VARIANTS OF THE HUNGARIAN METHOD FOR ASSIGNMENT PROBLEMS[1]

H. W. Kuhn
Bryn Mawr College

The author presents a geometrical model which illuminates variants of the Hungarian method for the solution of the assignment problem.

1. INTRODUCTION

The Hungarian Method [1] is an algorithm for solving assignment problems that is based on the work of D. König and J. Egerváry. In one possible interpretation, an assignment problem asks for the best assignment of a set of persons to a set of jobs, where the feasible assignments are ranked by the total scores or ratings of the workers in the jobs to which they are assigned. It is a special case of the transportation problem, which asks for the allocation of a homogeneous commodity from given supplies at a set of sources to satisfy prescribed demands at a set of destinations, while minimizing the total source-destination transportation costs. The extension of the Hungarian Method to transportation problems can be found in the work of L. R. Ford, Jr., and D. R. Fulkerson [2] and of J. Munkres [6]. Comprehensive reviews of these problems may be found in the papers of M. M. Flood [3] and T. S. Motzkin [4].

The original purpose of this note was to present a modification of the Hungarian Method, inspired by an algorithm of Marshall Hall, Jr. [5], for the choice of a system of distinct representatives for a family of sets. But is now seems that a synthesis of all of the available variants would be more valuable. Accordingly, the paper now presents a geometrical model that is adequate to define and compare the algorithms contained in [1], [2], [5], and [6]. In Section 2, the Hungarian Method is reviewed in outline so as to isolate the procedure that is altered in these variants. The modification suggested by Hall's work is presented in Section 3. A translation into graph-theoretical terms of the problem solved by all four versions is the subject of Section 4; this section is self-contained and can be read by itself. A final section compares these alternative solutions by means of this geometric model.

2. RESTATEMENT OF THE HUNGARIAN METHOD

As considered in this paper, the assignment problem asks: Given an n-by-n matrix $A = (a_{ij})$ of non-negative integers, find the permutation $j_1, \ldots, j_n$ of the integers $1, \ldots, n$

[1]The work reported by this note was supported by the Office of Naval Research Logistics Project, Department of Mathematics, Princeton University.

*Manuscript received January 16, 1956.

that minimizes the sum $a_{1j_1} + \ldots + a_{nj_n}$. (If a maximum permutation sum is demanded, set $a = \max_{i,j} a_{ij}$ and replace A by $A' = (a'_{ij})$ with $a'_{ij} = a - a_{ij}$, for all i and j.) The Hungarian method is then based on two assertions:

A. The problem is unchanged if the matrix A is replaced by $A' = (a'_{ij})$, with $a'_{ij} = a_{ij} - u_i - v_j$ for constants u_i and v_j, and $i, j = 1, \ldots, n$.

B. (König's theorem) The minimum number of lines (rows and/or columns) needed to contain all of the zeros of A is equal to the maximum number of zeros that can be chosen, with no two on the same line.

As presented in [1], the algorithm divides into two routines. Routine I is essentially an iterative procedure for determining the minimum number of lines and the maximum number of zeros referred to in König's theorem; it is the possible variations in this routine that form the main subject of this paper. The input of the routine is a subset of the zeros of A, marked by asterisks, with no two asterisks in the same line. If the number of asterisks is n, then the desired minimum is zero, and the problem is solved by $j_1, \ldots, j_n$ where the asterisk in row i appears in column j_i, for $i = 1, \ldots, n$. If the problem is not solved (i.e., if there are $k < n$ asterisks) then the output of a single application of Routine I is one of the following disjoint alternatives:

Ia. A new set of $k + 1$ asterisks on zeros in A, with no two on the same line; or

Ib. A set C of $k < n$ lines which contain all of the zeros of A.

Routine I is repeated until there are n asterisks (and the problem is solved) or until there is an occurrence of Alternative Ib. In the latter event, let $h > 0$ be the minimum entry not appearing in the lines of C. Routine II calls for the addition of h to all entries in each line of C (if an entry appears in two lines, then 2h is added) then for the subtraction of h from every entry in A. Routine II leaves A non-negative but decreases the sum of all entries by an amount $nh(n - k)$. The only property of Ib needed for the finite termination of the combined algorithm is that the set C contains less than n lines. The original Routine I produces the exact minimum k; in the modification of Routine I proposed in the next section, the set C contains exactly $n - 1$ lines at every occurrence of Alternative Ib.

