Notes on the Minimum Set Coverage Problem

Introduction

In the Minimum Set Coverage (MSC) problem, we are given m tasks and n non-empty subsets of these tasks. The objective is to select exactly k of these n subsets such that the number of tasks in the union is minimized. [8] establishes the following about MSC:

- MSC is a special case of the Minimum Sub-additive Join problem.
- MSC can be reduced in polynomial time to the Maximum Balanced Complete Bipartite Subgraph (MBCBS) problem.
- MBCBS is NP-hard and admits no PTAS [4].
- So, MSC is NP-hard and admits no PTAS.

For convenience, we assume that every subset is non-empty (otherwise it can be ignored) and that every task is in at least one subset (otherwise it can be removed from the list of tasks). We also assume that $k < n$ to make the problem non-trivial.

MIP Formulation

Next, we present a mixed integer linear programming formulation of MSC.

Notation:

- $M = \{1, \ldots, m\}$: index set of tasks
- $E_i \subseteq M$, $i = 1, \ldots, n$: subsets of tasks in M
- $N = \{1, \ldots, n\}$: index set of subsets
- k: number of sets to be selected, $1 \leq k < n$
- $z_i, i \in N$: variable that takes a value 1 if subset i is selected, 0 otherwise
- $x_j, j \in M$: variable that takes a value 1 if task j is selected, 0 otherwise

Model:

$$\text{Minimize} \sum_{j \in M} x_j$$

$$\text{such that } z_i \leq x_j, \quad i \in N, j \in E_i$$

$$\sum_{i \in N} z_i \geq k$$

$$x \in \{0,1\}^m, z \in \{0,1\}^n$$

Insights based on unimodularity

The structure of the constraint matrix has some interesting properties. First, note that if we eliminate the constraint with the number of sets, the constraint matrix is totally unimodular (TU) due to a result in [6]. Hence, every minor of this matrix will have a value in $\{-1, 0, 1\}$.

If we consider the constraint matrix including this constraint, the row with this constraint will be a sequence of -1 values followed by a sequence of 0 values. The number of -1 values in this row is the number of subsets in the problem. Hence, any minor constructed using values in this row cannot have a value that falls below $-k$ or exceeds k. We can construct examples to show that these bounds are tight. We can also construct examples to show that for a given k, there exist problem instances whose values are any of the integers $\{-k, -k+1, ..., 0, 1, ..., k\}$ (examples to be added later).

Some results relevant to matrices with a structure similar to this problem can be found in [3], [2], [5]. However, it seems that MSC is slightly more general than the classes of problems addressed in these publications. Also, Rico Zenklusen shared the following private communication: he has co-authored a publication that show that any ILP with a 2-modular constraint matrix can be solved in strongly polynomial time. However, it is open whether ILPs with a k-modular constraint matrix, for some fixed $k \geq 3$, can be solved efficiently.

Due to the hardness of MSC, there is no hope of using k-modularity results to solve the problem easily. We can only hope to exploit the boundedness of minors to In fact, we can use this relationship to say that there exist constraint matrices such that each of their minors have absolute determinant value of at most k, but the problem cannot be solved in polynomial time.

Consider the instance with 5 tasks numbered from 0 to 5 and 4 subsets $\{0, 1, 2\}$, $\{2, 3, 4\}$, $\{2, 4\}$, {0, 1, 3, 4}. The constraint matrix is:

$[1, 0, 0, 0, -1, 0, 0, 0, 0]$
$[1, 0, 0, 0, 0, -1, 0, 0, 0]$
$[1, 0, 0, 0, 0, 0, -1, 0, 0]$
$[0, 1, 0, 0, 0, 0, -1, 0, 0]$
$[0, 1, 0, 0, 0, 0, 0, -1, 0]$
$[0, 1, 0, 0, 0, 0, 0, 0, -1]$
$[0, 0, 1, 0, 0, 0, -1, 0, 0]$
$[0, 0, 1, 0, 0, 0, 0, 0, -1]$
$[0, 0, 0, 1, -1, 0, 0, 0, 0]$
$[0, 0, 0, 1, 0, -1, 0, 0, 0]$
$[0, 0, 0, 1, 0, 0, 0, -1, 0]$
$[0, 0, 0, 1, 0, 0, 0, 0, -1]$
$[-1, -1, -1, -1, 0, 0, 0, 0, 0]$

For every integer h in $\{-4, -3, ..., 3, 4\}$, the above matrix has a minor whose determinant

has value h.

Polyhedral properties of the MIP solution space

While we usually don't really desire a solution that selects more than k subsets, consider the following cases for any such solution:

- If any of the subsets is completely covered by the other subsets, it can be removed from the selection, reducing the number of subsets. This reduction can never reduce the number of selected subsets below k due to the lower bound.
- If a subset E_p contains a task not in any other subset in the solution, it can be dropped to reduce the objective and yield a better solution.

Hence, considering this relaxed version of the model will not yield a solution with a worse objective value than the one with a strict equality for the cardinality constraint. We use this model as it simplifies the polyhedral analysis of certain families of inequalities. There is also an indirect reason to consider this version. The randomized algorithm proposed in [7] seems to offer an efficient approximation algorithm for the minimization of a submodular function with a cardinality constraint. This algorithm only uses the cardinality constraint as a lower bound and not as a strict equality. It may be easier to relate MSC with that paper with the inequality constraint, although this is only speculative.

Let P denote the feasible region of the above MIP and $conv(P)$ its convex hull. For $j \in M$, let $A_j \subseteq N$ be the set of indices of E_i sets that contain task j. Throughout the rest of the article, we write any point in P in the order $(x_1, ..., x_m, z_1, ..., z_n)$. Minor note: the facets of $conv(P)$ for small instances were primarily identified using [1].

