P-14

On the Computational Complexity of Clustering and Related Problems

Teofilo F. Gonzalez
Programs in Mathematical Sciences
The University of Texas at Dallas
Richardson, Texas 75080/USA

Abstract

The problem of clustering a set of n points into k groups under various objective functions is studied. It is shown that under some objective functions clustering problems are NP-hard even when the points to be grouped are restricted to lie in the two dimensional euclidean space. Our results can be extended to show that their corresponding approximation problems are also NP-hard. It is shown that some restricted graph partition problems are also NP-hard. Keywords: NP-complete problems, approximation algorithms, clustering problems.

I. INTRODUCTION

The problem of clustering a set of objects arises in many disciplines. Because of the wide range of applications, there are many variations of this problem. The main difference between these clustering problems is in the objective function. Research in different fields of study during the past thirty years has produced a long list of clustering algorithms. However, very little is known about the merits of these algorithms. Even simple questions regarding to the computational complexity of most clustering problems have not yet been answered. In this paper, we study the computational complexity of typical clustering problems.

In what follows, we define some of the typical clustering problems we are interested in studying. Let G=(V,E,W) be a weighted undirected graph with vertex set V, edge set E and a disimilarity or weight function $W{:}E \to R_0^+$ (the set of non-negative reals). A k-split of the set of vertices V is a set of nonempty vertex subsets $B_1,B_2,\ldots,B_k$ such that $\cup B_i=V$. The sets B_i in a k-split are called clusters. The clusters are said to be nonoverlapping when $\sum|B_i| = |V|$. In what follows, we shall concentrate only on nonoverlapping clustering problems. An objective function, f: $B_1,B_2,\ldots,B_k \to R_0^+$, is defined for each k-split. For k-split $B_1,B_2,\ldots,B_k$, we define S_ℓ as the sum of the weights assigned to the edges adjacent to any pair of nodes in set B_ℓ, i.e., $S_\ell = \sum_{\substack{i,j\in B_\ell \\ \{i,j\}\in E}} W(\{i,j\})$. M_ℓ denotes the maximum weight assigned to an edge whose endpoints are vertices in cluster B_ℓ, i.e., $M_\ell = \max_{\substack{i,j\in B_\ell \\ \{i,j\}\in E}} \{W(\{i,j\})\}$. Some typical objective functions are shown in Table 1.

	$f(B_1,B_2,\ldots,B_k$
$(\sum\sum)$	$\sum_{\ell=1}^{k} S_\ell$
$(\sum 1/\lvert\cdot\rvert\sum)$	$\sum_{\ell=1}^{k} S_\ell/\lvert B_\ell\rvert$
$(M\sum)$	$\max_{1\le\ell\le k}\{S_\ell\}$
$(\sum M)$	$\sum_{\ell=1}^{k} M_\ell$
$(M\ M)$	$\max_{1\le\ell\le k}\{M_\ell\}$
(D)	$\sum_{\ell=1}^{k}\sum_{i\in B_\ell} \lvert\lvert x_i - m_\ell\rvert\rvert^2$, where the set of points (vertices) are in m-dimensional space, $\lvert\lvert x_k\rvert\rvert = \sqrt{{}_\ell\Sigma((x_k)_\ell)^2}$ and m_ℓ is the centroid, i.e., $m_\ell = (1/\lvert B_\ell\rvert)\sum_{i\in B_\ell} x_i$.

Table 1. Objective Functions

A clustering problem has one of the following forms:

(P1) Given a graph G, an objective function f and an integer k, find a k-split with least objective function value, i.e., find a k-split $(B_1^*,B_2^*,\ldots,B_k^*)$ such that $f(B_1^*,B_2^*,\ldots,B_k^*) = \min\{f(B_1,B_2,\ldots,B_k)|(B_1,B_2,\ldots,B_k)$ is a k-split for $G\}$.

(P2) Given a graph G, an objective function f and a real w, find for the least value of k a k-split with objective function value less than or equal to w, i.e., find a k-split $(B_1{}^*,B_2^*,\ldots,B_k^*)$ such that $f(B_1^*,B_2^*,\ldots,B_k^*) \le w$ and $f(B_1,B_2,\ldots,B_{k'}) > w$ for all k'-splits with $k' < k$.

(P3) Given a graph G, an objective function f, an integer k' and a real w. Is there a k-split $(B_1,B_2,\ldots,B_k)$ with objective function value $\le w$ for some $k \le k'$?

