.e., they each contain at least one vertex).

Case	γ is	path η in H is
1	γ (x_1 and x_2 not on γ)	γ
2	$\pi x_1\rho$ or $\pi x_2\rho$	$\pi x^*\rho$
3	$\pi x_1\rho x_2\sigma$ or $\pi x_2\rho x_1\sigma$	$\pi x^*\sigma$
4	$\pi x_1\rho x_2$ or $\pi x_2\rho x_1$	πx^*
5	πx_1 or πx_2	πx^*

Let η end at vertex u. Note that $u \in U$ and that η starts at vertex y. Since $\hat{H}$ is strong Hall with respect to $\hat{V}$, and H has fewer edges than G, there exists a path ω in H from y to u and matching $\hat{M}$ in $H-\omega$ which saturates all vertices in V that are not on ω. We will use this path and matching to construct similar ones in G. The constructions will differ depending on the structure of γ in G, but each will contradict

that G is a counterexample to the theorem . The cases below are described in the table above.

Cases 1–3. Here we have $x_1 \neq u \neq x_2$ and hence $u = x$.

If x^* is not on ω and x^* is not saturated by $\hat{M}$, then ω is a path in G from y to x and $M = \hat{M} \cup \{\{x_1,\hat{y}\}\}$ is a matching in G which saturates all vertices in Y that are not on ω, a contradiction.

If there exists a $v \in V$ such that $\{x^*,v\} \in \hat{M}$, then x^* is not on ω and hence ω is a path in G from y to x. Moreover, $\{z,v\} \in E$ for some $z \in \{x_1,x_2\}$. Let $\{w\} = \{x_1,x_2\}\backslash\{z\}$. Thus $M = (\hat{M}\backslash\{\{x^*,v\}\}) \cup \{\{z,v\}, \{w,\hat{y}\}\}$ is a matching in G which saturates all vertices in Y that are not on ω, a contradiction.

Therefore x^* is on ω and $\hat{M}$ is a matching in G. Let $\omega = \omega_1 x^* \omega_2$ and let v_1 and v_2 be the neighbors of x^* on ω such that ω_1 ends at v_1 and ω_2 starts at v_2. If $\{x_1,x_2\} \subseteq N_G(\{v_1,v_2\})$, then at least one of $\tau_1 = \omega_1 x_1 \hat{y} x_2 \omega_2$ and $\tau_2 = \omega_1 x_2 \hat{y} x_1 \omega_2$ is a path in G that starts at y and ends at x. Thus for some $i \in \{1,2\}$, $\hat{M}$ saturates all vertices in Y that are not on τ_i, a contradiction. Hence $\{x_1,x_2\} \not\subseteq N_G(\{v_1,v_2\})$. Thus since x^* is adjacent to v_1 and v_2, they are both adjacent to x_1 or to x_2 in G. If v_1 and v_2 are neighbors of x_1, then $\tau = \omega_1 x_1 \omega_2$ is a path in G from y to x, and $M = \hat{M} \cup \{\{x_2,\hat{y}\}\}$ is a matching in G that saturates all vertices in Y that are not on τ, a contradiction. Similarly, if v_1 and v_2 are neighbors of x_2, then $\tau = \omega_1 x_2 \omega_2$ is a path in G from y to x, and $M = \hat{M} \cup \{\{x_1,\hat{y}\}\}$ is a matching in G that saturates all vertices in Y that are not on τ, another contradiction. Therefore, Cases 1–3 cannot occur.

Cases 4 and 5. Note that $u = x^*$ and hence ω ends at x^*. Moreover, $\hat{M}$ is a matching in G. Let $\omega = \omega_1 x^*$ and let v_1 be the last vertex of the path ω_1. Since v_1 is adjacent to x^* in H, it is adjacent to either x_1 or x_2 in G. Thus either $\tau_1 = \omega_1 x_1$, $\tau_2 = \omega_1 x_2 \hat{y} x_1$, $\tau_3 = \omega_1 x_2$, or $\tau_4 = \omega_1 x_1 \hat{y} x_2$ is a path from y to x in G. If τ_2 is such a path in G, then $\hat{M}$ saturates all vertices in Y that are not on τ_2; similarly if τ_4 is a path from y to x in G. Otherwise, $\hat{M} \cup \{\{x_2, \hat{y}\}\}$ saturates all vertices in Y that are not on τ_1 or $\hat{M} \cup \{\{x_1, \hat{y}\}\}$ saturates all vertices in Y that are not on τ_3, a contradiction. Therefore Cases 4 and 5 cannot occur.