Lemma 1. *Let $T \subseteq M$ such that for $j \in T$, $|A_j| \geq n - k + 1$. Then $x_j = 1$, $j \in T$ for all feasible points in $conv(P)$.*

Proof. The proof follows from the Pigeon-hole principle. In other words, for any $j \in T$, there are at most $k - 1$ sets that do not contain j and we are required to select k sets. □

Theorem 2. $dim(P) = n + m - |T|$.

Proof. $dim(P) \leq n + m$ as at most $n + m$ variables are required to describe P. We also know that there are $|T|$ linearly independent hyperplanes that pass through $conv(P)$. These are the fixed-value constraints of Lemma 1. Hence, $dim(P) \leq n + m - |T|$. To establish equality, we will prove that any general hyperplane that passes through $conv(P)$ has to be a linear combination of the $|T|$ hyperplanes defining $conv(P)$. To that end, assume that a general hyperplane of the form

$$\sum_{j \in M} a_j x_j + \sum_{i \in N} b_i z_i = c \tag{1}$$

passes through $conv(P)$. Consider a point p_0 with all x_j and z_i set to 1. As $p_0 \in conv(P)$, substituting it in (1) yields

$$\sum_{j \in M} a_j + \sum_{i \in N} b_i = c.$$

Constructing p_1 by changing any z_e in p_0 to 0 for $e \in N$, we see that it remains feasible for P (as $k < n$). Substituting p_1 into (1) gives

$$\sum_{j \in M} a_j + \sum_{i \in N \setminus \{e\}} b_i = c.$$

From these two identities, we can see that $b_i = 0$ for all $i \in N$. Using this result and the hyperplanes that already define P, (1) should be a linear combination of the hyperplanes $x_j = 1$ for $j \in T$ and

$$\sum_{j \in M \setminus T} a_j x_j = c_1 \tag{2}$$

for some constant c_1. For any task $j \in M \setminus T$, we know that j can be in at most $n - k$ subsets, i.e. $|A_j| \leq n - k$. This means that there exist at least k subsets that do not contain task j. Given one such task $q \in M \setminus T$, construct a point p_2 with $x_q = 0$, $z_i = 0$ for $i \in A_q$ and all other variables set to 1. As $p_2 \in conv(P)$, substituting it in (2) yields

$$\sum_{j \in M \setminus (T \cup \{q\})} a_j = c_2.$$

Construct p_3 with $x_q = 1$ and every other variable has the same value as in p_2. As x_i variables do not have any variable upper bounds, p_3 remains feasible. Substituting it in (2), we get

$$a_q + \sum_{j \in M \setminus (T \cup \{q\})} a_j = c_2.$$

This implies that $a_q = 0$ for each $q \in M \setminus T$. This proves that any general hyperplane of the form (1) has to be a linear combination of the hyperplanes $x_j = 1, j \in T$ and concludes the proof. □

At this point, it is not clear whether even simple inequalities such as $x_i \leq 1$ and $z_i \geq 0$ are necessary to describe P. We try to clarify this in the following two theorems.

Theorem 3. *For $t \in M$, the inequality $x_t \leq 1$ is facet-defining for $conv(P)$ if $|A_t| \leq n - k$.*

Proof. Given $t \in M$, consider the face $F_t = \{(x, z) \in conv(P) : x_t = 1\}$ and a general hyperplane of the form

$$\sum_{j \in M} a_j x_j + \sum_{i \in N} b_i z_i = c \tag{3}$$

passing through F_t. As the points p_0 and p_1 used in Theorem 2 have all x_i variables set to 1, these points are also feasible for F_t. Hence, (3) has to be a linear combination of the hyperplanes $x_j = 1$ for $j \in T$, $x_t = 1$ and the hyperplane

$$\sum_{j \in M_t} a_j x_j = c_3$$

where $M_t = M \setminus (T \cup \{t\})$. Similar to the part of the proof of Theorem 2 where we used p_2, for every $q \in M_t$, we can build points that are feasible for F_t with $x_q = 0$ and $x_q = 1$. This pair of points and the above identity show that $a_q = 0$ for all $q \in M_t$. Hence, any hyperplane passing through F_t has to be a linear combination of only the hyperplanes $x_j = 1$ for $j \in T$ and $x_t = 1$. □

As the tasks in T will always be selected, they can be substituted out of the MIP. As sets that contain only tasks from T will always be selected, they can also be removed from the problem if the value of k is correspondingly reduced. For convenience, we assume from this point forward that $T = \emptyset$, i.e. $|A_j| \leq n - k$ for all $j \in M$. Note that $dim(P)$ becomes $n + m$ and the dimensions of its facets become $n + m - 1$ with this assumption. We also define the following additional notation that will be used in some proofs: for any subset index $i \in N$, let $\overline{E_i} = M \setminus E_i$ be the set of tasks not in E_i.

Theorem 4. *The inequality*

$$\sum_{i \in N} z_i \geq k \tag{4}$$

is facet-defining for conv(P).

Proof. Consider a general hyperplane

$$\sum_{j \in M} a_j x_j + \sum_{i \in N} b_i z_i = c \tag{5}$$

passing through the face C of P defined by the equality form of (4). As we can always build a point by selecting subsets E_i for $i \in K$ with $|K| = k$, we can assign $z_i = 1$ for $i \in K$, $x_j = 0$ for all j and all other variables to 0 for this point. As this point is feasible for C, substituting this point in (5) reduces it to

$$\sum_{i \in K} b_i z_i = c_1$$

for some constant c_1. As $k < n$, we can find a subset index $q \notin K$. Setting $z_r = 0$ for some $r \in K$ and $z_q = 1$ in the previous point gives a new feasible point that is also in K. Substituting this point in the previous identity yields

$$\sum_{i \in K \setminus \{r\}} b_i z_i = c_1.$$

This shows that $b_r = 0$ for any r in K. Hence, there is no hyperplane other than the equality form of (4) that passes through C. This makes (4) facet-defining for P. □

Interestingly, it turns out that $z_i \geq 0$ is not always facet-defining and requires certain restrictions on the subset E_i. Specifically, a sufficient and necessary condition for z_i to be facet-defining is that for every task j not in E_i, at most $n-k-1$ sets can contain j. A simple case of this condition is when E_i contains all tasks: $z_i \geq 0$ will define a facet for this subset as there is no task that is not in E_i. We formally state and prove this result in the following theorem.

