Matroidal Approach to the Structural Solvability of a System of Equations

Kazuo Murota* and Masao Iri*

August 1982

Abstract

Graph— or matroid—theoretic approach has been successfully applied to the design and analysis of large—scale systems, usually under certain "generality" assumptions. In this paper, the structural solvability of a system of equations is discussed under less restrictive and more realistic assumptions on the "generality" of the quantities involved. A necessary and sufficient condition for the structural solvability is stated in terms of the rank of a union matroid associated with the system of equations. Also given is the detailed description of an efficient algorithm for testing the structural solvability as well as for detecting the structural inconsistencies.

^{*} Department of Mathematical Engineering and Instrumentation Physics, Faculty of Engineering, University of Tokyo, Hongo, Tokyo 113, Japan

1. Introduction

Combinatorial approach has proven to be useful in the design and analysis of large-scale systems, especially in detecting structural inconsistencies and in decomposing the whole system into subsystems [10], [11]. For example, the solvability of a linear electric network is discussed in terms of combinatorial concepts connected with graphs and matroids, where a certain "generality" is usually imposed on the quantities involved in the network [12], [21], [22].

A number of graph-theoretic techniques have been successfully incorporated into chemical process simulators such as JUSE GIFS and DPS, developed in Japan [8], [14], [24], [28], [31], [32]. The ideas of applying graph-theoretic techniques to the solution of systems of equations are found in [1], [5], [6], [7], [15], [25], [26]. The mathematical basis for the techniques employed in JUSE GIFS and DPS for testing the structural solvability of a system of linear/nonlinear equations, as well as for decomposing the whole system into hierarchical subsystems, is discussed in [13], [14], [16], [17], [18] by means of graph theory under a certain "generality assumption" on the functional forms of the equations. However, the "generality assumption" is too stringent for some practical problems, so that the graph-theoretic techniques developed under the assumption are sometimes not justified in the strict sense of the word; in fact, a number of real problems have been encountered for which the graph-theoretic techniques fail to detect the structural inconsistencies.

In this paper, the structural solvability of a system of linear/nonlinear equations is formulated in a less restrictive and more realistic setting, i.e., under a weaker "generality assumption" on the

functions in the system and a necessary and sufficient condition for the structural solvability is given in terms of the rank of the union of two matroids associated with the system of equations. An efficient algorithm for testing the structural solvability is presented together with examples.

2. Structural Solvability of a System of Equations

We consider a system of equations in the following form, with unknowns $x_j \ (j=1,\dots,N) \ \text{and} \ u_k \ (k=1,\dots,K), \ \text{and parameters} \ y_i \ (i=1,\dots,M):$

$$\begin{cases} y_{i} = f_{i}(x, u) & (i=1,...,M), \\ u_{k} = g_{k}(x, u) & (k=1,...,K), \end{cases}$$
 (2.1)

where f_i (i=1,...,M) and g_k (k=1,...,K) are sufficiently smooth real-valued functions. This form is most natural and convenient when we treat a physical/engineering system represented by a set of functional relations among elemental state variables, where we want to adjust the values of x-and u-variables so as to meet the arbitrarily given values of y-variables [14].

We are concerned with whether the system (2.1) of equations <u>has a structure which admits a unique solution</u>. In the following we assume that M = N, since the number of equations must usually be equal to the number of unknowns in order for (2.1) to have a unique solution. We denote Jacobian matrix of (2.1) with respect to x and u by

$$J(x,u) = \begin{pmatrix} J(f,x) & J(f,u) \\ J(g,x) & J(g,u)-I_K \end{pmatrix}, \qquad (2.2)$$

where

$$J(f,x) = \left(\frac{\partial f_{i}}{\partial x_{j}}\right), \quad J(f,u) = \left(\frac{\partial f_{i}}{\partial u_{1}}\right),$$

$$J(g,x) = \left(\frac{\partial g_{k}}{\partial x_{j}}\right), \quad J(g,u) = \left(\frac{\partial g_{k}}{\partial u_{1}}\right).$$

Suppose that (2.1) has a solution (x, u) = (x, u) for some y = y. It follows from the theorem on implicit function that, if

$$\det J(x,u) \neq 0, \tag{2.3}$$

(2.1) has a unique solution (x, u) around (\hat{x}, \hat{u}) in accordance with an arbitrary perturbation of y in a neighborhood of \hat{y} . It should be noted also that, from a computational point of view, the condition (2.3) guarantees the feasibility of a Newton-like iterative method for the numerical solution of (2.1).

The above condition (2.3), however, depends not only on the functional forms of f_i and g_k but also on particular values of (\hat{x}, \hat{u}) , which are usually not known in advance. Hence we will consider a condition for the solvability of (2.1), i.e., the condition that the Jacobian, as a function in x_j (j=1,...,N) and u_k (k=1,...,K), does not vanish identically:

$$\det J(x,u) \neq 0. \tag{2.4}$$

More precisely, we shall assume that the partial derivatives of functions f_i and g_k can be regarded as elements of some extension field \underline{F} of the rational number field \underline{Q} . This assumption is fulfilled, for example, if f_i and g_k are functions rational in x_j (j=1,...,N) and u_1 (l=1,...,K), in which case the field of rational functions in x_j (j=1,...,N) and u_1 (l=1,...,K) may be taken as the field \underline{F} . We shall say that the system (2.1) of equations is structurally solvable if the Jacobian matrix J(x,u) of (2.1), as a matrix over \underline{F} , is nonsingular, i.e., if (2.4) holds in \underline{F} . In the following, we denote by D the set of partial derivatives of f_i and

g_k:

In [13], the structural solvability of (2.1) is discussed under the following assumption on the "generality" of the functional forms of \mathbf{f}_i and \mathbf{g}_k :

GA1: The nonvanishing elements of D are algebraically independent [29] over the rational number field Q.

Under this assumption, the structural solvability (2.4) of (2.1) is reduced to a condition on the "representation" graph of (2.1), which is a directed graph, or a kind of signal-flow graph, with vertices corresponding to variables x_j (j=1,...,N), u_k (k=1,...,K) and y_i (i=1,...,M) and arcs representing the functional dependences among the variables explicitly seen in f_i and g_k . That is, we have the following theorem [13].

Theorem 2.1. The system (2.1) of equations is structurally solvable iff there exists on the representation graph a Menger-type vertex-disjoint complete linking from $X=\{x_j | j=1,...,N\}$ to $Y=\{y_j | i=1,...,M\}$.

This result provides a mathematical basis for the graphical techniques incorporated in chemical process simulators JUSE GIFS and DPS, which have proven their effectiveness in industrial applications [24], [28], [31], [32]. The "generality" assumption GA1 is, however, sometimes too stringent in actual situations.