Since these cases exhaust all possibilities, we have a contradiction. Therefore no such counterexample G can exist and the theorem is proved. ■

4. Nonzero entries in $\mathcal{Q}$.

In section 2, we have identified positions in $\mathcal{Q}$ that must be zero for a variety of more or less subtle reasons. To complete the solution to our problem and see that we have identified all the necessarily zero positions in $\mathcal{Q}$, we must show that for each remaining position there is a full rank matrix $A \in \mathcal{A}$ whose QR factorization has a nonzero entry in Q in the position of interest. Our strategy is, for *each* nonzero position, to identify a particular full rank matrix A which produces a Q with a nonzero in that position and whose pattern is *contained in* that of $\mathcal{A}$. A simple perturbation argument then yields an A with the *exact* pattern of $\mathcal{A}$. We note that our argument is sufficient to prove that the zero/nonzero structure for $\mathcal{Q}$ is tight (i.e., the best possible).

We first consider a particularly simple form for $\mathcal{A}$. To describe our construction, we introduce the notation Q^u for a matrix that has orthogonal columns that are not in general of unit length (i.e. unnormalized); Q^u_j denotes the jth column of this matrix.

Lemma 4.1. Let $\mathcal{A}$ be an f–by–f pattern with bipartite graph $B_f(\mathcal{A}) = (R_f(\mathcal{A}), C_f(\mathcal{A}); E_f(\mathcal{A}))$, in which $R_f(\mathcal{A}) = \{r_1,\ldots,r_f\}$, $C_f(\mathcal{A}) = \{c_1,\ldots,c_f\}$ and the edges in $E_f(\mathcal{A})$ form a simple path of length $2f-1$ from c_f to some r_i, $1 \le i \le f$. There exists a nonsingular matrix $A \in \mathcal{A}$ such that if $A = QR$, then q_{if} is nonzero.

Proof: Assume that the path from c_f to r_i is

$$c_f = c_{j_f},\ r_{k_f},\ c_{j_{f-1}},\ \ldots,\ c_{j_2},\ r_{k_2},\ c_{j_1},\ r_{k_1} = r_i.$$

Let $A = [a_{de}]$ be the f-by-f (0,1) matrix with $a_{de} = 1$ iff $\{r_d, c_e\}$ lies on the path; thus $A \in \mathscr{A}$. Matrix A has exactly 2 entries equal to 1 in each column (except the last). Since A is permutation equivalent to a lower triangular matrix with ones on the main diagonal, A is nonsingular. The columns (except c_f) can occur in any order on the path, but the exact order is important to the following construction of Q_f^u.

Consider

$$Q_f^u = Q_{j_f}^u = f\,A_{j_f} - (f-1)A_{j_{f-1}} + \ldots \mp 2A_{j_2} \pm A_{j_1}, \qquad (4.1)$$

where the columns are from matrix A, the multiplying factors have alternating signs and f is the number of columns on the path. Clearly this $Q_{j_f}^u \in \text{span}\{A_{j_1},\ldots,A_{j_f}\}$, and we need to check the orthogonality conditions. By the construction of A, we have $A_{j_1} \cdot A_{j_1} = 2$, $A_{j_1} \cdot A_{j_2} = 1$ and $A_{j_1} \cdot A_{j_k} = 0$ for $k = 3, \ldots, f$. Thus $Q_f^u \cdot A_{j_1} = 0$. Similarly $A_{j_k} \cdot A_{j_k} = 2$ for $k = 2, \ldots, f-1$, and $A_{j_k} \cdot A_{j_{k-1}} = A_{j_k} \cdot A_{j_{k+1}} = 1$ for $k = 2, \ldots, f-1$, with A_{j_k} (combinatorially) orthogonal to all other columns. Thus $Q_f^u \cdot A_{j_k} = 0$ for $k = 1, \ldots, f-1$. The construction of A and the distinct multipliers in (4.1) mean that there can be no numerical cancellation, and so the ith entry of Q_f^u is nonzero. It is clear that if the column Q_f^u is normalized to have unit length, then column f of the matrix Q of the QR factorization of A is obtained, with $q_{if} \neq 0$. ■

Note that the construction in lemma 4.1 guarantees that all entries in column f of Q are nonzero, not just q_{if}. To illustrate lemma 4.1, consider an example in which $i = 2$ and $f = 3$.

Example (4.2). Consider the pattern $\mathcal{A}$ and bipartite graph $B_3(\mathcal{A})$ in Figure 2.

$$\mathcal{A} = \begin{bmatrix} * & * & 0 \\ 0 & * & 0 \\ * & 0 & * \end{bmatrix} \quad \text{with } B_3(\mathcal{A}):$$

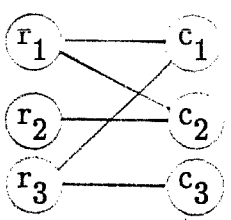


Figure 2.