Theorem 5. *For $p \in N$, $z_p \geq 0$ is facet-defining for $conv(P)$ if and only if:*

(i) $k < n-1$,

(ii) $|A_j| \leq n-k-1$ *for all* $j \in \overline{E_p}$.

Proof. Consider the case $k = n-1$. This means that for all $j \in \overline{E_p}$, $|A_j| \leq 0$ which is not possible as we have assumed that every task belongs to at least one set (otherwise the task can trivially be ignored). Hence, the second assumption of this theorem is only valid if $k < n-1$.
Given $p \in N$, consider the face $G_p = \{(x,z) \in conv(P) : z_p = 0\}$. Any general hyperplane of the form $ax + bz = c_0$ passing through G_p has to be a linear combination of $z_p = 0$ and the hyperplane

$$\sum_{j \in M} a_j x_j + \sum_{i \in N \setminus \{p\}} b_i z_i = c. \tag{6}$$

The point with all z_i and x_j variables set to 1 except z_p is feasible for G_p. Substituting this point in (7) yields

$$\sum_{j \in M} a_j + \sum_{i \in N \setminus \{p\}} b_i z_i = c.$$

As $k < n-1$, we can build a feasible point from the previous one by setting $z_q = 0$ for any $q \in N \setminus \{p\}$ without making the point infeasible for G_p. Substituting this point in (6) and using the above identity shows that $b_i = 0$ for all $i \in N \setminus \{p\}$. Therefore, (7) reduces to

$$\sum_{j \in M} a_j x_j = c_1.$$

To aid the next part of the proof, we write this equality as

$$\sum_{j \in E_p} a_j x_j + \sum_{j \in \overline{E_p}} a_j x_j = c_1. \tag{7}$$

For any $j_1 \in M$, if we can build a point in P that is feasible for G_p such that $x_{j_1} = 0$ and all other x_j variables being 1, this will result in a_{j_1} being zero. If we can find such a point for each $j \in M$, this will prove the sufficiency part of the Theorem. We now try to build points with this structure.
Consider any task $j_1 \in E_p$. As we have assumed that $|A_j| \leq n-k$ for all $j \in M$, we can find k subsets that do not contain j_1. Consider a point with z_i set to 1 for these k subsets, the other $n-k$ z_i variables set to 0, $x_{j_1} = 0$ and $x_j = 1$ for $j \neq j_1$. As E_p contains j_1, z_p

will be among the variables set to 0. Hence, this point is feasible. Substituting this point in (7) shows that $a_j = 0$ for $j \in E_p$. Hence, (7) reduces to

$$\sum_{j \in \overline{E_p}} a_j x_j = c_1. \tag{8}$$

Now, consider point in G_p with $x_q = 0$ for $q \in M \setminus E_p$. Hence, $|A_q| \leq n - k - 1$. This means that at most $n - k - 1$ z_i variables are forced to 0 other than z_p. Hence, there exist k z_i variables can be set to 1 meaning that such a point can be constructed. Substituting this point in (7) and the point obtained by changing x_q to 1 in the previous point in (7) we get $a_q = 0$ for $q \in M \setminus E_p$ meaning that $dim(G_p) = n + m - 1$. This proves the sufficiency part of the Theorem.
To prove necessity, consider p such that there exists $r \in \overline{E_p}$ with $|A_r| = n - k$. Now, if we would like to construct a point in G_p with $x_r = 0$, we need to set $n - k$ z_i variables to 0 in addition to z_p totaling to $n - k + 1$ z_i variables. This means that at most $k - 1$ z_i variables can be allowed to be 1 meaning that such a point cannot be in G_p as it cannot be in $conv(P)$. Hence, $x_r = 1$ for every point in G_p resulting in $dim(G_p) \leq n + m - 2$. This completes both sufficiency and necessity. □

Corollary 6. *For $p \in N$, $z_p \geq 0$ represents a face of dimension $n + m - 1 - d_p$ where d_p is the number of tasks $j \in \overline{E_p}$ such that $|A_j| = n - k$ (not completely sure).*

Proof. Proof seems like it should basically be the necessity part of the previous Theorem repeated for every task not in E_p. But this needs to be verified. □

Now, we try to explore the situation where the conditions of Theorem 5 are not satisfied.

Theorem 7. *For $t \in M$ and $p \in N$ such that E_p does not contain t, the inequality*

$$1 - x_t \leq z_p \tag{9}$$

is valid for P if $|A_t| = n - k$. Further, (9) is facet-defining for $conv(P)$ if $k < n - 1$ and t is the only task not in E_p with $|A_t| = n - k$. The condition $k < n - 1$ is also necessary for (9) to be facet-defining.

Proof. To prove validity, we need to analyze the following two cases for any point in P:

- $x_t = 1$: (9) adds no restrictions for these points and remains valid.
- $x_t = 0$: as $n - k$ subsets contain t, the z_i variables corresponding to these subsets are forced to be zero. As there are only k subsets remaining, the z_i variables corresponding to them are forced to be one for the point to remain feasible for P. As E_p is a set that does not contain t, z_p is forced to 1. This point also satisfies (9).