It can be easily shown that the decision problem P3 is computationally not harder than P1 and P2, i.e., any algorithm which solves P1 or P2 can be used to solve P3. This relation implies that if problem P3 is NP-complete then both P1 and P2 are NP-hard. In what follows when we refer to optimization clustering problems, it is implied that we refer to problems of the form P1. Whenever we wish to consider problems in the form P2, we shall state it explicitly.

An m-dimensional clustering problem is one in which the vertices of G are points in the m-dimensional euclidean space, the set of edges is complete and the weight of each edge is given by the euclidean distance between the two points it joins, i.e., $W(x_i,x_j) = \lvert\lvert x_i - x_j\rvert\rvert$ where $\lvert\lvert x_k\rvert\rvert = \sqrt{{}_\ell\Sigma\,((x_k)_\ell)^2}$.

We shall refer to a clustering problem as an $\alpha - \beta\gamma$ problem, where $\alpha \in \{2,3,\ldots, k\}$ is the number of clusters; β means that it is either a β-dimensional clustering problem ($\beta \in \{1,2,\ldots,m\}$) or that the problem has been defined over an arbitrarily weighted graph (β=g); and $\gamma \in \{\Sigma\Sigma, \Sigma 1/|\cdot|\Sigma, M\Sigma, \Sigma M, MM, D\}$ is the objective function (see Table 1). For example, k-2ΣΣ indicates that the number of clusters k is an input to the problems; it is a 2-dimensional euclidean problems; and the objective function is ΣΣ (see Table 1). Note that any algorithm which solves the k-2ΣΣ problem will also solve the 2-2ΣΣ problem, but the converse is not true. In the 2-2ΣΣ problem, the set of vertices in G is always partitioned into two clusters whereas in the k-2ΣΣ problem the set of vertices in G will be partitioned into k clusters, where k could be any integer greater than 1.

Let us now define the k-maxcut problem. This problem is similar to the k-gΣΣ problem, but instead of finding a nonoverlapping k-split minimizing the sum of the weights of the edges inside a cluster, the objective is to find a nonoverlapping k-split maximizing the sum of the weights of the edges between clusters [SG,K and GJS].

A reader not familiar with NP-complete problems and approximate solutions is referred to [HS,GJ2 and K]. Our notation is that of [HS].

It is simple to prove that for any k, the k-gΣΣ problem is computationally identical to the k-maxcut problem, i.e., any algorithm solving one of these problems will also solve the other problem. The k-maxcut problem for k = 2 was shown to be NP-hard in [K]; in [SG] it was shown to be NP-hard for k > 2; and in [GJS] it was shown to be NP-hard for k = 2 even when the weight of every edge is zero or one. Hence, k-gΣΣ is NP-hard. Sahni and Gonzalez [SG] showed that there is an efficient (1/k)-approximation algorithm for the k-maxcut problem, whereas the k-gΣΣ ε-approximation problem is NP-hard. Using the same approach as the one in [SG], one can show that k-gΣM, k-gMΣ, k-gMM, k-gΣ1/|•|Σ and their corresponding ε-approximation problems are also NP-hard.

Fisher [F] showed that the k-1D problem can be solved in polynomial time. This was shown by first proving that in every problem instance there exists an optimal solution with the property that the convex hulls of every pair of distinct clusters are disjoint. This reduces the problem to one that can be solved by dynamic programming procedures. Bodin [Bd] extended this approach to solve other clustering problems. A similar approach was used by Brucker [Br] to show that the k-1ΣM, k-1MM and k-1Σ1/|•|Σ can be solved in polynomial time. The k-1ΣM problem can also be solved by reducing it to the problem of finding the largest k gaps [Br], which can be solved in O(n log n) time. When k is some fixed constant, finding the largest k gaps can be solved in linear time [G1]. Bock [Bk] showed that the 2-mΣ1/|•|Σ problem can be solved in polynomial time. This was shown by first proving that every instance of the k-mΣ1/|•|Σ problem has an optimal solution with the property that the convex hulls of every pair of distinct clusters are disjoint. The 2-gMM problem can be solved efficiently by reducing the problem to that of testing whether a graph is bipartite or not [Br]. Gonzalez [G2] showed that the k-2(*) problem can be solved efficiently when k is some fixed constant and (*) represents objective functions with some given properties.

For general graphs, most clustering problems are NP-hard. On the other hand, 1-dimensional clustering problem can be solved efficiently. The complexity of most 2-dimensional clustering problems is not known. In this paper, we study the computational complexity of exact and approximate solutions to these problems.