The edges in $E_3(\mathcal{A})$ form one path from column c_3 to r_2, namely c_3, r_3, c_1, r_1, c_2, r_2. Take the matrix

$$A = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \in \mathcal{A}.$$

From (4.1), $Q_3^u = 3A_3 - 2A_1 + A_2 = (-1,1,1)^T$, (note the order of the A_i in this equation is determined by the path), and the (2,3) entry of Q^u is equal to 1. Thus if $A = QR$, column 3 of Q must be a scalar multiple of Q_3^u (by the uniqueness of the QR factorization), and in fact

$$Q = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{3}} \\ 0 & \frac{2}{\sqrt{6}} & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{3}} \end{bmatrix}. \quad \blacksquare$$

Example (4.3). In the event that $\mathcal{A}$ has bipartite graph $B_f(\mathcal{A})$ with edges forming a path traversing the columns in order $c_f, c_{f-1}, \ldots, c_1$, then $A \in \mathcal{A}$ is bidiagonal. For example when $f = 5$,

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} \in \mathcal{A},$$

the corresponding Q^u is the full upper Hessenberg matrix

$$Q^u = \begin{bmatrix} 1 & -1 & 1 & -1 & 1 \\ 1 & 1 & -1 & 1 & -1 \\ 0 & 2 & 1 & -1 & 1 \\ 0 & 0 & 3 & 1 & -1 \\ 0 & 0 & 0 & 4 & 1 \end{bmatrix}$$

and note that R is bidiagonal upper triangular. $\blacksquare$

To prove our result for more general patterns, we need the fact that Q depends continuously on A. More specifically, if A is a matrix with full column rank and $A = QR$, then there exists a neighborhood of A such that every matrix in this neighborhood has full column rank and the orthogonal matrix of its factorization depends continuously on the entries. This qualitative fact (which is all we need) may be straightforwardly proven by

elementary means, but precise bounds have been given in [S, Theorem 3.1]. We now use continuity to prove our main result on nonzero entries in $\mathcal{Q}$.

Theorem 4.4. For $m \geq n$, let $\mathcal{A}$ be an m-by-n pattern with the Hall property. If $r_i \in p_j$, $1 \leq j \leq n$, then there is an $A \in \mathcal{A}$ of rank n such that $A = QR$ with q_{ij} nonzero.

Proof: Without loss of generality, we can assume that each diagonal entry of $\mathcal{A}$ is *. (Thus, as noted in section 2, $s_j = \{i: 1 \leq i \leq j$ and $\mathcal{A}_i \in S_j\}$ and consequently if some column $c_e \notin C_j(\mathcal{A})$, where $1 \leq c_e < j$, then $r_e \notin R_j(\mathcal{A})$.) As $r_i \in p_j$, there exists at least one path in $B_j(\mathcal{A})$ from c_j to r_i. Thus, by Theorem 3.2, there exists a path, say

$$c_j = c_{j_f}, r_{k_f}, c_{j_{f-1}}, \ldots, c_{j_2}, r_{k_2}, c_{j_1}, r_{k_1} = r_i \qquad (4.2)$$

from c_j to r_i of length $2f-1$ $(1 \leq f \leq j)$ and a matching M in $B_j(\mathcal{A})$ of all columns in $B_j(\mathcal{A})$ that are not on the path (4.2) to some subset of the rows in $B_j(\mathcal{A})$ that are not on the path (4.2). For any nonzero ϵ, let $A_\epsilon = [a_{de}]$ be the m-by-j matrix whose pattern $\mathcal{A}_\epsilon$ is the same as the first j columns of $\mathcal{A}$ and such that $a_{de} \in \{0,1,\epsilon\}$ with nonzero terms given by the following construction:

$$a_{de} = \begin{cases} 1 \begin{cases} \text{if the edge } \{r_d, c_e\} \text{ is on the path (4.2)} \\ \underline{\text{or}} \text{ if the edge } \{r_d, c_e\} \in M \\ \underline{\text{or}} \text{ if } d = e \text{ and } c_e \notin C_j(\mathscr{A}), \text{ the column vertex set of } B_j(\mathscr{A}) \end{cases} \\ \epsilon \text{ if the (d,e) entry of } \mathscr{A} \text{ is } * \text{ and } a_{de} \text{ has not been set to 1 above} \end{cases}$$

(Note that $\mathscr{A}_{j+1}, \ldots, \mathscr{A}_n$ do not enter here.)