3. MODIFIED ROUTINE I

Search each column of A in turn for a 0^*. If a 0^* is found, proceed to the next column (if no columns remain then there are n asterisks and the problem is solved). If a 0^* is not found in the column, then the column is called pivotal and is searched for all of its 0's. If no 0 is found in the pivotal column, then the remaining $n - 1$ columns contain all of the zeros of A (Modified Alternative Ib). If there are 0's in the pivotal column, then their rows are searched in turn for a 0^*. If such a row contains no 0^*, then the 0 in that row and in the pivotal column is marked with an asterisk (Alternative Ia). If each 0 in the pivotal column has a 0^* in its row, then these rows are listed in any order: $i_1, \ldots, i_t$. We now construct a sequence of rows based on this initial segment. At step s of the construction, we search row i_s for a 0^*. If a 0^* is found, then the row indices of all 0's in its column are added to the sequence in any order and any indices already present are omitted. If no 0^* is found in row i_s, then a transfer (see [1] and Section 4 below) is possible. Namely, i_s was added to the sequence via a 0 in the column of a 0^* in an earlier row i_r of the sequence. If the asterisk is transferred from row i_r to row i_s in this column, then we have the same situation earlier in the sequence. Continuing, we ultimately free an asterisk in a row of the initial segment, $i_1, \ldots, i_t$. But then an asterisk can be marked in this row of the pivotal column (Alternative Ia). In the remaining case, a finite sequence $i_1, \ldots, i_v$ is constructed with the following properties: Each row in the sequence contains a 0^* in some column. These columns, together with the pivotal column ($v + 1$ columns in all), contain 0's (marked or not) only in rows $i_1, \ldots, i_v$. Hence, the remaining $n - v - 1$ columns combined with the rows $i_1, \ldots, i_v$ constitute $n - 1$ lines that contain all zeros (Modified Alternative Ib).

In a rough attempt to evaluate the advantages of this variant, it may simplify the search for a transfer in Routine I at the possible expense of smaller changes in the sum of the entries in A on the next application of Routine II.

4. A GRAPH-THEORETICAL EQUIVALENT OF ROUTINE I

In this section, the problem solved by Routine I (and by the algorithm of Section 3) is given an independent statement and then rephrased as a problem in graph theory.

Let A be an n-by-n square array in which certain of the places are occupied by zeros. An assignment is a subset of k zeros, distinguished by asterisks, such that no two asterisks lie in the same line (row or column) of A. An assignment is complete if every zero lies in the line of some asterisk. A cover is a set of lines of A that contain all of the zeros in A; if a complete assignment is available, only the lines in which asterisks appear will be used in a cover. König's theorem asserts that the largest number of asterisks in an assignment is equal to the smallest number of lines in a cover. Such an assignment is called maximal; such a cover is called minimal. A transfer is possible when there is a sequence of asterisks at $(i_1, j_1), \ldots, (i_r, j_r)$ in A such that there are zeros at $(i_2, j_1), \ldots, (i_r, j_{r-1})$ and a zero at (i_1, j_0), for some j_0, with no asterisk in its column. The transfer removes the asterisks from $(i_1, j_1), \ldots, (i_r, j_r)$ and assigns them to $(i_1, j_0), \ldots, (i_r, j_{r-1})$. If there is a zero at (i_{r+1}, j_r), for some i_{r+1}, with no asterisk in its row, then the result of the transfer is an incomplete assignment (and another asterisk can be placed at (i_{r+1}, j_r)).

To each complete assignment for A, we shall associate an oriented graph G by the following rules:

The nodes V of G are in a 1-1 correspondence with the asterisks. A directed edge joins V_1 to V_2 if there is a zero in intersection of the column of the asterisk-node V_1 and the row of the asterisk-node V_2. The node V is distinguished as a source (sink) if there is a zero in the row (column) of V with no asterisk in its column (row). Note that a node can be both a source and a sink.