For optimization problems of the form P2, one can show that the k-g$\Sigma\Sigma$ ε-approximation problem is computationally identical to the k-maxcut ε-approximation problem. For k-g$\Sigma\Sigma$, k-maxcut, k-g$\Sigma 1/|\cdot|\Sigma$, k-gMΣ, k-gΣM and k-gMM the 1-approximation problem is NP-hard. The proof follows the same approach as the one in [SG] but uses the result in [GJ1], which states that the 1-approximation problem for graph coloration is NP-hard.

Algorithms for other clustering problems appear in [AM], [JL], [S1], [S2], [FV], [DH], [M], [Sh] and [R].

In section II, we show that the k-2MM problem is NP-hard. The same reduction is then used in section III to show that the following problems are also NP-hard: k-2MΣ, k-2MM 1.36-approximation and k-2MΣ 1.16-approximation.

II. The Complexity of the k-2MM Decision Problem

In this section it is shown that the k-2MM decision problem is NP-complete. This result is obtained by reducing a restricted version of the exact cover by three sets problem to it.

The exact cover by three sets (XC3) problem was shown to be NP-complete in [GJ3] and is defined as follows:

> Exact Cover by Three Sets(XC3): Given a finite set of elements $X=\{x_1,x_2,\ldots x_{3q}\}$ and a collection of 3-element subsets of X, $C=\{(x_{i_\ell},x_{j_\ell},x_{k_\ell})\,|\,1 \leq \ell \leq m\}$, in which no element in X appears in more than three subset. The problem consists of determining whether C has an exact cover for X, i.e., a subcollection $C' \subseteq C$ such that every element in X occurs in exactly one member of C'.

The restricted version of this problem, to be used in our reduction, is denoted RXC3. This problem is exactly like the XC3 problem, except that each element in X appears in exactly three subsets of C. RXC3 is shown to be NP-complete in [G2].

In order to simplify the presentation of our result, we begin by showing that the k-gMM decision problem is NP-complete (lemma 1). The construction used in this lemma is then modified to show thet the k-gMM decision problem is NP-complete even when the input graph, after deleting all edges with weight different than one, is planar and no node is of degree greater than six (lemma 2). We then show how this result can be used to prove that the k-2MM decision problem is NP-complete (theorem 1). The reduction RXC3 α k-gMM is identical to the one in [GJ2], which was used to show that partition of a graph into triangles is NP-complete.

Lemma 1: The decision problem k-gMM is NP-complete.

Proof: It is simple to show that the decision problem k-gMM can be solved in nondeterministic polynomial time. We now show that RXC3 α k-gMM.

Given an instance, (X,C), of the restricted exact cover by 3-sets problem, we construct an instance of the k-gMM decision problem which we denote KG. KG=(G=(V,E,W),k,d) is defined as follows:

> Vertex set: There is a vertex (v_i) for each element of set X and nine vertices, ($a_{\ell,1}, b_{\ell,1}, c_{\ell,1}, \ldots, a_{\ell,3}, b_{\ell,3}, c_{\ell,3}$), are introduced for each 3-element subset of X in C.
>
> Edge set: The set of edges is complete,i.e., for every pair of vetices i≠j edge {i,j} is in E.
>
> Weights: For each 3-element subset of X in C, eighteen edges will get a weight of one. The edges introduced for ($x_{i_{\ell,1}}, x_{i_{\ell,2}}, x_{i_{\ell,3}}$) C are shown in figure 1. All other edges are given the weight of two.

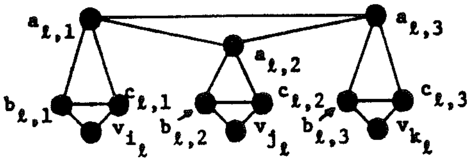


figure 1.

The maximum number of clusters, k, is 3m+q. The maximum weight for an edge inside a cluster, d, is one.

In order to complete the proof of the lemma it is only required to show that KG has a k-split with objective function value $\leq$ d iff (X,C) has an exact cover, since the construction of KG can be carried out in polynomial time.

Claim: KG has a k-split with objective function value $\leq$ d iff (X,C) has an exact cover.