Let A_0 denote the matrix A_ϵ with all ϵ entries set to zero, and let $\tilde{A} = [\tilde{a}_{de}]$ be the (0, 1) square submatrix of A_0 restricted to rows $r_{k_1}, \ldots, r_{k_f}$ and columns $c_{j_1}, \ldots, c_{j_f}$, with $\tilde{a}_{de} = 1$ if and only if $a_{de} = 1$. Let $\tilde{\mathscr{A}}$ be the pattern of $\tilde{A}$; then $\tilde{\mathscr{A}}$ satisfies the conditions of lemma 4.1 and $\tilde{A}$ is the matrix of the construction in lemma 4.1. Thus if $\tilde{A} = \tilde{Q}\tilde{R}$, then $\tilde{q}_{ij}$ is nonzero.

Now we consider the matrix A_0, and note that $\tilde{A}$ is embedded in A_0. By construction, columns of A_0 not on the path (4.2) contain distinct unit vectors, so A_0 has full column rank. If $A_0 = Q_0 R_0$, then $\tilde{Q}$ is embedded in Q_0 in the same way as $\tilde{A}$ is embedded in A_0, with corresponding unit column vectors in columns not on the path. Thus the (i, j) entry of Q_0 is equal to $\tilde{q}_{ij}$ and is nonzero.

Finally, consider $A_\epsilon = Q_\epsilon R_\epsilon$, which has the same column rank as A_0, namely j, for sufficiently small ϵ. By the continuity statement above, for sufficiently small ϵ, the matrix Q_ϵ has its (i,j) entry nonzero. Thus any full rank matrix $A \in \mathscr{A}$ with the submatrix in its first j columns equal to A_ϵ has q_{ij} nonzero, where $A = QR$, since column j of Q is independent of $A_{j+1}, \ldots, A_n$. ■

To illustrate this construction, consider an example.

Example (4.5). Consider the pattern $\mathcal{A}$ and bipartite graph $B_4(\mathcal{A})$ in Figure 3.

$$\mathcal{A} = \begin{bmatrix} * & * & * & 0 & 0 \\ 0 & * & 0 & 0 & 0 \\ 0 & 0 & * & 0 & * \\ * & 0 & 0 & * & 0 \\ 0 & 0 & * & 0 & * \\ 0 & * & 0 & 0 & 0 \end{bmatrix} \quad \text{with} \quad B_4(\mathcal{A}):$$

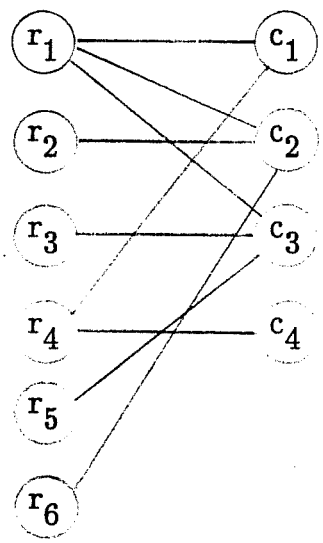


Figure 3.

There is one path in $B_4(\mathcal{A})$ from c_4 to r_2, namely c_4, r_4, c_1, r_1, c_2, r_2, implying that the (2,4) entry of $\mathcal{Q}$ is *. As this is the only path in $B_4(\mathcal{A})$ from c_4 to r_2, this must be the path (4.2) in the proof of theorem 4.4, and the matching $\mathbf{M}$ can be chosen as either the edge $\{r_3,c_3\}$ or the edge $\{r_5,c_3\}$. If the first of these, then

$$A_\epsilon = \begin{bmatrix} 1 & 1 & \epsilon & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & \epsilon & 0 \\ 0 & \epsilon & 0 & 0 \end{bmatrix} \quad \text{and} \quad \tilde{A} = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}.$$

From the proof of lemma 4.1, the third column of $\tilde{Q}$ is a scalar multiple of

$(-1,1,1)^T$, implying that the fourth column of Q_0 is a scalar multiple of $(-1,1,0,1,0,0)^T$. Thus, by continuity, the (2,4) entry of Q_ϵ is nonzero, implying that the (2,4) entry of $\mathcal{L}$ is nonzero. ■

Consider now a further example, with a pattern having more than one path from c_j to r_i (cf. example (4.5) with a change in column 1).