The equivalence of various concepts in the matrix A and the graph G is displayed in the table below:

Matrix A	Graph G
Transfer	Directed path (possibly void) originating at a source.
Transfer yielding incomplete assignment	Directed path (possibly void) from a source to a sink.
Cover	Assignment of + and/or - to each node so that every source (sink) is marked + (-), and no edge starts at a + and ends at a - .

The first two equivalences follow directly from the definitions. To prove the last equivalence, suppose that a cover is given for the matrix A in which the assignment is complete. Each asterisk is contained in a line of the cover; if it is covered by a row (column), mark the corresponding node by a + (-). An asterisk-node V is a source (sink) if, and only if, there is a zero in the row (column) of V without an asterisk in its column (row). This zero is covered; hence the node is marked with a + (-) if it is a source (sink). Each edge in G corresponds to an unmarked zero, and leads from an asterisk-node in its column to an asterisk-node in its row. For this zero to be covered, either the former is marked by a - or the latter by a + (Recall that when the available assignment is complete, only lines through asterisks appear in a cover.) To show the other half of the equivalence, consider the set of the rows of the asterisk-nodes marked with a + and of the columns of the asterisk-nodes marked with a - ; under the conditions listed, it is asserted that this set is a cover. Clearly, only zeros without asterisks need be considered. If there is an asterisk in the row (column) of such a zero, but no asterisk in its column (row), then this asterisk-node is a source (sink) and is marked by a + (-). Hence the row (column) of the zero without an asterisk is in the set. If there is an asterisk in both the row and column of an unmarked zero, then there is an edge from the latter to the former. Since either the asterisk-node in the row is marked with a + or the asterisk-node in the column is marked with a - , the unmarked zero is covered.

THEOREM 1: Let A be a matrix with a complete assignment and let G be the associated oriented graph. Suppose each node of G is assigned + or - (and not both), so that every source (sink) is marked + (-) and no edge starts at a + and ends at a - . Then the corresponding cover is minimal.

PROOF: If there are k asterisks assigned in A, no cover can contain less than k lines. The cover corresponding to the given distribution of one + or - to each node of G contains exactly k lines. Q.E.D.

We now prove a purely graph-theoretic theorem that will form the basis of our discussion of Routine I (and, incidentally, prove König's theorem).

THEOREM 2: Let G be a directed graph in which two disjoint (and possibly void) sets of nodes have been distinguished as <u>sources</u> and <u>sinks</u>, respectively. Then exactly one of the following alternatives holds:

(a) There is a directed path in G from a source to a sink.

(b) Each node of G can be marked with one sign (+ or -), so that every source (sink) is marked + (-) and no edge starts at a + and ends at a - .

PROOF: Both alternatives cannot hold since, for any distribution of signs and any directed path from a source (+) to a sink (-), there must be an edge that starts at a + and ends at a - .

Let S be the set of nodes that are connected to some source by a directed path (starting at the source). If a sink lies in S then (a) holds. If a sink does not lie in S, then assign + to the nodes in S and - to the nodes not in S. This marks the sources and sinks as required by (b). Suppose that an edge starts at a node marked +, and hence in S. If the directed path from a source to this node is extended by the edge in question, the terminus is connected to the source by the extended path. Hence the terminus is in S and is marked +. Q.E.D.

Remark 1: A proof of König's theorem now results as follows: It is a simple matter to construct a complete assignment of asterisks to the zeros in A (say, by assigning an asterisk to the first zero in each row which is not in the column of an asterisk that has been placed previously). If alternative (a) of Theorem 2 holds, a transfer yields an incomplete assignment and another asterisk can be placed. Repeating, alternative (a) can hold only a finite number of times (surely no more than $n - k$). When alternative (b) first occurs, we have a minimal cover with the same number of lines as there are assigned asterisks, and König's theorem is proved.

Remark 2: If there are no sources (sinks) then the theorem is trivially satisfied by marking all nodes - (+).