To meet realistic situations, we have to mitigate the generality assumption. Specifically, we shall consider the rank of the Jacobian

matrix, when the elements of D are classified into two subsets, \tilde{T} and $\tilde{D}\setminus \tilde{T}$, in such a way that \tilde{T} is algebraically independent over the field $\underline{Q}(\tilde{D}\setminus \tilde{T})$, the adjunction of $\tilde{D}\setminus \tilde{T}$ to \underline{Q} [29]. By so doing, we can treat more realistic situations than we could under the assumption GA1, since GA1 obtains as a particular case where $\tilde{T}=D$. In the next section, we establish Theorem 3.1, which is a basic lemma concerning the rank of a matrix consisting of two kinds of entries as above.

3. Rank of a Matrix

3.1. Additive partition of a matrix

Let \underline{K} be a field and \underline{F} an extension field of \underline{K} . Consider an m by n matrix $A = (a_{ij})$ over \underline{F} . We shall denote by A also the set of entries $\{a_{ij} | i=1,\ldots,m; j=1,\ldots,n\}$ of the matrix A. Suppose the matrix A is expressed as the sum of two matrices $T = (t_{ij})$ over \underline{F} and $Q = (q_{ij})$ over \underline{K} : $A = T + Q, \tag{3.1}$

where the nonvanishing entries of T are algebraically independent over \underline{K} . As is readily seen, if we define

 $T_0 = \{a_{ij} | a_{ij} \text{ is a transcendental over } \underline{K}(A \setminus \{a_{ij}\})\}$, then $t_{ij} \neq 0$ implies $a_{ij} \in T_0$. Conversely, when given a matrix A over \underline{F} , we can get an additive partition (3.1) of A by setting

$$t_{ij} = a_{ij}, q_{ij} = 0$$
 if $a_{ij} \in T_0$,
 $t_{ij} = 0, q_{ij} = a_{ij}$ otherwise.

Example 3.1. Let $\underline{K} = \underline{Q}$ and $\underline{F} = \underline{Q}(x,u,e^{X})$. For

$$A = \begin{pmatrix} 2^{1/2} & x^3 + 1 & u + x \\ x^2 & 1 & e \\ 0 & e^x & 2 \end{pmatrix}$$

we have partition (3.1) with

$$T = \begin{bmatrix} 0 & 0 & u + x \\ 0 & 0 & e \\ 0 & e^{x} & 0 \end{bmatrix} \quad \text{and} \quad Q = \begin{bmatrix} 2^{1/2} & x^{3} + 1 & 0 \\ x^{2} & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix}.$$

Let $R = \{1, ..., m\}$ and $C = \{1, ..., n\}$ be the column set and the row set, respectively. For $I \subseteq R$ and $J \subseteq C$, we denote by A(I, J) the submatrix of A with rows I and columns J. The rank and the term-rank of A are denoted by r(A) and t(A), respectively. The following identity plays an important role in the subsequent arguments.

Theorem 3.1 (Rank Identity). For a matrix A of the form (3.1), we have

$$r(A) = m a x \{t(T(I,J))+r(Q(R\setminus I,C\setminus J))\}.$$

$$I \subseteq R, J \subseteq C$$
(3.2)

Proof: First we show the more obvious inequality:

 $r(A) \leq \text{m a x } \{t(T(I,J)) + r(Q(R\setminus I,C\setminus J))\}.$ $I^{\subset}R,J^{\subset}C$ Take a nonsingular submatrix A(R',C') = T(R',C') + Q(R',C') of A with |R'| = |C'| = r(A). Since det A(R',C'), viewed as a polynomial in nonvanishing $t_{ij}\text{'s over }\underline{K}, \text{ does not vanish, there exists a nonvanishing term, say}$ $\Pi_{i\in I}t_{i,j(i)}\text{ with }I^{\subset}R', \text{ whose coefficient is equal, up to a sign, to }$ det $Q(R'\setminus I,C'\setminus J)$ with J=j(I). Hence, we have t(T(I,J)) = |I| and

 \leq t(T(I,J))+r(Q(R\I,C\J)). Note that the algebraic independence of t_{i,i}'s

 $r(Q(R\backslash I,C\backslash J)) > r(Q(R'\backslash I,C'\backslash J)) = |R'|-|I| = r(A)-|I|$. Thus we obtain r(A)

is not used in this part.

Next we show

$$r(A) \ge \max_{I \subseteq R, J \subseteq C} t(T(I,J)) + r(Q(R\setminus I,C\setminus J))$$

Take (I,J) which attains the maximum on the right-hand side. Then there exists a square nonsingular submatrix $T(I^{\dagger},J^{\dagger})$ of T(I,J) such that

$$t(T(I',J')) = |I'| = |J'| = t(T(I,J)).$$

Similarly there exists a square nonsingular submatrix Q(I'',J'') of $Q(R\setminus I,C\setminus J)$ such that

$$r(Q(I'',J'')) = |I''| = |J''| = r(Q(R\setminus I,C\setminus J)).$$

The submatrix A(I'+I'',J'+J'') is nonsingular, since the generalized Laplace expansion of det A(I'+I'',J'+J'') with respect to rows I' contains a term det Q(I'',J'') det T(I',J'),

which cannot be cancelled out by virtue of the algebraic independence of the nonvanishing entries of T over K. Therefore we obtain

$$r(A) \ge r(A(I'+I'',J'+J''))$$

= $|I'| + |I''|$
= $t(T(I,J)) + r(Q(R\setminus I,C\setminus J))$. Q.E.D.

3.2. Matroids associated with the additive partition (3.1)

Theorem 3.1 may be rephrased in terms of linking systems [23] that the linking system defined by matrix A is the union of the two linking systems defined by matrices T and Q. It may also be remarked that (3.2) can be regarded as an extension of the 2-block rank introduced in [9].

With matrix A of the form (3.1), we will associate two matroids on S = R+C, which are nothing but the matroids corresponding to the linking systems defined by T and Q, respectively [23]. To be specific, we define two functions τ and $\rho:2^S--->Z$ by

$$\tau(I+J) = t(T(R\setminus I,J)) + |I|, \tag{3.3}$$

$$\rho(I+J) = r(Q(R\setminus I, J)) + |I|,$$
(3.4)

where I \subset R and J \subset C. Since τ and ρ enjoy the properties of rank functions of matroids [30], we denote the corresponding matroids by M(τ) and M(ρ), respectively. Then it follows from (3.2), (3.3) and (3.4) that

$$r(A) = \max \left\{ \tau(S') + \rho(S\backslash S') \right\} - |R|,$$

$$S' \subseteq S$$

and by the definition of a union matroid, we obtain

$$r(A) = (\tau \lor \rho)(S) - m$$

where $M(\tau^{\vee}\rho)$ is the union matroid of $M(\tau)$ and $M(\rho)$, and m=|R|. On recalling the well-known relation between the rank of the union matroid and the maximum size of a common independent set, we obtain the following theorem, where $M(\rho^*)$ is the dual of matroid $M(\rho)$.