This proves the validity of (9) for P. We now prove the sufficient condition for (9) to be facet-defining for P. Let F represent the face of $conv(P)$ in which (9) is satisfied at equality. Assume that the general hyperplane

$$\sum_{j\in M} a_j x_j + \sum_{i\in N} b_i z_i = c. \tag{10}$$

passes through F. Consider a point with all variables set to 1 except z_p which is set to zero. As this point is feasible for F, substituting it in (10) yields

$$\sum_{j\in M} a_j + \sum_{i\in N\setminus\{p\}} b_i = c. \tag{11}$$

Substituting the expression for c from (11) in (10) modifies it to

$$\sum_{j\in M} a_j(x_j - 1) + \sum_{i\in N\setminus\{p\}} b_i(z_i - 1) + b_p z_p = 0. \tag{12}$$

Now construct the point v_2 in which $z_p = 1$, $x_t = 0$, $z_i = 0$ for $i \in A_t$ and all other variables are set to 1. Clearly, $v_2 \in F$. Substituting this point in (12) yields

$$b_p = a_t + \sum_{i\in A_t} b_i.$$

Based on this relationship, (12) can be written as

$$\sum_{j\neq t} a_j(x_j - 1) + \sum_{i\neq p, i\notin A_t} b_i(z_i - 1) + a_t(x_t + z_p - 1) + \sum_{i\in A_t} b_i(z_i + z_p - 1) = 0. \tag{13}$$

Now consider the point v_3^q in which $z_p = 0$, $z_q = 0$ for some $q \neq p$ and all other variables are 1. This point is feasible for P as our assumption ($k < n - 1$) allows us to de-select at least 2 subsets. Also, $v_3^q \in F$ as $z_p + x_t = 1$. Substituting this point in (12) yields $b_q = 0$ for any $q \neq p$. Using this in (13) reduces it to

$$\sum_{j\neq t} a_j(x_j - 1) + a_t(x_t + z_p - 1) = 0.$$

Hence, any general hyperplane has to be a linear combination of the equality that defines the face F and the hyperplane

$$\sum_{j\neq t} a_j(x_j - 1) = 0. \tag{14}$$

Consider the set of points in P for which $x_t = 1$ and $z_p = 0$. These points are all in F as they lie on its defining hyperplane. Among these points, select a point for which $x_j = 0$ for some $j \in E_p$ and all other x_j values are 1. For this point, at most $n - k - 1$ z_i variables other than z_p are forced to zero as $|A_j| \leq n - k$ for all j and E_p contains j. Substituting this point in (14) gives us $a_j = 0$ for $j \in E_p$. This reduces (14) to

$$\sum_{j\in \overline{E_p}\setminus\{t\}} a_j(x_j - 1) = 0. \tag{15}$$

We know that for every task j not in E_p other than t, $|A_j| \leq n - k - 1$. So, we can build a point with $x_t = 1$, $z_p = 0$, $x_j = 0$, $z_i = 0$ for $i \in A_j$ and all other variables with

value 1. Substituting this point in (15) yields $a_j = 0$ for $j \in \overline{E_p} \setminus \{t\}$ and concludes the proof.

Now, we consider the case $k = n - 1$. In this case, A_t contains a single index. Let this index be q. Note that the inequality $z_p + z_q \geq 1$ is valid for P in this case as at most one z_i can be set to zero. Now, for any point lying on (9), there are two cases:

- $z_p = 1$: This implies that $x_t = 0$ which means $z_q = 0$. This point satisfies $z_p + z_q = 1$.
- $z_q = 0$: This implies that $z_p = 1$. Hence, this point also satisfies $z_p + z_q = 1$.

In other words, any point feasible for P and lying on (9) also lies on another valid inequality for P. This proves that (9) cannot be facet-defining for $conv(P)$ when $k = n - 1$. □

In the next Theorem, we establish certain inequalities that arise when the conditions of Theorem 7 are relaxed even further by allowing multiple tasks outside E_p to have $|A_t| = n - k$. However, this requires some additional notation. For any subset index $p \in N$, let $\overline{E_p}' \subseteq \overline{E_p}$ be the set of indices such that for $j \in \overline{E_p}'$, $|A_j| = n - k$.

Now, we would like to classify $\overline{E_p}'$ in a very specific way. Given a sequence of tasks in $\overline{E_p}'$, build up the sets B_p and $B'_p = \overline{E_p}' \setminus B_p$ as follows:

- Initialize B_p as the empty set.
- For each task t, add t to B_p if either B_p is empty, or $A_t \neq A_j$ for any $j \in B_p$.
- Otherwise, add t to B'_p.

Note: as will use B_p to define some facets, this looks suspiciously like sequence-dependent lifting. Later, we will define even messier sets C_p and D_p for which even validity becomes painful. If we can isolate and clarify this lifting procedure, it seems that we may be able to we should be able to define a nice step-by-step process to specify variables that can be lifted into an inequality like $z_i \geq 0$ that is not facet-defining.

Theorem 8. *Consider any $B_p \subseteq \overline{E_p}'$ such that for $j_1, j_2 \in B_p$, $A_i \neq A_j$ and for any $j_1 \in \overline{E_p}' \setminus B_p$, $A_{j_1} = A_{j_2}$ for some $j_2 \in B_p$. The inequality*

$$\sum_{j \in B_p} (1 - x_j) \leq z_p \tag{16}$$

is valid for P.

Proof. For any point in P, there are two possible cases:

(i) If $z_p = 0$, this means that E_p cannot be selected. For the point to be feasible, at least k z_i variables among the remaining $n - 1$ z_i variables should be selected. If any of the variables x_j in B_p is zero in this point, it means the $n - k$ z_i variables (for $i \in A_j$) are forced to zero. As p is not in A_j, this results in $n - k + 1$ variables being zero at this point, or only $k - 1$ z_i variables that can be 1. As this violates the cardinality lower bound, every x_j corresponding to $j \in B_p$ must be one. Hence, (16) is satisfied by points in P with $z_p = 0$.