Example (4.6). Consider the pattern $\mathcal{A}$ and bipartite graph $B_3(\mathcal{A})$ in Figure 4.

$$\mathcal{A} = \begin{bmatrix} * & * & * \\ 0 & * & 0 \\ 0 & 0 & * \\ 0 & 0 & 0 \\ 0 & 0 & * \\ * & * & 0 \end{bmatrix} \quad \text{with} \quad B_3(\mathcal{A}):$$

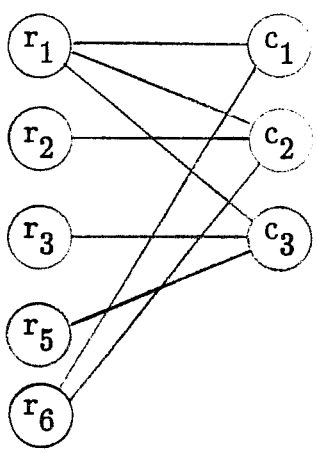


Figure 4.

Focussing on the (6, 3) entry, there are 2 paths, namely c_3, r_1, c_1, r_6 and c_3, r_1, c_2, r_6, from c_3 to r_6. The path (4.2) in the proof of theorem 4.4 (i.e., the path τ in theorem 3.2) is the first of these, and the matching M is the edge $\{r_2,c_2\}$. Note that for the other path from c_3 to r_6, there is no matching in the bipartite graph of c_1 with any row vertex not on this path (i.e., r_2, r_3, r_4 or r_5). This illustrates the care required in the selection of a path (4.2) in the proof of theorem 4.4; there must exist an associated matching M so that A_ϵ is constructed such that A_0 has full

column rank. The result of theorem 3.2 ensures that this is always possible.

With the choice of the path c_3, r_1, c_1, r_6 above, we obtain

$$A_\epsilon = \begin{bmatrix} 1 & \epsilon & 1 \\ 0 & 1 & 0 \\ 0 & 0 & \epsilon \\ 0 & 0 & 0 \\ 0 & 0 & \epsilon \\ 1 & \epsilon & 0 \end{bmatrix} \in \mathcal{A}.$$

For sufficiently small ϵ, if $A_\epsilon = A = QR$, then q_{63} is nonzero. ■

We now combine the results of theorems 2.1 and 4.4 to solve the problem stated in the introduction (as far as $\mathcal{Q}$ is concerned) in terms of the row index partition defined in section 2.

Theorem 4.7. For $m \geq n$, let $\mathcal{A}$ be an m-by-n pattern with the Hall property. Then the (i, j) entry of $\mathcal{Q}$ is zero if and only if $i \in s_{j-1} \cup t_j \cup u_j$. Equivalently, the (i, j) entry of $\mathcal{Q}$ is nonzero if and only if $i \in p_j$.

The next two corollaries indicate how $\mathcal{Q}$ can be computed. In the strong Hall case, $S_{j-1} = \emptyset$, $2 \leq j \leq n$, and $\mathcal{Q}_j$ is the union of $\mathcal{A}_j$ and certain columns $\mathcal{A}_w$, $1 \leq w \leq j-1$.

Corollary 4.8. For $m \geq n$, let $\mathcal{A}$ be an m-by-n pattern with the strong Hall property. For any fixed j, $1 \leq j \leq n$, partition $\mathcal{A}_1, \ldots, \mathcal{A}_j$ into two disjoint sets $\mathcal{X}_j$ and $\mathcal{Y}_j$, for which $\mathcal{A}_j \in \mathcal{X}_j$, every column of $\mathcal{X}_j$ is combinatorially orthogonal to every column of $\mathcal{Y}_j$, and $|\mathcal{Y}_j|$ is as large as possible. Then $\mathcal{Q}_j = \cup\, \mathcal{A}_w$, $\mathcal{A}_w \in \mathcal{X}_j$.

Proof: Since $S_{j-1} = \phi$, $2 \le j \le n$, the bipartite graph $B_j(\mathcal{A})$ has $C_j(\mathcal{A}) = \{c_1,\ldots,c_j\}$ and $R_j(\mathcal{A}) = \{r_i = i$: there exists a * in row i in one of the columns of $C_j(\mathcal{A})\}$. For $k \le j$, column $c_k \in \mathcal{X}_j$ if and only if there exists a path from c_j to some r_i where column c_k has a * in row r_i. By definition of p_j, this is true if and only if $r_i \in p_j$; and by theorem 4.7, $r_i \in p_j$ if and only if q_{ij} is nonzero. Together these give exactly the union statement above. ■

For $\mathcal{A}$ having the Hall property, the essential modification to the result of the above corollary is that in the union of columns $\mathcal{A}_w$ specifying $\mathcal{Q}_j$, the column vectors in S_{j-1} and nonzeros in rows s_{j-1} must be omitted.