Theorem 3.2. For a matrix A of the form (3.1), we have $r(A) = \operatorname{rank} \text{ of } M(\tau \vee \rho) - m$ $= \operatorname{maximum} \text{ size of a common independent set of } M(\tau) \text{ and } M(\rho^*).$

Remark 3.1. The dual matroid M(ρ *) of M(ρ) corresponds to the transpose of the matrix Q. That is,

$$\rho^*(I+J) = r(Q(I,C\setminus J)) + |J|.$$

Since, as described later, $M(\tau)$ can be represented by a bipartite graph, and $M(\rho)$ as well as $M(\rho^*)$ by a matrix over \underline{K} , we can compute the rank of a matrix of the form (3.1), on the basis of Theorem 3.2, by utilizing any established efficient algorithm for finding the union of $M(\tau)$ and $M(\rho)$, or for finding a maximum common independent set of $M(\rho)$ and $M(\rho^*)$. The significance of Theorem 3.2 lies in the fact that it

enables us to determine the rank of a matrix A over \underline{F} with graph manipulations as well as arithmetic operations in the subfield \underline{K} , not in the extension field \underline{F} .

In connection with the Rank Identity (3.2), the following holds.

Theorem 3.3. For a maximizer (I,J) of (3.2) which is maximal with respect to set inclusion, we have $|R\setminus I| = |C\setminus J|$ and

A CHILL FOR A SHIP TO BE A SHIP OF A SHIP OF A SHIP

det $A(R\setminus I, C\setminus J) \in K^*$ $(K^*=K-\{0\}).$

<u>Proof:</u> Put $\overline{I}=R\setminus I$ and $\overline{J}=C\setminus J$. Suppose $|\overline{I}|>r(Q(\overline{I},\overline{J}))$. Then there exists i in \overline{I} such that $(I+\{i\},J)$ is also a maximizer of (3.2), which contradicts the assumption that (I,J) is maximal. Similarly for $|\overline{J}|$. Hence $|\overline{I}|=|\overline{J}|=r(Q(\overline{I},\overline{J}))$, that is, $Q(\overline{I},\overline{J})$ is nonsingular, and a fortiori $A(\overline{I},\overline{J})$ is nonsingular, i.e., det $A(\overline{I},\overline{J})\neq 0$.

Suppose det $A(\overline{I},\overline{J}) \notin \underline{K}$. Then there exist $I' \subseteq \overline{I}$ and $J' \subseteq \overline{J}$ ($I' \neq \emptyset$, $J' \neq \emptyset$) such that both T(I',J') and $Q(\overline{I}\setminus I',\overline{J}\setminus J')$ are nonsingular, which implies that (I+I',J+J') is also a maximizer of (3.2), a contradiction. Q.E.D.

It is known [19] that if det $A \in \underline{K}^*$ for a square matrix A over \underline{F} of the form (3.1), the matrix A, with suitable permutations of its rows and columns, can be decomposed into LU-factors with a unit lower triangular matrix over \underline{F} and a nonsingular upper triangular matrix over \underline{K} . This means, in practical situations, that the submatrix $A(R\setminus I,C\setminus J)$ in Theorem 3.3 has one and the same U-factor whatever values $t_{i,j}$'s may take.

4. Condition for the Structural Solvability

4.1. Structural solvability under generality assumptions

Suppose the partial derivatives D of f_i and g_k in (2.1) are classified into two subsets, \tilde{T} and $D\setminus \tilde{T}$, in such a way that \tilde{T} is algebraically independent over $Q(D\setminus \tilde{T})$ (see section 2). Accordingly, we express the Jacobian matrix (2.2) as

$$\mathbf{J} = \mathbf{T} + \mathbf{Q}_{\bullet}$$

where the partial derivatives belonging to T are taken for the nonvanishing entries of T. This expression is of the form (3.1) with $\underline{K}=\underline{Q}(D\setminus T)$ in the previous section. Let us denote by $M(\tau)$ and $M(\rho)$, respectively, the matroids defined by (3.3) and (3.4). From Theorem 3.2 with A=J, we immediately obtain a necessary and sufficient condition for the structural solvability (2.4) of the system (2.1) of equations, as stated below.

Theorem 4.1. Suppose that M = N in (2.1). The following three conditions are equivalent.

- (i) The system (2.1) of equations is structurally solvable.
- (ii) The union matroid $M(T \lor \rho)$ has the rank 2(N+K).
- (iii) The maximum size of a common independent set of M(τ) and M(ρ^*) is equal to N+K.

In case we can assume

<u>GA2:</u> Those elements of D which do not belong to the rational number field \underline{Q} are algebraically independent over \underline{Q} ,

on the generality of the partial derivatives D, we may take T to be the set of all the partial derivatives which are not rational constants. Then we

have $\underline{K}=\underline{Q}(D\setminus T)=\underline{Q}$, which implies that the structural solvability can be determined by arithmetic operations on rational numbers as well as by graph manipulations.

Another useful assumption is that

GA3: Those elements of D which do not belong to the real number field \underline{R} are algebraically independent over \underline{R} .

In this case, we can take for \tilde{T} the set of non-constant derivatives and put $\underline{K=R}$, and consequently, the structural solvability is determined by arithmetic operations on real numbers as well as graph manipulations. The generality assumption GA3 reflects the situation where we regard the nonlinear part of the functions f_i and g_k as independent.

Some comments would be in order here on the mutual relations among the generality assumptions GA1, GA2 and GA3. First of all, GA2 is weaker than GA1; that is, if GA1 holds, GA2 holds, too. No other implications exist, as illustrated below, where $\underline{F} = \underline{R}(x, e^X)$ and \overline{T}_0 is determined with $\underline{K} = \underline{Q}$.

	GA1	GA2	GA3	T ₀
$\{e^{x}, x, e^{\sqrt{2}}, e^{\sqrt{3}}\}$	о.к.	o.ĸ.	O.K.	$\{e^{x}, x, e^{\sqrt{2}}, e^{\sqrt{3}}\}$
$\{e^{x}, x, e^{\sqrt{2}}x, e^{\sqrt{3}}\}$	о.к.	0.K.	X	$\{e^{x}, x, e^{\sqrt{2}}, e^{\sqrt{3}}\}$
{e ^X ,x,π,1}	х	O.K.	O.K.	{e ^x ,x,π}
$\{x, \pi, \pi^2, 2^{1/2}\}$	х	x	O.K.	{x}
$\{e^{X}, x, \pi x, 1\}$	х	О.К.	X	{e ^X ,x,πx}
{e ^X ,x, πx, π}	x	X	X	{e ^X }

When given a system of equations, we can test the structural solvability by applying Theorem 4.1, once we classify the elements of D into two subsets T and D \backslash T. In actual applications to real problems, the

classification is in some sense at our disposal, or in other words, it depends on how we model the system, as will be seen in the example below.