(ii) $z_p = 1$: The main thing to observe in this case is as follows: for any $(j_1, j_2) \in B_p$, $A_{j_1} \neq A_{j_2}$ by definition. In other words, for any $(j_1, j_2) \in T$, $|A_{j_1} \cup A_{j_2}| \geq n - k + 1$. Now, (16) will not be valid only if at least two x_j variables are 0 for some feasible point in P with $z_p = 1$. Let the two variables be x_{j_1} and x_{j_2}. Due the constraints of P, we have $z_i = 0$ for $i \in A_{j_1} \cup A_{j_2}$. However, this means that the total number of sets available for selection is at most $(n-1) - (n-k+1) = k-2$ while we need $k-1$ sets to be selected for the corresponding point to be feasible. This is a contradiction.

□

Theorem 9. *The inequality* (16) *is facet-defining for conv*(P) *if and only if* $k < n - 1$.

Proof. First we assume that $k < n - 1$. Let F represent the face of P in which (16) is satisfied at equality. Consider a general hyperplane

$$\sum_{j \in M} a_j x_j + \sum_{i \in N} b_i z_i = c \tag{17}$$

that passes through every point in F. We will prove that (17) has to be a scalar multiple of the equality form of (16). Consider the point v_p in which $z_p = 0$ and all other variables are 1. This point lies on (16) and is feasible for P. So, $v_p \in F$. Subsituting this point in (17) reduces it to

$$\sum_{j \in M} a_j (x_j - 1) + \sum_{i \in N \setminus \{p\}} b_i (z_i - 1) + b_p z_p = 0. \tag{18}$$

As $k < n - 1$, at least one other z variable (say z_q) can be set to 0 in v_p without affecting its feasibility for F. Substituting this new point in (18) yields $b_q = 0$. Hence, (18) reduces to

$$\sum_{j \in M} a_j (x_j - 1) + b_p z_p = 0. \tag{19}$$

Now, consider the x_j variables in (19) with $j \in E_p$. Corresponding to any such variable (say x_r), we can construct a point v_2 with $x_r = 0$, $z_i = 0$ for $i \in A_r$ and all other variables equal to 1. As $|A_r| \leq n - k$ such a point will be feasible for P. Also, as $z_p = 0$ (due to p being in A_r) and all x_i variables in B_p are 1, this point also lies on (16). Substituting this point in (19) shows that $a_r = 0$ for any $r \in E_p$. Hence, (19) reduces to

$$\sum_{j \in \overline{E_p}} a_j (x_j - 1) + b_p z_p = 0. \tag{20}$$

Next, consider any $r \in \overline{E_p}$ such that $|A_r| \leq n - k - 1$ and construct a point with $x_r = 0$, $z_i = 0$ for $i \in A_r$ and all other variables set to 1. As $|A_r| \leq n - k + 1$, at most $n - k - 1$ sets out of $n - 1$ sets have been set to zero, meaning that at least $k + 1$ sets are available for selection (including E_p). This point is feasible for P. Setting $z_p = 0$ in this point means that it lies on (16). Substituting such points successively in (20) reduces it to

$$\sum_{j \in \overline{E_p}'} a_j (x_j - 1) + b_p z_p = 0. \tag{21}$$

For any $r \in B_p$, construct a point with $x_r = 0$, $z_i = 0$ for $i \in A_r$ and all other variables equal to 1. As $|A_r| = n - k$, such a point is feasible for P and lies on (16) as $z_p = 1$. Substituting this point in (14) gives $a_r = b_p$. Substituting all such points successively in (14) reduces it to

$$b_p \left(\sum_{j \in B_p} x_j + z_p - |B_p| \right) + \sum_{j \in \overline{E_p}' \setminus B_p} a_j (x_j - 1) = 0. \tag{22}$$

We know that for any $r \in \overline{E_p}' \setminus B_p$, there exists an index $t \in B_p$ such that $A_r = A_t$. Hence, corresponding to every index r, set $x_r = 0$, $x_t = 0$ $z_i = 0$ for $i \in A_r (= A_t)$ and all other variables equal to 1. As $|A_r| = n - k$, only $n - k$ z variables have been set to 0, meaning that k variables can be set to 1 making this point feasible for F. Substituting this point in (22) gives $a_r = 0$, which concludes the proof.

Now, we consider the case $k = n - 1$. In this case, for any $i_1, i_2 \in N$, the inequality $z_{i_1} + z_{i_2} \geq 1$ is valid for P as at most one z_i can be set to zero. Also, $|A_r| = 1$ for $r \in B_p$. Let this single index be q_r. Now, for any point lying on (16), there are two cases:

(i) $z_p = 1$: This implies that $x_j = 0$ for some $j \in B_p$ which means $z_{q_j} = 0$. This point also lies on $z_p + z_{q_j} \geq 1$.

(ii) $z_p = 0$: This implies that $x_j = 1$ for all $j \in B_p$. This point also lies on the inequalities $x_j \leq 1, j \in B_p$.

In other words, every point lying on (16) also lies on another valid inequality for P. Therefore, (16) cannot be a facet. □

Please check the appendix for some inequalities that relax the conditions of the above inequalities even further. But they do get much messier.

Alternative perspective on MSC based on submodular functions

Given a finite set Ω and a function $f : 2^\Omega \to \mathbb{R}$ is said to be submodular if any of the following three equivalent conditions are satisfied:

1. For every $X, Y \subseteq \Omega$ with $X \subseteq Y$ and every $x \in \Omega \backslash Y$, we have that $f(X \cup \{x\}) - f(X) \geq f(Y \cup \{x\}) - f(Y)$.
2. For every $S, T \subseteq \Omega$, $f(S) + f(T) \geq f(S \cup T) + f(S \cap T)$.
3. For every $X \subseteq \Omega$ and $x_1, x_2 \in \Omega \backslash X$, $f(X \cup \{x_1\}) + f(X \cup \{x_2\}) \geq f(X \cup \{x_1, x_2\}) + f(X)$.