Corollary 4.9. For $m \ge n$, let $\mathcal{A}$ be an m-by-n pattern with the Hall property. For any fixed j, consider the subpattern of $\mathcal{A}$ in columns $\{\mathcal{A}_1,\ldots,\mathcal{A}_j\} - S_{j-1}$ and rows $v_j = \{1,\ldots,m\} - s_{j-1}$; this is an $(m-|s_{j-1}|)$-by-$(j-|s_{j-1}|)$ pattern. Partition the columns of this subpattern into 2 disjoint sets, $\mathcal{X}_j$ and $\mathcal{Y}_j$, with $\mathcal{A}_j[v_j] \in \mathcal{X}_j$, every column of $\mathcal{X}_j$ is combinatorially orthogonal to every column of $\mathcal{Y}_j$ and $|\mathcal{Y}_j|$ is as large as possible. Then $\mathcal{Q}_j[s_{j-1}] = 0$ and $\mathcal{Q}_j[v_j] = \cup\, \mathcal{A}_w[v_j]$, with $\mathcal{A}_w[v_j] \in \mathcal{X}_j$.

Proof: By definition, p_j is the set of all row indices r_i such that

(i) $r_i \notin s_{j-1}$

(ii) there exists a column c_k, $1 \le k \le j$, such that $\mathcal{A}_k \notin S_{j-1}$ and there is a * in row r_i of c_k

and (iii) there exists a path in $B_j(\mathcal{A})$ from c_j to r_i.

Thus the positions of nonzeros in vectors $\mathcal{A}_w[v_j] \in \mathcal{X}_j$ correspond precisely to entries in p_j, and the result follows from theorem 4.7. ∎

The following algorithm implements the construction in corollary 4.9.

Algorithm 4.10. Computation of the Pattern $\mathcal{Q}$.

Notation:

$\mathcal{A}$ is an m-by-n pattern, $m \geq n$, satisfying the Hall property.

$\mathcal{A}_j$ is the j-th column of $\mathcal{A}$, $1 \leq j \leq n$.

$\beta(\mathcal{A}_j) = \{i |$ the i-th entry of $\mathcal{A}_j$ is $*\}$.

S_j, s_j are as defined in section 2.

1. For $1 \leq j \leq n-1$, compute S_j and s_j.
2. Initialization

   $a_1 = \beta(\mathcal{A}_1)$

   $\beta(\mathcal{Q}_1) = a_1$

   $a_2 = \beta(\mathcal{A}_2) - s_1$

   If $a_1 \cap a_2 \neq \emptyset$ then $a_2 = a_2 \cup a_1$

   $\beta(\mathcal{Q}_2) = a_2$
3. For $j = 3, 4, \ldots, n$ do

   3.1 If $s_{j-1} \neq s_{j-2}$ then do

   For $k = 1, 2, \ldots, j-1$ do

   If $k \notin s_{j-1}$ then do

   $a_k = \beta(\mathcal{A}_k) - s_{j-1}$

   For $i = 1, 2, \ldots, k-1$ do

   If $i \notin s_{j-1}$ and $a_k \cap a_i \neq \emptyset$ then $a_k = a_k \cup a_i$

   3.2 If $a_j = \beta(\mathcal{A}_i) - s_{j-1}$

3.3 For $i = 1, 2, \ldots, j-1$ do

If $i \notin s_{j-1}$ and $a_j \cap a_i \neq \emptyset$ then $a_j = a_j \cup a_i$

3.4 $\beta(\mathcal{Q}_j) = a_j$ ■

We illustrate this algorithm with the pattern $\mathcal{A}$ of example (4.5). The only nonempty Hall set S_j, $1 \leq j \leq 4$, is $S_4 = \{\mathcal{A}_1, \mathcal{A}_4\}$ giving $s_4 = \{1,4\}$. For $j = 4$, $s_2 = s_3 = \emptyset$, so computation of $\mathcal{Q}_4$ proceeds from step 3 of algorithm 4.10 as follows. In step 3.2, $a_4 = \beta(\mathcal{A}_4) = \{4\}$. In step 3.3, for $i = 1, 2, 3$, accumulate $\beta(\mathcal{A}_1)$, $\beta(\mathcal{A}_2)$, $\beta(\mathcal{A}_3)$ into a_4, giving $a_4 = \{1,2,3,4,5,6\}$, and so $\mathcal{Q}_4$ is full. For $j = 5$, as $s_4 \neq s_3$ in step 3.1, update $a_2 = \{2,6\}$ and $a_3 = \{3,5\}$. In step 3.2, $a_5 = \beta(\mathcal{A}_5) - s_4 = \{3,5\}$. This remains unchanged in step 3.3 because now $a_5 \cap a_2 = \emptyset$ and $a_5 \cap a_3 = a_5$. Thus step 3.4 gives $\beta(\mathcal{Q}_5) = a_5 = \{3,5\}$. The entire pattern $\mathcal{Q}$ is given by

$$\mathcal{Q} = \begin{bmatrix} * & * & * & * & 0 \\ 0 & * & * & * & 0 \\ 0 & 0 & * & * & * \\ * & * & * & * & 0 \\ 0 & 0 & * & * & * \\ 0 & * & * & * & 0 \end{bmatrix}.$$