Proof: First of all it is shown that if (X,C) has an exact cover, then KG has a k-split ($B_1, B_2, \ldots, B_k$) with objective function value $\leq$d=1. Let C' be any exact cover for (X,C). Assume without loss of generality that ($x_{i_{1,1}}, x_{i_{1,2}}, x_{i_{1,3}}$),..., ($x_{i_{q,1}}, x_{i_{q,2}}, x_{i_{q,3}}$) are the elements in C which form an exact cover C'. Let $B_{\ell,j}=\{b_{\ell,j}, c_{\ell,j}, v_{i_{\ell,j}}\}$ for $1 \leq j \leq 3$ and $1 \leq \ell \leq q$; let $B_{\ell,4}=\{a_{\ell,1}, a_{\ell,2}, a_{\ell,3}\}$ for $1 \leq \ell \leq q$; and let $B_{\ell,j}=\{a_{\ell,j}, b_{\ell,j}, c_{\ell,j}\}$ for $1 \leq j \leq 3$ and $q+1 \leq \ell \leq m$. It is simple to show that ($B_{1,1}, \ldots, B_{q,4}, B_{q+1,1}, \ldots, B_{m,3}$) is a k-split with objective function value equal to d for KG.

In order to complete the proof of the claim it is only required to show that if KG has a k-split with objective function $\leq$d=1, then (X,C) has an exact cover. Let $B_1, B_2, \ldots, B_k$ be a k-split with objective function value $\leq$d for KG. Since no four nodes are completly connected by edges with a weight of one and since the number of nodes in KG is 3*k, we have that each B_i must have exactly three nodes. Let

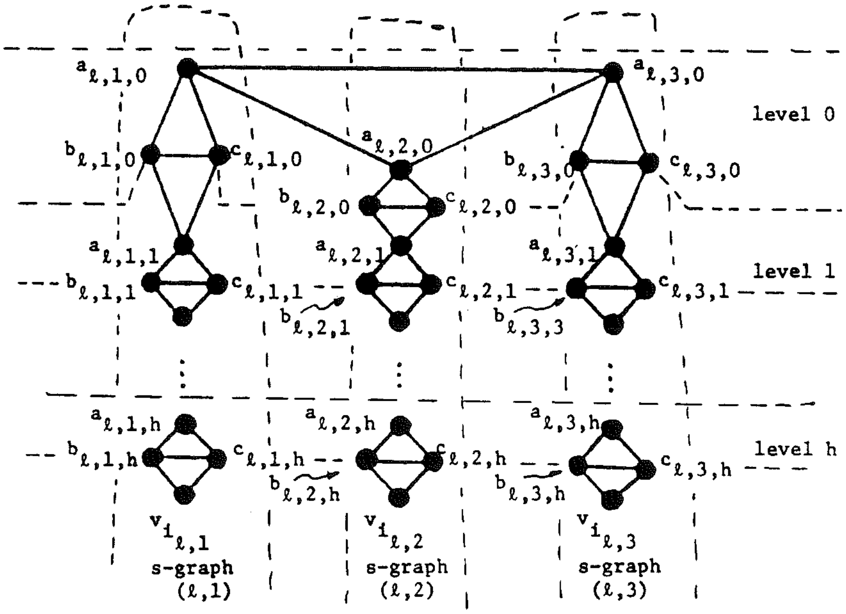
$\gamma_z=(\ell,j)$, for z=1,...,3q, if vertex v_{i_z} is in the same cluster with $b_{\ell,j}$ and $c_{\ell,j}$. It can be easily shown that: i) If $\gamma_z=(\ell,j)$ for some z, then $a_{\ell,j}$ is not in the same cluster with $b_{\ell,j}$ and $c_{\ell,j}$.

and ii) If for all z $\gamma_z\neq(\ell,j)$ then $a_{\ell,j}$, $b_{\ell,j}$ and $c_{\ell,j}$ are in the same cluster.

We now show that if for some z, $\gamma_z=(\ell,j)$ then there exists z_i, $1\leq i\leq 3$, such that $\gamma_{z_i}=(\ell,i)$. Let j_1 and j_2 be such that $\{j_1,j_2,j\}=\{1,2,3\}$. The construction rules together with the fact that each cluster has exactly three nodes implies that $a_{\ell,j}$ can be in a cluster only with either $b_{\ell,j}$ and $c_{\ell,j}$ or a_{ℓ,j_1} and a_{ℓ,j_2}. Since i) holds true for $\gamma_z=(\ell,j)$, it must then be that $a_{\ell,1}$, $a_{\ell,2}$ and $a_{\ell,3}$ are in the same cluster. This fact together with i) and ii) imply that there exists z_i, $1\leq i\leq 3$, such that $\gamma_{z_i}=(\ell,i)$.