4.2. Example --- Ethylene dichloride production system

Consider a hypothetical system (Fig. 1) for the production of ethylene dichloride (${^{\rm C}_2}{^{\rm H}_4}{^{\rm Cl}_2}$), which is slightly modified from the one given in "Users Manual of Generalized Interrelated Flow Simulation" of "The Service Bureau Co."

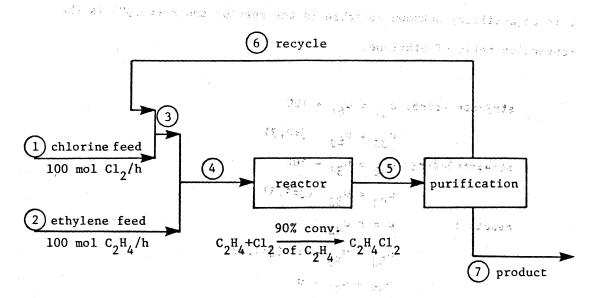
Feeds to the system are 100 mol/h of pure chlorine (${\rm Cl}_2$) (stream 1), and 100 mol/h of pure ethylene (${\rm C}_2{\rm H}_4$) (stream 2). In the reactor, 90% of the ethylene is converted into ethylene dichloride according to the reaction formula

In the purification stage, the product ethylene dichloride is recovered and the unreacted chlorine and ethylene are separated for recycle. The purification is described in terms of component recovery ratios a_1 , a_2 and a_3 of chlorine, ethylene and ethylene dichloride, respectively. These are the fractions of components in the unit feed (stream 5) that are recovered in stream 6.

The problem considered here is as follows.

<u>Problem:</u> Given the component recoveries a_1 and a_2 of chlorine and ethylene, determine the recovery $x=a_3$ of ethylene dichloride with which a specified value y mol/h of ethylene dichloride is produced.

Let u_{i1} , u_{i2} and u_{i3} mol/h be the component flow rates of chlorine, ethylene and ethylene dichloride in stream i, respectively. The system of



errent (8) is an empty appropriate and in any set for nearly at his entrance

Fig. 1. Hypothetical ethylene dichloride production system

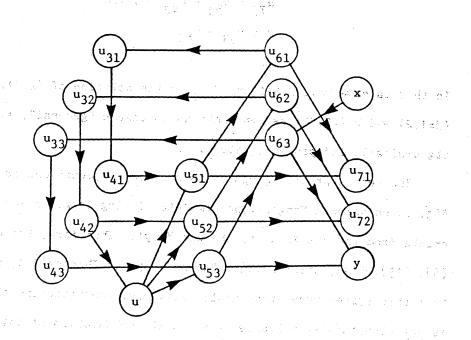


Fig. 2. Representation graph of the example problem

equations to be solved may be put in the following standard form [8], where u is an auxiliary unknown variable in the reactor and r (=0.90) is the conversion ratio of ethylene.

str3=str1+str6:
$$u_{31} = u_{61} + 100$$

 $u_{3j} = u_{6j}$ (j=2,3)
str4=str2+str3: $u_{42} = u_{32} + 100$
 $u_{4j} = u_{3j}$ (j=1,3)
reactor: $u = r u_{42}$
 $u_{5j} = u_{4j} - u$ (j=1,2)
 $u_{53} = u_{43} + u$
purification: $u_{6j} = a_{j} u_{5j}$ (j=1,2)
 $u_{63} = x u_{53}$
 $u_{7j} = u_{5j} - u_{6j}$ (j=1,2)
 $y = u_{53} - u_{63}$

In this case we have M=N=1 and K=15 in the notation of (2.1). We regard a j (j=1,2) and r (=0.90) as constants which assume "general", i.e., algebraically independent, values.

The representation graph of this system of equations is given in Fig. 2, on which a Menger-type complete linking (e.g., $x \rightarrow u_{63} \rightarrow y$) exists from the set $X=\{x\}$ to the set $Y=\{y\}$. Therefore, the existing method [8], [13], [14], which assumes GA1 and applies Theorem 2.1, would conclude that this system were structurally solvable, contradicting the fact that it is <u>not</u> structurally solvable, i.e., that the Jacobian of this system (Fig. 3) identically vanishes. This contradiction stems from the assumption GA1, which, in this case, fails to hold. In fact a large part

of the equations are made up of simple additions and subtractions so that most of the partial derivatives are constant of values such as 1 and -1, which are too special to be considered "general"; consequently, the Jacobian matrix is singular in spite of the fact that it has term-rank of N+K=16, or equivalently [13], that the representation graph has a Menger-type complete linking.

On the other hand, the new method based on Theorem 4.1 can detect the structural inconsistency of this system of equations. We consider a_1 , a_2 , r, x and u_{53} to be algebraically independent over \underline{Q} and, putting $\overline{T}=\{a_1,a_2,r,x,u_{53}\}$, express the Jacobian matrix in the form of (3.1). The maximum size of a common independent set of $M(\tau)$ and $M(\rho^*)$, which can be found by the algorithm given in the next section, turns out to be equal to 15. It follows from Theorem 4.1 that the system of equations is not structurally solvable.

: . !	x	~ ^u 31 ^{-u} 32 ^{-u} 33 ^{-u} 41 ^{-u} 42	43 ^u 51 ^u 52 ^u 53 ^u 61 ^u 62 ^u 63 ^u 71 ^u 72 ^u 72
y	1.79	The date upperfixe to	and the control of the property of the second
u ₃₁		-1 -1 -3 -3 -3 -4 -4 -5 -5 -5 -5 -5 -5 -5 -5 -5 -5 -5 -5 -5	
^u 32	,	-1 7%+1 / 128 to	
^u 33		-1 I	in the first of the second of
^u 41 ^u 42	sed vi	1 -1	
u ₄₃			
^u 51			The state of the s
u ₅₂			1 * * * * * * * * * * * * * * * * * * *
^u 53 ^u 61	1, 1,		
u ₆₂			a ₂ -1
^u 63	^u 53	1	x -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1
u ₇₁			1 –1 –1
u ₇₂			1 -1 ,-1
u		r !	_1

Fig. 3. Jacobian matrix of the example problem

5. Algorithm for Testing the Structural Solvability

5.1. Representation of the matroids M(τ) and M(ρ^*)

In this section, we will describe an algorithm for determining the rank of an m by n matrix A = T + Q of the form (3.1) on the basis of Theorem 3.2. For convenience, we put the column set $C=\{1,\ldots,n\}$ and the row set $R=\{n+1,\ldots,n+m\}$ and $S=C+R=\{1,\ldots,n+m\}$, and understand throughout that I and J denote a subset of R and C, respectively. Two functions, i.e., the closure cl^Q and the fundamental circuit C^Q are defined in $M(\rho^*)$:

 ${\rm cl}^Q(X) = \{v | \rho^*(X + \{v\}) = \rho^*(X)\} \quad \text{for } X \subseteq S,$ and for $v \in {\rm cl}^Q(X) \setminus X$ with X independent

 $C^{\mathbb{Q}}(v|X) = \{u|(X+\{v\})\setminus\{u\} \text{ is independent}\}.$

The closure ${\rm cl}^T$ and the fundamental circuit ${\rm C}^T$ in M(${}^\tau$) are defined similarly.