We now summarize a graph theoretic way to compute the Hall sets required by step 1 of algorithm 4.10. Firstly, a maximal transversal in the m-by-n pattern $\mathcal{A}$ is obtained; see, e.g. [LP, p. 15] for an algorithm, or [D] for the case $m = n$. Without loss of generality, this maximum matching gives a square n-by-n pattern in which all diagonal entries are *. Secondly, for the digraph associated with this square pattern, find the strongly connected components using an algorithm of Tarjan (see, e.g. [RND]), and use these to construct its condensation digraph. The Hall sets of maximal cardinality can

be determined from this condensation digraph by traversing paths starting at vertices of indegree zero.

The time complexity of this procedure to determine the Hall sets is dominated by the computation of a maximal transversal, and is $O(\tau(m+n)^{\frac{1}{2}})$, where τ is the number of nonzero entries in $\mathcal{A}$ (see, e.g. [LP]). Step 3 of algorithm 4.10 has time complexity $O((h+1)mn^2)$ where h is the number of distinct (nonempty) maximal Hall sets S_j, $1 \leq j \leq n-1$.

5. Pattern $\mathcal{R}$.

Given a pattern $\mathcal{A}$, theorem 4.7 determines the pattern $\mathcal{Q}$, and we now consider the pattern $\mathcal{R}$. As noted in the introduction, the pattern $\mathcal{R}$ is invariant under row permutations of $\mathcal{A}$.

Theorem 5.1. For $m \geq n$, let $\mathcal{A}$ be an m-by-n pattern with the Hall property. If the pattern $\mathcal{Q}$ is determined as in theorem 4.7, then the pattern $\mathcal{R}$ is given by the upper triangular part of $\mathcal{Q}^T\mathcal{A}$.

Proof: Without loss of generality, we can assume that each diagonal position of $\mathcal{A}$ is *. Since $i \in p_i$, $1 \leq i \leq n$, the (i,i) entry of $\mathcal{Q}$ is * by theorem 4.7. Thus, since the (i, i) entry of $\mathcal{A}$ is *, the (i, i) entry of $\mathcal{Q}^T\mathcal{A}$ is *. The diagonal entries of $\mathcal{R}$ are all * (by definition of the unique QR factorization), thus the patterns of $\mathcal{Q}^T\mathcal{A}$ and $\mathcal{R}$ agree on the diagonal.

Now consider in $\mathcal{R}$ a fixed (i, j) entry with $i < j \leq n$. If $\mathcal{Q}_i$ is combinatorially orthogonal to $\mathcal{A}_j$, then $\mathcal{Q}_i^T\mathcal{A}_j = 0$. Also, for any $A \in \mathcal{A}$, if $A = QR$, then $R = Q^TA$ and so $r_{ij} = Q_i^TA_j = 0$ in this case of combinatorial orthogonality.

If $\mathcal{Q}_i$ is *not* combinatorially orthogonal to $\mathcal{A}_j$, then $\mathcal{Q}_i^T\mathcal{A}_j$ is *. We must show that there exists a matrix $A \in \mathcal{A}$ such that if $A = QR$, then $r_{ij} \neq 0$. Suppose that the (k,j) entry of $\mathcal{A}$ is * for some k, $1 \leq k \leq m$, and the (k, i) entry of $\mathcal{Q}$ is *. Then $k \in p_i$, and by the construction in theorem 4.4, the submatrix in the first i columns of A can be determined so that in its QR factorization, the entry $q_{ki} \neq 0$. (Note that in the QR factorization of any $A \in \mathcal{A}$, Q_i is completely determined by the first i columns of A.) Since $\mathcal{Q}_i$ and $\mathcal{A}_j$ are not combinatorially

orthogonal, we may choose $A_j \in \mathcal{A}_j$ so that A_j is not orthogonal to the already determined Q_i. Thus $r_{ij} = Q_i^T A_j$ is nonzero, completing the proof. (Note that only columns $A_1, \ldots, A_i$ and A_j have been specified; the other columns of A are arbitrary subject to $A \in \mathcal{A}$.) ■

Returning again to example (1.1), we see that theorem 4.7 predicts the true pattern for $\mathcal{Q}$, that is as in (1.2) with $\otimes = 0$. Theorem 5.1 then predicts the correct pattern for $\mathcal{R}$, again (1.2) with $\otimes = 0$. For example, the (4, 6) entry of $\mathcal{R}$ is zero, as $\mathcal{Q}_4$ is combinatorially orthogonal to $\mathcal{A}_6$.