Remark 3: The construction of the proof can be altered in the following manner. Let S be as above. Let T be the set of nodes that are connected to a sink by a directed path (ending at the sink). If there is a node both in S and T, then (a) holds. Otherwise, mark the nodes in S (T) with + (-), and mark all nodes that are in neither S nor T with the same sign (either all + or all -).

5. VARIANTS OF ROUTINE I ON THE GRAPH G

The methods described below all amount to a construction of the set S or T in different manners. They will be defined in informal geometric terms, and the translation back into terms of the matrix A will be left for the reader.

1. (Kuhn [1]). In the search for a path from a source to a sink, start at any source and continue as far as possible along a directed path. If a sink is not reached, backtrack to the nearest branch point and try again. When the directed paths originating at one source are exhausted, try the next source. If all sources are exhausted and no sink has been reached, the set of all nodes that have been encountered is S.

2. (Ford-Fulkerson [2]). Start simultaneously from all sources and construct all directed paths with one edge (or less) originating at the sources. Extend these to all directed paths with two edges (or less). Continue fanning out from all sources until a sink is encountered or until S is exhausted.

3. (Hall [5]). Pick a column of A without an asterisk and call it pivotal. If there is no 0 in this column, then the remaining columns form a cover of $n - 1$ lines. Otherwise, each 0 is in the row of a source. Fan out from these sources (as in 2, above) until a sink is encountered or until the set of nodes connected to these sources by directed paths is exhausted. The cover then consists of the rows of the asterisk-nodes thus encountered completed by all columns except the pivotal column and the columns of asterisk-nodes already covered by rows ($n - 1$ lines in all).

4. (Munkres [6]). Mark all nodes both + and -. Then examine the nodes that have two signs, according to some preassigned order (say, from left to right in the matrix A). If such a node is a source (sink) or is connected from (to) a node marked only + (-) by a directed edge, remove the - (+) from the node and prime the edge used in the criterion (if any). If both signs are removed, a unique path from a source to a sink consisting of primed edges has been found. Otherwise, the examination is repeated until no more removals are indicated. At this point, the nodes marked + (-) constitute S (T) and the remaining double signs can be changed to -. (See Remark 3, above.)

Through the use of the graphical model, the reader can easily concoct variants of his own. Moreover, simple graphical examples show that none of the four methods given above is the best for all matrices A.

The method of Munkres sketched in Section 5 above is replaced in the revised version of his paper 6 by the following variant:

5. (Munkres 6). Mark all nodes - . Examine each source, sink, and edge in some pre-assigned order (since each of these objects is associated with one or more unassigned zeros in A, the natural order of these zeros can be followed). If a source marked - , or an edge connecting a + to a - , is found then change the - to a + (and prime the edge). If a sink marked + is found then a unique path consisting of primed edges from a source to a sink has been discovered. (We now note that, if the rows and columns associated with the current marks + and - are covered, each of the objects for which we are searching is associated with one or more uncovered zeros in A, and we can follow the natural order of these zeros.) Otherwise the process is repeated until all sources are marked + and no edge connects a + to a - (Alternative (b) of Theorem 2).

REFERENCES

[1] H. W. Kuhn, "The Hungarian Method for the Assignment Problem," Naval Research Logistics Quarterly, Vol. 2, Nos. 1 and 2 (1955), pp. 83-97.

[2] L. R. Ford, Jr., and D. R. Fulkerson, "A Simple Algorithm for Finding Maximal Network Flows and an Application to the Hitchcock Problem," RAND Corporation, Santa Monica, P-743, 26 September 1955.

[3] M. M. Flood, "The Traveling-Salesman Problem," Operations Research, Vol. 4, (1956), pp. 61-75.

[4] T. S. Motzkin, "The Assignment Problem," Proc. of Symposia in Applied Math., Vol. VI - Numerical Analysis, pp. 109-125, McGraw-Hill, New York, 1956.

[5] Marshall Hall, Jr., "An Algorithm for Distinct Representatives," Amer. Math. Monthly, Vol. 63, (1956), pp. 716-717.

[6] J. Munkres, "Algorithms for the Assignment and Transportation Problems," to appear in the Journal of the Society for Industrial and Applied Mathematics (SIAM).

* * *