The matroid M(τ) can be represented by a bipartite graph G_T associated with matrix T; the vertex sets are R and C, and there is an edge (i,j) iff t_{ij} $\neq 0$. The subset I+J is independent in M(τ) iff, for some subset J_* of R\I, there exists a complete matching between J_* and J on G_T . With a fixed complete matching between J_* and J for an independent set I+J, is associated the auxiliary directed graph G_T , with vertices R+C, which has an arc (i,j) if the edge (i,j) in G_T is out of the matching and an arc (j,i) if the edge (i,j) is in the matching, where i \in R and j \in C.

The basic functions, cl T and T , of M(T) may be characterized in terms of the reachability on \tilde{G}_T , as is stated in the following propositions, where K = R\(I+J_*) and u-*->v means that v is reachable from u.

<u>Proposition</u> 5.1. For $v \notin I+J$ with I+J independent in $M(\tau)$,

 $v \notin cl^{T}(I+J)$ iff $u-\overset{*}{=}-v$ on \tilde{G}_{T} for some $u \in K$.

<u>Proposition 5.2.</u> For $u \in I+J$, $v \in cl^T(I+J) \setminus (I+J)$ with I+J independent in M(τ), $u \in C^T(v|I+J)$ iff $u-\overset{*}{=}-v$ on \widetilde{G}_T .

<u>Proposition</u> 5.3. For $u \in I+J$, $v \notin I+J$ with I+J independent in $M(\tau)$, u-#->v on G_T implies that $I+J+\{v\}\setminus\{u\}$ is independent.

On the other hand, the matroid $M(\rho^*)$ can be represented by n-dimensional row vectors over \underline{K} . Let q_i be the i-th row vector of $Q(i=n+1,\ldots,n+m)$ and e_j be the j-th row vector of the unit matrix $I_n(j=1,\ldots,n)$. For convenience, we write

$$r_{i} = \begin{cases} e_{i} & (i=1,...,n), \\ q_{i} & (i=n+1,...,n+m). \end{cases}$$

Then I+J is independent in M(ρ *) iff the row vectors $\{r_j | j \in I+J\}$ are linearly independent over \underline{K} .

Let s be a permutation of S={1,...,n+m} such that $\{r_{s(j)}|j=1,...,n\}$ is a base of the n-dimensional vector space, that is, B={s(j)|j=1,...,n} is a base of M(ρ *). Then $r_{s(i)}$'s (i=n+1,...,n+m) can be represented as a linear combination of $r_{s(j)}$'s (j=1,...,n), and we shall write the coefficient matrix of the combination by m by n matrix $P=(p_{ij})$ over \underline{K} such that

$$r_{s(i)} = \sum_{j=1}^{n} p_{ij} r_{s(j)}$$
 (i=n+1,...,n+m).

In terms of the matrix P, cl^Q and C^Q can be characterized as follows. For $s(i) \not\in I+J$,

 $s(i) \in cl^Q(I+J) \ \, \text{iff n<i and p}_{i,j} = 0 \ \, \text{for any j with } s(j) \in B \backslash (I+J).$ For $s(i) \in cl^Q(I+J) \backslash (I+J)$,

$$s(j) \in C^{\mathbb{Q}}(s(i)|I+J)$$
 iff $j \le n$ and $p_{i,j} \ne 0$.

5.2. Description of the algorithm

By specializing the algorithm described in [11], [27], we will find a maximum common independent set of M(τ) and M(ρ *) instead of determining the rank of the union matroid of M(τ) and M(ρ) [4]. To be specific, we consider a maximum independent matching problem on a bipartite graph with the vertex set V_T+V_Q, each of V_T and V_Q being a copy of S, and the edges connecting the vertices of V_T and V_Q corresponding to the same element of S. To V_T is attached the matroid M(τ) and to V_Q the matroid M(ρ *). Let I+J be a common independent set of M(τ) and M(ρ *). The algorithm works with the auxiliary graph G, on which augmenting paths are to be found, and an m by n matrix P=(p_{ij}) (i=n+1,...,n+m; j=1,...,n) along with a permutation s of {1,...,n+m} such that I+J \in B={s(j)|j=1,...,n}.

The auxiliary graph \widetilde{G} is determined as follows. The vertex set of \widetilde{G} is the union of $V_T=R_T+C_T$ and $V_Q=R_Q+C_Q$, where R_T and R_Q (C_T and C_Q , resp.) are copies of R (C_T , resp.). The copies of V_T are denoted by $V_T\in V_T$ and $V_Q\in V_Q$. The subgraph of \widetilde{G} on V_T is the auxiliary graph \widetilde{G}_T , mentioned above, with respect to a matching on G_T . The subgraph of \widetilde{G} on V_Q expresses $M(\rho^*)$ with the help of the matrix P and the permutation S in the following manner: An arc (U_Q,V_Q) ($U_Q,V_Q\in V_Q$) exists iff $P_{ik}=0$ for all K such that $S(K) \not\in I+J$ and $P_{ij} \not= 0$, where U=S(i) and V=S(j).

The algorithm starts with a maximum matching on G_T , the bipartite graph associated with the matrix T. This means that it starts with a common independent set J, the subset of C which is covered by that matching. With the aid of the auxiliary graph \widetilde{G} , the algorithm increases the size of the common independent set I+J one by one. Associated with the auxiliary graph \widetilde{G} is a common independent set I+J:

$$I = \{i \in R | (i_Q, i_T) \text{ exists on } G\},$$

$$J = \{j \in C | (j_Q, j_T) \text{ exists on } G\}.$$

A vertex in R_T is called an entrance vertex if it has no incoming edge on \widetilde{G} . A vertex u_Q in R_Q+C_Q is called an exit vertex if it is not in the closure of the independent set I+J in $M(\rho^*)$, i.e., if

either $u \in \{s(j) | j=1,...,n\} \setminus (I+J)$ or $p_{ij} \neq 0$ for some j with $s(j) \neq I+J$, where u=s(i).

Algorithm

1) s(k):=k for k=1,...,n+m.

P:=Q.