The (i,j) entry of $\mathcal{R}$ is * if and only if there is an index k such that $a_{kj} \neq 0$, and a path in $B_i(\mathcal{A})$ from column i to row k (passing only through columns numbered less than i). If $\mathcal{A}$ has the strong Hall property and $A \in \mathcal{A}$, this is equivalent (ignoring cancellation) to the existence of a path in the graph of A^TA from i to j passing only through vertices less than i. As the existence of such a path is equivalent to a generically nonzero entry in the (i,j) position of U in the LU factorization of A^TA, our result for $\mathcal{R}$, specialized to the strong Hall case, yields the same result as that of [CEG] mentioned in our introduction. We note that, since the columns of $\mathcal{Q}$ are determined sequentially (without appeal to previously determined columns), the rows of $\mathcal{R}$ may be determined serially. Thus, if only the pattern of $\mathcal{R}$ is of interest, it may be determined without necessity for storage of $\mathcal{Q}$. In fact, if only row i of $\mathcal{R}$ is desired, it may be determined from $\mathcal{Q}_i$ (and $\mathcal{A}$) for which only the Hall set S_{i-1} and $\mathcal{B}_i(\mathcal{A})$ are necessary.

Remarks:

We thank F. Ruskey, M. Gillespie and the referees for useful comments. We also note that lemma 2.5 of [Gi] (which was brought to our attention after preparation of this manuscript) is closely related in substance to our theorem 3.2. The two proofs are quite different, and the underlying concept seems to be an important one for exhibiting nonzeros in factorizations of sparse matrices. The report [Gi] does not consider strong Hall critical matrices, as in our theorem 3.1.

In addition, A. Pothen [private communication in response to our manuscript] proposed an alternate proof of our theorem 4.4, based on the Dulmage–Mendelsohn decomposition and a statement equivalent to lemma 2.5 of [Gi]. Pothen also notes that the DM decomposition can be used to construct the required Hall sets, and notes (via an argument based on algebraic indeterminates) that the collection of all entries that may individually be nonzero in Q may simultaneously be nonzero for some (in fact, almost all) allowed instances of A.

References

[CEG] T. Coleman, A. Edenbrandt and J. Gilbert, Predicting fill for sparse orthogonal factorization, J. Assoc. for Comp. Mach. 33(1986), 517–532.

[D] I.S. Duff, On algorithms for obtaining a maximum transversal, ACM Transactions on Mathematical Software 7(1981), 315–330.

[Ge] W.M. Gentleman, Row elimination for solving sparse linear systems and least squares problems, in Proc. of the 1975 Dundee Conference on Numerical Analysis, G.A. Watson (ed.), Lecture Notes in Mathematics 506, Springer–Verlag, New York, 1976, 122–133.

[GH] A. George and M. Heath, Solution of sparse linear least squares problems using Givens rotations, LAA 34(1980), 69–83.

[GLN] A. George, J. Liu, and E. Ng, Row–ordering schemes for sparse Givens transformations I: Bipartite Graph Model, LAA 61(1984), 55–81.

[GN] A. George and E. Ng, Symbolic factorization for sparse Gaussian elimination with partial pivoting, SIAM J. Sci. Stat. Comput. 8(1987), 877–898.

[Gi] J.R. Gilbert, An efficient parallel sparse partial pivoting algorithm, Chr. Michelsen Inst. Dept. of Science and Technology Report 88/45052–1, Bergen, Norway (1988).

[HJ] R. Horn and C.R. Johnson, Matrix Analysis, Cambridge Univ. Press, N.Y., 1985.

[JOD] C.R. Johnson, D.D. Olesky and P. van den Driessche, Inherited matrix entries: LU factorizations, SIAM J. of Matrix Analysis 10(1989), 94–104.

[LP] L. Lovász and M.D. Plummer, Matching Theory, North–Holland, Amsterdam, 1986.

[RND] E.M. Reingold, J. Nievergelt and N. Deo, Combinatorial Algorithms: Theory and Practice, Prentice–Hall, Inc., New Jersey, 1977.

[S] G.W. Stewart, Perturbation bounds for the QR factorization of a matrix. SIAM J. Numer. Anal. 14(1977), 509–518.