A submodular set function f is said to be *monotone* if for $S \subseteq T \subseteq \Omega$, $f(S) \leq f(T)$. Given $j \in S$, the *marginal value* of j with respect to S is defined as $f_S(j) = f(S \cup \{j\}) - f(S)$. The *total curvature* of f is defined as

$$c = 1 - \min_{S, j \notin S} \frac{f_S(j)}{f_\varnothing(j)}.$$

Minimum Set Coverage Problem as a Submodular Function

Consider the set of subset indices N as the finite set Ω. Any set from 2^{Ω} will be a set of subset indices. Define a function based on these subsets as

$$f(K) = \left|\cup_{i \in K} E_i\right|, K \subseteq N.$$

Lemma 10. *f is submodular.*

Proof. Given $X \subseteq Y \subseteq N$ and any subset index $x \in N \setminus Y$, we need to prove that

$$\left|\cup_{i \in X \cup \{x\}} E_i\right| - \left|\cup_{i \in X} E_i\right| \geq \left|\cup_{i \in Y \cup \{x\}} E_i\right| - \left|\cup_{i \in Y} E_i\right|. \tag{23}$$

Given the subset index x, consider the corresponding subset $E_x \subseteq M$ and a task $t \in E_x$. The following cases are possible for t:

- $t \in E_{i_1}$ for some $i_1 \in X$: as $X \subseteq Y$, i is also in Y, meaning that this task will not change the value of any of the four terms in (23).
- $t \notin E_i$ for any $i \in X$, but $t \in E_{i_1}$ for some $i \in Y$: LHS increases by 1 as this is a new task w.r.t X, but RHS is unchanged as the task is already accounted for with Y. Hence, for such a task, LHS increases while RHS remains the same.
- $t \notin E_i$ for any $i \in Y$: both LHS and RHS increase by 1.

Hence, LHS is always greater than or equal to RHS for any set of tasks E_x. □

Lemma 11. *f is monotone.*

Proof. Given $S \subseteq T \subseteq N$, $\left|\cup_{i \in S} E_i\right| \leq \left|\cup_{i \in T} E_i\right|$ is always true as T has all subset indices in S and possibly more. □

Hence, the minimum set coverage problem is the problem of minimizing a monotone submodular function over a cardinality constraint.

Curvature of f

Given $S = \{1, \ldots s\} \subseteq N$ and $t \notin S$, we would like to find the value of $c = 1 - \min_{S, j \notin S} \frac{f_S(j)}{f_\varnothing(j)}$ where the function $f_S(t)/f_\varnothing(t)$ is given by

$$\frac{|E_1 \cup \ldots \cup E_s \cup E_t| - |E_1 \cup \ldots \cup E_s|}{|E_t|}.$$

Some cases:

- E_t is completely disjoint from $E_1, \ldots, E_s$: in this case, the function evaluates to $|E_t|/|E_t|$ which is 1. So, the total curvature of f is 0 if all sets are disjoint.

- Every task in E_t belongs to one of the sets $E_1, \ldots, E_s$: in this case, the numerator evaluates to 0 meaning that the curvature is 1. As f is simply a counting function and always evaluates to a non-negative value, the curvature of f is 1 even if there exists one such set.

- Consider the case $N = \{1, 2\}$ with two subsets $E_1 \subseteq M$ and $E_2 \subseteq M$ with $E_1 \neq E_2$, $|E_1| = |E_2| = e$. Let E_1 have u tasks not in E_2 and E_2 have u tasks not in E_1. Also assume that $h > 0$ is the number of common elements of E_1 and E_2. Note that $e = u + h$ by construction (it is important for h to be non-zero for this construction to work).

  For these sets, we have:

$$\frac{|E_1 \cup E_2| - |E_1|}{|E_2|} = \frac{|E_1 \cup E_2| - |E_2|}{|E_1|} = \frac{(2u + h) - e}{e} = \frac{u}{e}.$$

Hence, there exist instances for which the curvature of f can be any rational number in $[0, 1]$.

Research Questions

- Confirm that the problem is NP-hard (most likely true due to [8]).
- Confirm that the problem has no PTAS (most likely true due to [8] and [4]).
- Is this problem well-researched in the submodular community? It seems to simply be a submodular function with a cardinality constraint. Can the special properties of this function be used to do something more?
- The physical structure of the facets of small instances show that $z_i \geq 0$ can be lifted to build better facets. Can we build a general family like this?
- Computationally, can CPLEX and some submodular algorithm be compared?
- Can the submodular approach yield some insight into the polyhedral properties of the MIP form?
- What sizes of problems do we need for CPLEX to start struggling?

References

[1] Ewgenij Gawrilow and Michael Joswig. polymake: a framework for analyzing convex polytopes. In Gil Kalai and Günter M. Ziegler, editors, *Polytopes — Combinatorics and Computation*, pages 43–74. Birkhäuser, 2000.

[2] Dmitry Gribanov and Sergey Veselov. On integer programing with bounded determinants. *arXiv preprint arXiv:1505.03132*, 2015.

[3] DV Gribanov. On integer program with bounded minors and flatness theorem. *arXiv*, 2013.

[4] Subhash Khot. Ruling out ptas for graph min-bisection, dense k-subgraph, and bipartite clique. *SIAM Journal on Computing*, 36(4):1025–1071, 2006.

[5] Balazs Kotnyek. *A generalization of totally unimodular and network matrices.* PhD thesis, London School of Economics and Political Science (United Kingdom), 2002.

[6] George L Nemhauser and Laurence A Wolsey. Integer programming and combinatorial optimization. *Wiley, Chichester. GL Nemhauser, MWP Savelsbergh, GS Sigismondi (1992). Constraint Classification for Mixed Integer Programming Formulations. COAL Bulletin*, 20:8–12, 1988.

[7] Zoya Svitkina and Lisa Fleischer. Submodular approximation: Sampling-based algorithms and lower bounds. *SIAM J. Comput.*, 40(6):1715–1737, December 2011.

[8] Staal A. Vinterbo. A stab at approximating minimum subadditive join. In Frank Dehne, Jörg-Rüdiger Sack, and Norbert Zeh, editors, *Algorithms and Data Structures: 10th International Workshop, WADS 2007, Halifax, Canada, August 15-17, 2007. Proceedings*, pages 214–225. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.