Find a maximum matching on G_T ; set J equal to the subset of C such that the corresponding vertices in C_T are covered by it. Let the arc set of G consist of the following:

$$\begin{split} &\{(\mathbf{i}_T,\mathbf{j}_T)|\,\mathbf{i}\in R,\mathbf{j}\in C;(\mathbf{i}_T,\mathbf{j}_T) \text{ is out of the matching on } G_T\}\\ &+\{(\mathbf{j}_T,\mathbf{i}_T)|\,\mathbf{i}\in R,\mathbf{j}\in C;(\mathbf{i}_T,\mathbf{j}_T) \text{ is in the matching on } G_T\},\\ &\{(\mathbf{i}_T,\mathbf{i}_Q)|\,\mathbf{i}\in R\} \,+\, \{(\mathbf{j}_T,\mathbf{j}_Q)|\,\mathbf{j}\in C\backslash J\} \,+\, \{(\mathbf{j}_Q,\mathbf{j}_T)|\,\mathbf{j}\in J\},\\ &\{(\mathbf{i}_Q,\mathbf{j}_Q)|\,\mathbf{i}\in R,\mathbf{j}\in C;p_{\mathbf{i}k}=0 \text{ for any } k\in C\backslash J \text{ and } p_{\mathbf{i}\mathbf{j}}\neq 0\}. \end{split}$$

- 2) Look for a shortest path \widetilde{P} (i.e., with minimum number of arcs) on \widetilde{G} from an entrance vertex to an exit vertex.
 - i) If such a path P does not exist (including the case where there is no entrance or no exit vertex), stop with I+J being a maximum independent set.
- ii) If such a path P exists, do the following. For all pairs u,v such that the arc (u_Q, v_Q) exists on P, perform the pivoting operation on P with the pivot (k,l), where u=s(k) and v=s(l):

$$\begin{cases} p_{ij} := p_{ij} - p_{i1} p_{kj} / p_{kl} & (i \neq k, j \neq l), \\ p_{kj} := -p_{kj} / p_{kl} & (j \neq l), \\ p_{il} := p_{il} / p_{kl} & (i \neq k), \\ p_{kl} := 1 / p_{kl}, & (i \neq k) \end{cases}$$

Let v_Q be the exit vertex on P and v=s(k). If k>n, then do pivoting above with pivot $p_{kl} \neq 0$ such that $u=s(1) \neq I+J$. Reverse the orientations of all the arcs on \tilde{P} . (Update I+J accordingly.)

Referring to the new matrix P, update the arcs of G connecting two vertices within V_Q , i.e., make an arc (u_Q, v_Q) iff $p_{ik} = 0 \text{ for all } k \text{ with } s(k) \neq I+J \text{ and } p_{ij} \neq 0,$ where u=s(i), v=s(j).

The following example illustrates how the algorithm works.

Example 5.1. Consider a square matrix

$$A = \begin{pmatrix} t_1 & 1 & t_4 & 1 & -1 & 0 & 2 \\ 0 & t_2 & 1 & 1 & 1 & 2 & 0 \\ 0 & 0 & t_3 & t_5 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 1 & 1 \\ 1 & -1 & 0 & 0 & 1 & 1 & -1 \\ 0 & 0 & 1 & 1 & 1 & 2 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 \end{pmatrix},$$

where $\{t_i | i=1,...,5\}$ are algebraically independent over the rational number field \underline{Q} . The partition (3.1) yields A = T + Q with

It is easy to see that r(A) = 6, r(T) = t(T) = 3 and r(Q) = 4.

The algorithm starts with a maximum matching on G_T , e.g., $\{(8,1), (9,2), (10,3)\}$, or $\{t_1, t_2, t_3\}$. This means $I+J=\{1, 2, 3\}$. The initial state of the auxiliary graph $\widetilde{G}^{(0)}$ is given in Fig. 4 (i) together with that of the matrix $P=P^{(0)}=Q$. The entrance is $\{11_T, 12_T, 13_T, 14_T\}$ and the exit is $\{4_Q, 5_Q, 6_Q, 7_Q, 8_Q, 9_Q, 11_Q, 12_Q, 13_Q, 14_Q\}$.

Then an augmenting path $\tilde{P}^{(0)}=11_T$ —>11 $_Q$ is found on $\tilde{G}^{(0)}$ and the pivoting of P with pivot (k,1)=(11,4) is done. (We may choose (11,6) or (11,7) instead of (11,4) as the pivot.) According to the updated matrix $P=P^{(1)}$, we change the arcs connecting two vertices in V_Q , to obtain the auxiliary graph $\tilde{G}^{(1)}$ in Fig. 4 (ii). Now I+J = {1, 2, 3, 11}.

On $\tilde{G}^{(1)}$, an augmenting path $\tilde{P}^{(1)}=12_T-->12_Q$ is found, yielding $\tilde{G}^{(2)}$ in Fig. 4 (iii) with I+J = {1, 2, 3, 11, 12}. The entrance of $\tilde{G}^{(2)}$ is {13 $_T$, 14 $_T$ } and the exit is {4 $_Q$, 5 $_Q$, 6 $_Q$, 7 $_Q$ }. An augmenting path $\tilde{P}^{(2)}=14_T-->14_Q-->3_Q-->3_T-->10_T-->4_T-->4_Q$ is found. For the arc (14 $_Q$,3 $_Q$) and the exit vertex 4 $_Q$, we perform the pivoting, once with (k,1)=(14,3) and once with (k,1)=(11,6) (or (11,7)). The auxiliary graph $\tilde{G}^{(2)}$ is updated to $\tilde{G}^{(3)}$ in Fig. 4 (iv) with I+J = {1, 2, 4, 11, 12, 14}. It turns out that no path exists on $\tilde{G}^{(3)}$ from the entrance {13 $_T$ } to the exit {5 $_Q$, 6 $_Q$, 7 $_Q$ }. Therefore the I+J is a maximum common independent set. It follows from

Theorem 3.2 that r(A) = |I+J| = 6. From the maximum independent set I+J, we also know that $T(R\setminus I,J)$ and $Q(I,C\setminus J)$ gives a partition of matrix A that attains the maximum in the right-hand side of (3.2).

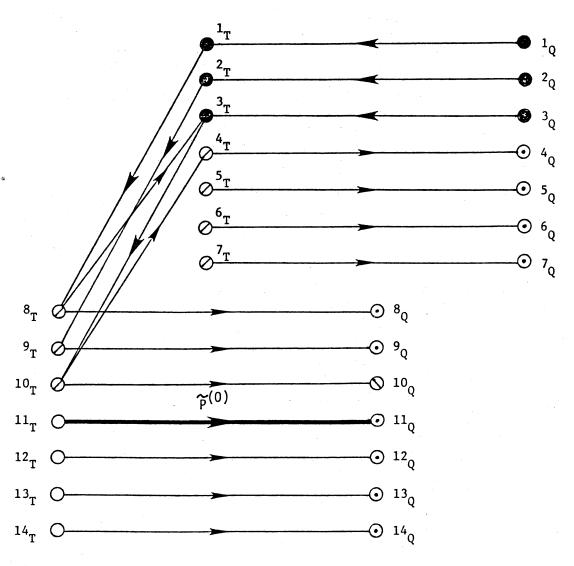


Fig.4 (i). Auxiliary graph $\widetilde{G}^{(0)}$.