Now, let $A=\{\ell\mid\gamma_z=(\ell,j)\}$. Clearly $|A|=q$. Also, it is simple to see that $C'=\{(x_{i_{\ell,1}},x_{i_{\ell,2}},x_{i_{\ell,3}})\mid\ell\epsilon A\}$ is an exact cover for C. This completes the proof of the claim and the lemma. □

Before proving our next result, we outline the construction to be used in it. First of all, the construction in lemma 1 (figure 1) is replaced by the one given in figure 2.



<u>figure 2</u>

The subgraph induced by the set of nodes $a_{\ell,j,z},b_{\ell,j,z},c_{\ell,j,z},v_{i_{\ell,j}}$ $0\leq z\leq h$ is called <u>s-graph</u>(ℓ,j). For z=0,1,...,h, nodes $a_{\ell,j,z}$, $b_{\ell,j,z}$ and $c_{\ell,j,z}$ are said to be in <u>level z</u>. The weight assigned to all edges introduced by the rule implied in by figure 2 is one.

It is simple to show that not all the graphs constructed by using the above rule, starting with an instance of RXC3, are planar. In order to guarantee planarity, we shall modify our construction rule. h is selected in such a way that at each level z ($z \geq 1$) only two adjacent s-graphs cross and after level h all the s-graphs that include node v_j are adjacent to each other. The crossing of the two s-graphs at level z is handled by applying the transformation shown in figure 3.

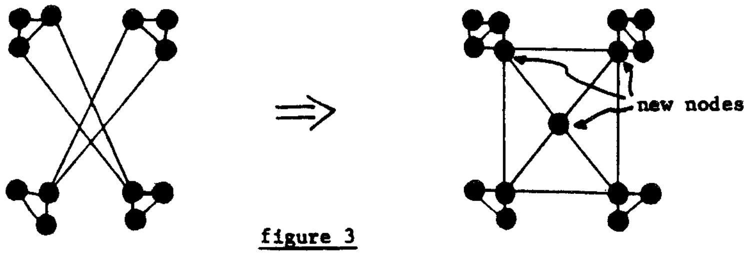


figure 3

Lemma 2: The decision problem k-gMM is NP-complete even when the input graph, after deleting all the edges with a weight different than one, is planar and no node is of degree greater than six.

Proof: The construction is as outlined above and the proof is similar to the proof of lemma 1. □

The subgraphs, in figure 2, consisting of two triangles placed side by side are called diamonds. The ends are the two nodes of degree two in it. It should be clear that two diamonds connected in series can replace any diamond and the resulting construction can also be used in lemma 2. This transformation can be carried out any number of times, as long as the total transformation takes polynomial time.

In the final transformation we replace the constructions implied in figures 1,2 and 3 by the one in figure 4.

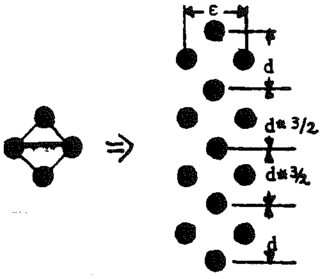
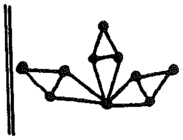
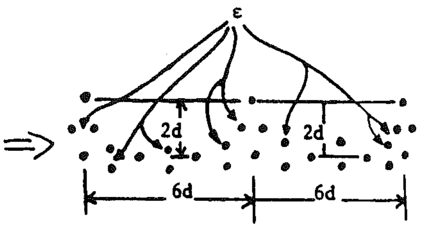
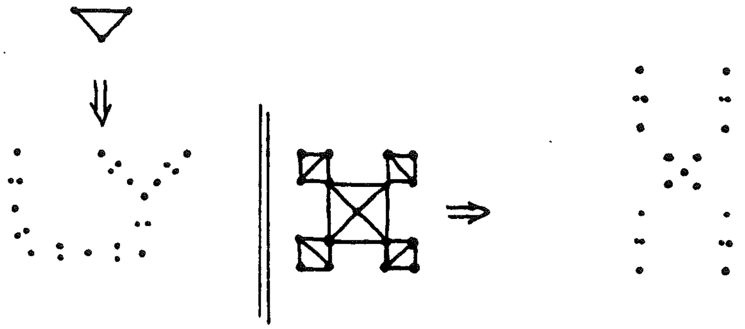


figure 4a







where {•• are at a distance ε
• • are at a distance 2*d

figure 4b.