Appendix

Note: the proofs for these inequalities seem to work, but still need to be verified if they will be added to the paper. Hopefully, a lifting procedure will make them obsolete.

Theorem 12. *For $p \in N$, let $E'_p \subseteq E_p$ such that for all $j_1, j_2 \subseteq E'_p$, $|(A_{j_1} \setminus \{p\}) \cup (A_{j_2} \setminus \{p\})| \geq n - k$. Let $C_p \subseteq E'_p$ such that for all $j_1, j_2 \in C_p$, $A_i \neq A_j$ and for all $j_1 \in E'_p \setminus C_p$, there exists $j_2 \in C_p$ such that $A_{j_1} = A_{j_2}$. Then, the inequality*

$$\sum_{j \in C_p} (x_j - 1) \geq z_p - 1 \tag{24}$$

is valid for P and facet-defining for $conv(P)$ if $k < n - 1$.

Proof. For validity we consider two cases:

(i) $z_p = 1$: forces $x_j = 1$ for all $j \in C_p (\subseteq E_p)$ which is true.

(ii) $z_p = 0$: the inequality is violated when at least two x_j variables are 0. Let any two such variables be x_{j_1}, x_{j_2}. We have $|(A_{j_1} \setminus \{p\}) \cup (A_{j_2} \setminus \{p\})| \geq n - k$ by definition, which means that $n-k+1$ variables have been forced to zero. This makes the resulting point infeasible as at least k z_i variables have to be equal to 1.

Let F be the face of P in which (24) holds at equality. Consider a general hyperplane

$$\sum_{j \in M} a_j x_j + \sum_{i \in N} b_i z_i = c \tag{25}$$

passing through F. The point in P with all variables equal to 1 lies on (24). Substituting in (25) results in a reduction to

$$\sum_{j\in M} a_j(x_j-1)+\sum_{i\in N} b_i(z_i-1)=0. \tag{26}$$

As $|A_j|\leq n-k$ for all $j\in M$ and $k<n-1$, any one z_q $(q\neq p)$ can be set to zero with all other variables equal to 1 to generate a point in F. Substituting any such point in (26) results in $z_q=0$. Hence, (18) becomes

$$\sum_{j\in M} a_j(x_j-1)+b_p(z_p-1)=0. \tag{27}$$

For any $r\neq E_p$, we can construct a point in F with $x_r=0$ and $z_i=0$ for $i\in A_r$ and all other variables equal to 1. Using such points, we can reduce (27) to

$$\sum_{j\in E_p} a_j(x_j-1)+b_p(z_p-1)=0. \tag{28}$$

For any $r\in C_p$, we have $|A_r|\leq n-k$. Hence, the point with $x_r=0$, $z_i=0$ for $i\in E_r$ and all other variables equal to 1 is in F. Substituting this point in (20) gives us the relation $a_r=-b_p$. Based on such relationships for all $j\in C_p$, we get the following reduction of (28):

$$b_p\left(\sum_{j\in C_p}(x_j-1)-(z_p-1)\right)+\sum_{j\in E_p\setminus C_p} a_j(x_j-1)=0. \tag{29}$$

For all $r_1\in E_p\setminus E'_p$ we have $|(A_{r_1}\setminus\{p\})\cup(A_{r_2}\setminus\{p\})|\leq n-k-1$ for some $r_2\in E'_p$. Hence, we can construct a point in F with $x_{r_1}=x_{r_2}=z_p=0$ and all other variables equal to 1. Substituting this point in (29) yields $a_{r_1}=0$ for $r_1\in E_p\setminus E'_p$. Hence, (29) becomes

$$b_p\left(\sum_{j\in C_p}(x_j-1)-(z_p-1)\right)+\sum_{j\in E'_p\setminus C_p} a_j(x_j-1)=0. \tag{30}$$

For the final step of the reduction, we observe that for all $r_1\in E'_p\setminus C_p$, there exists $r_2\in C_p$ such that $A_{r_1}=A_{r_2}$. Hence, the point with only the variables x_{r_1}, x_{r_2}, z_p and z_i for $i\in A_{r_1}$ equal to 0 is feasible for F and can be substituted into (30), resulting in $a_j=0$ for $j\in E'_p\setminus C_p$. This concludes the proof. □

Consider any $p_1,p_2\in N$ and construct a set $D_{p_1p_2}$ with the following conditions:

(i) For all $j\in D_{p_1p_2}$, $|A_j\setminus\{p_1,p_2\}|\geq n-k-1$,

(ii) For $j_1,j_2\in D_{p_1p_2}$, $A_{j_1}\setminus\{p_1,p_2\}\neq A_{j_2}\setminus\{p_1,p_2\}$,

(iii) For $j_1\in M\setminus D_{p_1p_2}$ such that $|A_j\setminus\{p_1,p_2\}|\geq n-k-1$, there exists $j_2\in D_{p_1p_2}$ such that $A_{j_1}\setminus\{p_1,p_2\}\neq A_{j_2}\setminus\{p_1,p_2\}$,

(iv) There exists at least one $j_1\in D_{p_1p_2}$ such that $p_1,p_2\notin A_r$.

Let D_1, D_2 be a partition of $D_{p_1p_2}$ such that for $j \in D_1$, $|A_r \setminus \{p_1, p_2\}| = n - k - 1$. For the sake of convenience, we refer to the set $D_{p_1p_2}$ as simply D. The following assumptions and observations will be used in proofs for inequalities that can be obtained using these sets:

(i) We assume that P is full-dimensional (if not, any equality defining the affine hull is of the form $x_j = 1$ and can be used to reduce the dimension of P by one).

(ii) Due to this assumption, $|A_j| \leq n - k$ for all $j \in M$.

(iii) Any $r \in D$ can belong to at most one of the sets E_{p_1} or E_{p_2} by definition of D.