O:entrance; : exit

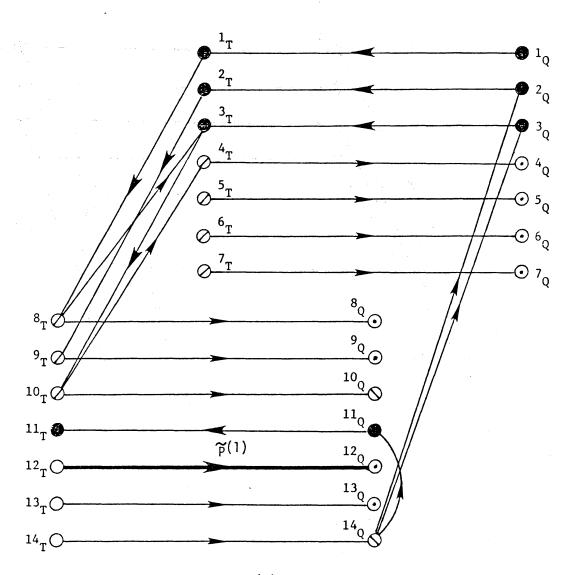


Fig.4 (ii). Auxiliary graph $\widetilde{G}^{(1)}$.

O:entrance :exit

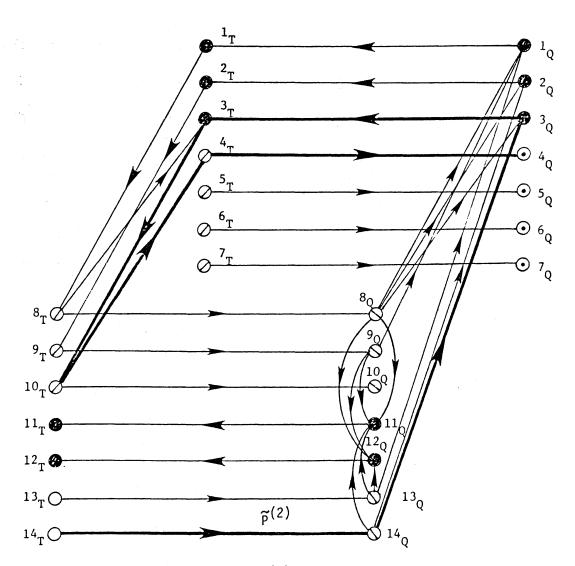


Fig.4 (iii). Auxiliary graph $\tilde{G}^{(2)}$.

○:entrance ; ○:exit

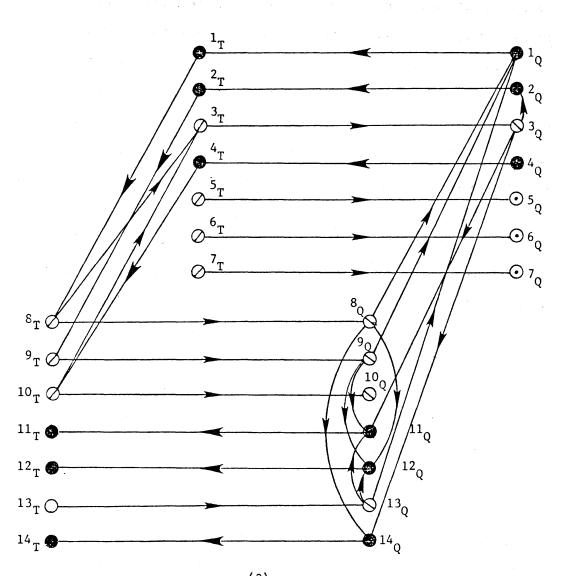


Fig.4 (iv). Auxiliary graph $\tilde{g}^{(3)}$.

○:entrance ; •:exit

6. Concluding Remarks

Graph-theoretic techniques for testing the structural solvability naturally lead to the decompositions of the whole system into hierarchical subsystems [13], [16], [17], [18], [25]. For example, the decomposition of a bipartite graph [2], [3] due to A. L. Dulmage and N. S. Mendelsohn applied to the Jacobian matrix (Fig. 3) of the ethylene dichloride production system yields the block-triangular form shown in Fig. 5. Obviously, the term-rank of this matrix is equal to 16. Once the matrix is put in the block-triangular form, it will not be difficult, in this case, to detect the singularity of the matrix by inspection; in fact, the diagonal block of size 4 in the middle, i.e., the one corresponding to $\{y, u_{33}, u_{43}, u_{53}\} \times \{u_{53}, u_{63}, u_{33}, u_{43}\}$, is easily seen to have rank equal to 3.

On the other hand, the matrix A of Example 5.1 is irreducible by the Dulmage-Mendelsohn decomposition. And it would be essential to resort to the proposed method to determine the rank of this matrix efficiently.

The algorithm, described in section 5, for finding a maximum common independent set of M(τ) and M(ρ *) also gives a partition of R+C with the partial order structure, which is called the principal partition [10], [11], [20] with respect to the pair of matroids (M(τ), M(ρ *)).

We conclude with the following theorem.

Theorem 6.1. Suppose that the matrix A of the form (3.1) is nonsingular. The partition of R+C induced by the principal partition with respect to (M($^{\tau}$), M(ρ *)) is a refinement, inclusive of the partial order, of that induced by the Dulmage-Mendelsohn decomposition of the matrix A.

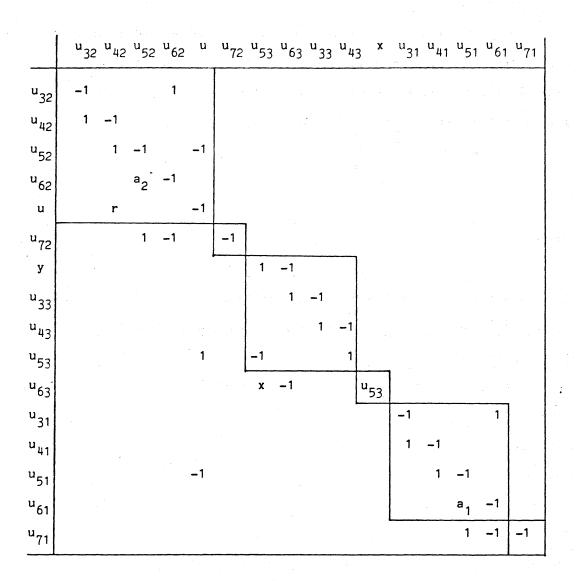


Fig. 5. Block-triangular form of the Jacobian matrix in Fig. 3

Acknowledgement

The authors would like to thank Messrs. J. Tsunekawa and S. Kobayashi of the Institute of the Union of Japanese Scientists and Engineers and S. Abe of Nissan Chemical Industries, Ltd. for useful discussions and suggestions from the practical point of view.