After taking care of some simple details, one can show that two points are at a distance $\leq$ d+ε iff these two points had an edge between them with a weight of one in the construction used in lemma 2 (after adding several diamonds as shown in figure 4).

Theorem 1: The k-2MM problem is NP-complete.

Proof: The construction used in this proof follows the rules shown in figure 4 and the proof follows the same lines as the proof of lemma 2. □

III. The Complexity of Related Problems.

After a careful examination of the construction rules shown in figure 4, one can show that the closest three points not at a distance $\leq$ d + ε of each other, are at least $1/\sqrt{2}$ units apart. Using this fact together with the techniques used in [SG], one can prove the following theorem.

Theorem 2: The k-2MM $(1/\sqrt{2})$-approximation problem is NP-hard. □

The proofs and constructions of the next two theorems are similar to the ones in theorems 1 and 2. For brevity they will not be included.

Theorem 3: The k-2MΣ decision problem is NP-complete. □

Theorem 4: The k-2MΣ (1.16)-approximation problem is NP-hard. □

The construction used in section II can be easily adapted to show that partition of a graph into triangles is NP-complete even when the graphs to be partitioned are planar and no node is of degree ≥ 6.

(The formal proofs of our theorems will appear in a subsequent paper.)

References

[AM] Augustson, J. G. and J. Minker, "An Analysis of Some Graph Theoretical Cluster Techniques," J.ACM, 17,571-588,(October 1970).

[Bk] Bock, H. H., "Automatische Klassifikation," Vandenhoek und Ruprecht, Gottingen, 1974.

[Bd] Bodin, L. D., "A Graph Theoretic Approach to the Grouping of Ordering Data," Networks, 2, 307-310, (1972).

[Br] Brucker, P. "On the Complexity of Clustering Problems," in R. Henn, B. Korte and W. Oletti (eds), Optimiening and Operations Research, Lecture Notes in Economics and Mathematical Systems, Springer, Berlin (1977).

[DH] Duda, R. and P. Hart, "Pattern Classification and Scene Analysis," John Wiley and Sons, New York, 1973.

[FV] Fisher, L. and J. Van Ness, "Admissible Clustering Procedures," Biometrica, 58:91-104, 1971.

[F] Fisher, W. D., "On Grouping for Maximum Homogeneity, "JASA, 53:789-798,1958.

[G1] Gonzalez, T., "Algorithms on Sets and Related Problems," Technical Report 75-15, The University of Oklahoma, 1975.

[G2] Gonzalez, T.,Manuscript in preparation.

[GJ1] Garey, M. R. and D. S. Johnson, "The Complexity of Near-Optimal Graph Coloring," JACM, 23, 1, 43-69, (Jan 1976).

[GJ2] Garey, M. R. and D. S. Johnson, "Computers and Intractability: A Guide to the Theory of NP-Completeness," W. H. Freeman and Company, San Francisco, 1980.

[GJ3] Garey, M. R. and D. S. Johnson, Unpublished results referenced in [GJ2].

[HS] Horowitz, E. and S. Sahni, "Fundamentals of Computer Algorithms," Computer Science Press, Inc., 1978.

[JL] Johnson, D. B. and J. M. Lafuente, "Controlled Single Pass Classification Algorithm with applications to Multilevel Clustering," Scientific Report #ISR-18, Information Science and Retreival, Cornell University, Oct 1970.

[R] Rohlf, F. J. "Single Link Clustering Algorithms," RC 8569 (#37332) Research Report, IBM, T. J. Watson Research Center, Nov. 1980.

[K] Karp, R. M., "Reducibility Among Combinatorial Problems," In Complexity of Computer Computations, R. E. Miller and J. W. Thatcher, Eds, Plenum Press, N. Y. 1972, p.p. 85-104.

[M] Meisel, W. S., "Computer-Oriented Approaches to Pattern Recognition," Academic Press, New York, 1972.

[SG] Sahni, S. and T. Gonzalez, "P-Complete Approximation Problems," JACM, 23, 555-565, 1976.

[S1] Salton, G. "The Smart Retreival System, Experiments in Automatic Document Processing," Prentice-Hall, New Jersey (1971).

[S2] Salton, G., "Dynamic Information and Library Processing," Prentice-Hall, New Jersey (1975).

[Sh] Shamos, M.I., "Geometry and Statistics: Problems at the Interface," in J. F. Traub (ed), Algorithms and Complexity: New Directions and Recent Results, Academic Press, New York, 251-280, 1976.