(iv) For any $r \in D_1$, $|A_r| = n - k$ if $r \in E_{p_1}$ or $r \in E_{p_2}$.

(v) For any $r \in D_1$, $|A_r| = n - k - 1$ if $r \notin E_{p_1} \cup E_{p_2}$.

(vi) For any $r \in D_2$, $|A_r| = n - k$ and $r \notin E_{p_1} \cup E_{p_2}$.

Theorem 13. *The inequality*

$$\sum_{j \in D_1} (x_j - 1) + \sum_{j \in D_2} 2(x_j - 1) + z_{p_1} + z_{p_2} \geq 0 \tag{31}$$

is valid for P.

Proof. **(this proof is not correct, but some inequalities with this structure were observed. So, there may still be a way to fix this proof.)** For any point in P, three cases are possible:

(i) $z_{p_1} = z_{p_2} = 0$: The inequality (31) is not valid even if one $x_j = 0$ for $j \in D$. If there is such a point, $x_j = 0$ forces at least $n - k - 1$ z_i variables to zero, making the point infeasible.

(ii) $z_{p_1} + z_{p_2} = 1$: Here, (31) can be violated if either $x_{j_1} = x_{j_2} = x_{j_3} = 0$ for $j_1, j_2, j_3 \in D_1$ or if $x_{j_1} = 0$ for $j_1 \in D$ and $x_{j_2} = 0$ for $j_2 \in D_2$. Validity for these subcases can be proved as follows:

  (a) In the first subcase, even if $j_1, j_2, j_3 \notin E_{p_1} \cup E_{p_2}$, setting the corresponding x_j variables to zero forces at least $n - k - 1 + 2 = n - k + 1$ z_i variables to zero as no two pairs among $A_{j_1} \setminus \{p_1, p_2\}$, $A_{j_2} \setminus \{p_1, p_2\}$, $A_{j_3} \setminus \{p_1, p_2\}$ are identical, making the point infeasible.

  (b) In the second subcase, the worst case we can construct is when $j_1 \in D_1$, $j_2 \in D_2$ and $A_{j_1} \subset A_{j_2}$. Even for such a point, $n - k$ subsets are forced to zero in addition to either z_{p_1} or z_{p_2}, which causes infeasibility.

(iii) $z_{p_1} = z_{p_2} = 1$: Proof is similar to the previous ideas.

□

Theorem 14. *The inequality* (31) *is facet-defining for conv(P) if $k < n - 2$.*

Proof. (only a rough sketch is provided) Consider a general hyperplane

$$\sum_{j\in M} a_j x_j + \sum_{i\in N} b_i z_i = c \tag{32}$$

passing through the face F defined by (31). Set $z_{p_1} = z_{p_2} = 0$ and all other variables to 1 in (32) to reduce it to

$$\sum_{j} a_j(x_j - 1) + \sum_{i\neq p_1,p_2} b_i z_i + b_{p_1} z_{p_1} + b_{p_2} z_{p_2} = 0. \tag{33}$$

As $k < n-2$, any one other z_q can be set to 0 in the previous point. Substituting in (33), we get

$$\sum_{j} a_j(x_j - 1) + b_{p_1} z_{p_1} + b_{p_2} z_{p_2} = 0. \tag{34}$$

Consider the $r \in D$ with $p_1, p_2 \notin A_r$ and $|A_r| = n-k-1$. Set $x_r = 0$, $z_i = 0$ for $i \in A_r$, $z_{p_1} = 0$ and all other variables to 1 in (34). This gives $a_r = b_{p_1}$. In this point, change z_{p_1} to 1 and z_{p_2} to 0. This gives $a_r = b_{p_2}$. Let $b_{p_1} = b_{p_2} = h$. Further, consider any $r \in D_1$ with one of p_1, p_2 in A_r. As $|A_r| = n-k$, we can set $x_r = 0$, $z_i = 0$ for $i \in A_r$ and all other variables to 1 to generate a point feasible for F. Substituting this point in (34) gives $a_r = d$. The same construction can be done for any other $r \in D_1$ too and can be used to obtain the relation $a_r = d$ for $r \in D_1$. Also, performing the same construction for $r \in D_2$ gives the relation $a_r = 2d$. Using these identities reduces (34) to

$$d\left(\sum_{j\in D_1}(x_j - 1) + \sum_{j\in D_2} 2(x_j - 1)\right) + \sum_{j\notin D} a_j(x_j - 1) = 0. \tag{35}$$

If $r \notin D$ and $|A_r \setminus \{p_1, p_2\}| \geq n-k-1$, there exists an $r_1 \in D$ such that $A_r = A_{r_1}$. As we can generate a feasible point in F with $x_{r_1} = 0$ for any $r_1 \in D$, we can simply generate this point and set $x_r = 0$ without affecting its feasibility for F. Substituting this point in (28) yields $a_r = 0$ for any such r.
If $r \notin D$ and $|A_r \setminus \{p_1, p_2\}| \leq n-k-2$, there are two cases:

(i) $r \notin E_{p_1} \cup E_{p_2}$: we can generate a point feasible for F with $x_j = 1$ for $j \in D$, $z_{p_1} = z_{p_2} = 0$ and $x_r = 0$. Substituting this point in (28) gives $a_r = 0$.

(ii) $r \in E_{p_1}$ or $r \in E_{p_2}$: In this case, we generate a feasible point with $x_r = 0$ and $x_{r_1} = 0$ where r_1 is the unique index that must necessarily be in D_1. This allows either z_{p_1} or z_{p_2} to be set to zero as needed and forces $a_r = 0$.

(iii) $r \in E_{p_1}$ and $r \in E_{p_2}$: simply set $x_r = 0$, $i = 0$ for $i \in A_r$ and all other variables to 1. This point when substituted in (35) forces $a_r = 0$.

This proves that any general hyperplane passing through F has to be a scalar multiple of the hyperplane form of (31). □