References

- [1] Christensen, J. H., and Rudd, D. F.: Structuring Design Computations.
- A. I. Ch. E. (American Institute of Chemical Engineers) Journal, Vol.15 (1969), pp.94 100.
- [2] Dulmage, A. L., and Mendelsohn, N. S.: A Structure Theory of Bipartite Graphs of Finite Exterior Dimension. <u>Transactions of the Royal Society of</u>
 Canada, Section III, Vol.53 (1959), pp.1 13.
- [3] Dulmage, A. L., and Mendelsohn, N. S.: On the Inversion of Sparse Matrices. Mathematics of Computation, Vol.16 (1962), pp.494 496.
- [4] Edmonds, J.: Minimum Partition of a Matroid into Independent Subsets.

 Journal of National Bureau of Standards, Vol.69B (1965), pp.67 72.
- [5] Harary, F.: A Graph-Theoretic Approach to Matrix Inversion by
- Partitioning. Numerische Mathematik, Vol.4 (1962), pp.128 135.
- [6] Himmelblau, D. M.: Decomposition of Large Scale Systems --- I.

 Systems composed of lumped parameter elements. Chemical Engineering

 Science, Vol.21 (1966), pp.425 438.
- [7] Himmelblau, D. M.: Decomposition of Large Scale Systems --- II.

 Systems containing nonlinear elements. Chemical Engineering Science,

- Vol.22 (1967), pp.883 895.
- [8] Institute of the Union of Japanese Scientists and Engineers: JUSE-L-GIFS User's Manual, Version 3 (in Japanese), 1976.
- [9] Iri, M.: The Maximum-Rank Minimum-Term-Rank Theorem for the Pivotal Transforms of a Matrix. <u>Linear Algebra and Its Applications</u>, Vol.2 (1969), pp.427 446.
- [10] Iri, M.: A Review of Recent Work in Japan on Principal Partitions of Matroids and Their Applications. Annals of the New York Academy of Sciences, Vol.319 (1979), pp.306 319.
- [11] Iri, M., and Fujishige, S.: Use of Matroid Theory in Operations

 Research, Circuits and Systems Theory. <u>International Journal of Systems</u>

 Science, Vol.12 (1981), pp.27 54.
- [12] Iri, M., and Tomizawa, N.: A Unifying Approach to Fundamental Problems in Network Theory by Means of Matroids. Electronics and Communications in Japan, Vol.58-A (1975), pp.28 35.
- [13] Iri, M., Tsunekawa, J., and Murota, K.: Graph-Theoretic Approach to Large-Scale Systems --- Structural Solvability and Block-Triangularization (in Japanese). <u>Transactions of Information Processing Society of Japan</u>, Vol.23 (1982), pp.88 95. (English translation available).
- [14] Iri, M., Tsunekawa, J., and Yajima, K.: The Graphical Techniques Used for a Chemical Process Simulator "JUSE GIFS". <u>Information Processing 71</u> (Proceedings of the IFIP Congress 71), Vol.2 (Applications), (1972), pp.1150 1155.
- [15] Lee, W., Christensen, J. H., and Rudd, D. F.: Design Variable

 Selection to Simplify Process Calculations. A. I. Ch. E. (American

 Institute of Chemical Engineers) Journal, Vol.12 (1966), pp.1104 1110.

 [16] Murota, K.: Decomposition of a Graph Based on the Menger-type

- Linkings on It (in Japanese). <u>Transactions of the Information Processing</u>
 Society of Japan, Vol.23 (1982), pp.280 287.
- [17] Murota, K.: Structural Analysis of a Large-Scale System of Equations by Means of the M-Decomposition of a Graph (in Japanese). <u>Transactions of the Information Processing Society of Japan</u>, Vol.23 (1982), No.5, to appear.
- [18] Murota, K.: Menger-Decomposition of a Graph and Its Application to the Structural Analysis of a Large-Scale System of Equations. <u>Kokyuroku</u>, Kyoto University, No.453 (1982), pp.127 173.
- [19] Murota, K.: LU-Decomposition of a Matrix with Entries of Different Kinds. Kokyuroku, Kyoto University, No.463 (1982), pp.23 32.
- [20] Nakamura, M., and Iri, M.: A Structural Theory for Submodular Functions, Polymatroids and Polymatroid Intersections. Research Memorandum RMI 81-06, Department of Mathematical Engineering and Instrumentation Physics, University of Tokyo, 1981.
- [21] Recski, A.: Sufficient Conditions for the Unique Solvability of Linear Memoryless 2-Ports. <u>Circuit Theory and Applications</u>, Vol.8 (1980), pp.95 103.
- [22] Recski, A., and Iri, M.: Network Theory and Transversal Matroids.

 Discrete Applied Mathematics, Vol.2 (1980), pp.311 326.
- [23] Schrijver, A.: <u>Matroids and Linking Systems</u>. Mathematical Centre Tracts, Vol.88, Amsterdam, 1978.
- [24] Sebastian, D. J.G., Noble, R. G., Thambynayagam, R. K. M., and Wood, R.K.: DPS —— A Unique Tool for Process Simulation. 2nd World Congress of Chemical Engineering, Montreal, 1981.
- [25] Steward, D. V.: On an Approach to Techniques for the Analysis of the Structure of Large Systems of Equations. SIAM Review, Vol.4 (1962), pp.321

- 342.
- [26] Steward, D. V.: Partitioning and Tearing Systems of Equations. <u>SIAM</u>

 Journal on Numerical Analysis, Series B2 (1965), pp.345 365.
- [27] Tomizawa, N., and Iri, M.: An Algorithm for Determining the Rank of a Triple Matrix Product AXB with Application to the Problem of Discerning the Existence of the Unique Solution in a Network. Electronics and Communications in Japan, Vol.57-A (1974), pp.50 57.
- [28] Thambynayagam, R. K. M., Wood, R. K., and Winter, P.: DPS --- An Engineer's Tool for Dynamic Process Analysis. The Chemical Engineer, No.365 (1981), pp.58 65.
- [29] van der Waerden, B. L.: Algebra. Springer-Verlag, Berlin, 1955.
- [30] Welsh, D. J. A.: Matroid Theory. Academic Press, London, 1976.
- [31] Yajima, K., Tsunekawa, J., and Kobayashi, S.: On Equation-based

 Dynamic Simulation. Proc. of 2nd World Congress of Chemical Engineering,

 Vol.V, pp.469 480, 1981.
- [32] Yajima, K., Tsunekawa, J., Shono, H., Kobayashi, S., and Sebastian, D. J.: On Graph-Theoretic Techniques for Large-Scale Process Systems.

 International Symposium on Process Systems Engineering, Kyoto, August, 1982.