# Improved Performance of the Greedy Algorithm for the Minimum Set Cover and Minimum Partial Cover Problems

Petr Slavík *

October 15, 1995

**Abstract**

We establish significantly improved bounds on the performance of the greedy algorithm for approximating *minimum set cover* and *minimum partial cover*. Our improvements result from a new approach to both problems. In particular,

(a) we improve the known bound on the performance ratio of the greedy algorithm for general covers without cost, showing that it differs from the classical harmonic bound by a function which approaches infinity (as the size of the base set increases);

(b) we show that, for covers without cost, the performance guarantee for the greedy algorithm is significantly better than the performance guarantee for the randomized rounding algorithm;

(c) we prove that the classical bounds on the performance of the greedy algorithm for complete covers with costs are valid for partial covers as well, thus lowering, by more than a factor of two, the previously known estimate.

**Keywords:** Approximation Algorithms, Combinatorial Optimization, Greedy Algorithm.

*Department of Mathematics, State University of New York at Buffalo, Buffalo, NY 14214, USA.
E-mail: `slavik@math.buffalo.edu`

# 1 Introduction

In the *minimum set cover* problem, the goal is to find a smallest subcover of a cover of a finite "base set". This can be generalized for covers with costs, where the goal is to find a subcover with minimum cost. These problems are NP-hard (see [6] or [3]), hence efficient approximation algorithms are of great interest. One of the best known algorithms for approximating the *minimum set cover* is the so called greedy algorithm. The classical bound on its performance ratio (see e.g. [1], [5], or [8]) is

$$\frac{c_{greedy}}{c_{min}} \leq H(d), \tag{1}$$

where $c_{greedy}$ and $c_{min}$ are the costs of a minimum subcover and the subcover output by the greedy algorithm, respectively, $d$ bounds the size of the sets in the cover, and $H(d) = 1 + \cdots + 1/d$. For general covers (a priori unknown bound on the size of the covering sets), this reduces to

$$\frac{c_{greedy}}{c_{min}} \leq H(m), \tag{2}$$

where $m$ is the size of the base set.

Recently, other algorithms for approximating the *minimum set cover* were introduced, among them the randomized rounding technique of Raghavan and Thompson ([11]). In the case of covers with constant costs, Srinivasan in [12] showed that the approximation guarantee for the randomized rounding algorithm is

$$c_{rra} \leq c^*_{min} \left( \ln(\frac{m}{c^*_{min}}) + O(\ln\ln(\frac{m}{c^*_{min}})) + O(1) \right), \tag{3}$$

making it appear at that point that the performance ratio for the randomized rounding technique was better than the performance ratio for the greedy algorithm. (Here $c^*_{min} \leq c_{min}$ is the optimum cost of the LP relaxation of the *minimum cover problem.*)

The *minimum partial cover* problem further generalizes the *minimum set cover* problem. The goal here is to find a subcollection of sets covering at least a $p$-fraction, $0 < p \leq 1$, of the base set with minimum cost. This problem is also NP-hard, since it contains the *minimum set cover* as a special case ($p = 1$, $c_1 = c_2 = \cdots = c_n = 1$), and the solution can again be approximated by a greedy algorithm ([7]). Even though this algorithm (see Preliminaries) is a straightforward modification of the greedy algorithm for complete covers, its output is less predictable, hence the approach used by Chvátal, Johnson, or Lovasz for proving (1) and (2) cannot be generalized. In [7, p. 69], Kearns proves that for $0 < p \leq 1$

$$\frac{c_{greedy}}{c_{min}} \leq 2H(m) + 3.$$

This is much worse than the bounds for *minimum set cover* and only shows how complicated the behavior of the "greedy" partial cover can be.

### Our Results

In Section 2, we discuss the *minimum partial cover* problem for covers with costs. We introduce a new simple but powerful technique based on the fact that the worst possible "greedy" partial cover can be made of one-element sets. This enables us to prove that (1) is indeed true for *minimum partial cover* as well (given that the size of sets in the cover is bounded by $d$), and that

$$\frac{c_{greedy}}{c_{min}} \leq H(\lceil pm \rceil) \tag{4}$$

for general covers. Hence we establish a classical bound for a non-classical problem. These results clearly contain, as a special case, the bounds proved by Chvátal and Johnson, and significantly improve Kearns's estimate.

Section 3 is devoted to general covers with equal costs, i.e. $c_1 = c_2 = \cdots = c_n$. Lund and Yannakakis in [9] proved that even in this special case, assuming $DTIME[n^{\ln^{O(1)} n}] \neq NTIME[n^{\ln^{O(1)} n}]$, no polynomial time algorithm can approximate *minimum set cover* to a performance ratio of better than $c \log_2 m$, for any $c < 1/4$. This hardness result has renewed interest in improving the performance ratio. Recently, Halldórsson ([4]) improved the performance guarantees (1) and (2) by $11/42$ using local improvement modification of the greedy algorithm, but the bound of $H(m)$ for the classical greedy algorithm has remained the best that was known (for general covers) despite a fair amount of work by many people. In this paper, we are able to show that the greedy algorithm performs better than was previously thought; in particular we prove that the performance ratio of the greedy algorithm differs from the harmonic bound by a function that goes to infinity as the size of the base set approaches infinity. We will prove this in the setting of partial covers, and show, among other things, that, for any $0 < p \leq 1$,

$$\frac{c_{greedy}}{c_{min}} \leq \ln u - f(u), \tag{5}$$

where $u = \lceil pm \rceil$ and $f(u)$=$\Theta(\ln \ln u)$ (hence $\lim f(u) = \infty$). This improves not only our inequality (4) for the *minimum partial cover* problem but also the classical bound (2) for the *minimum (complete) cover* problem ([1],[5],[8]).

In proving the results, we introduce two new methods which should be generally useful for attacking other similar problems. For example, our approach to analyzing the greedy algorithm for covers with equal costs yields the bound

$$c_{greedy} \leq \left(c^*_{min} - \frac{1}{2}\right) \ln\left(\frac{m}{c^*_{min}}\right) + c^*_{min}, \tag{6}$$

which is clearly stronger than (3). Hence the performance guarantee for the greedy algorithm is significantly better than that for the randomized rounding technique.

## Preliminaries

**Definition 1** *Let $U$ be a finite set, $\mathbf{S} = \{S_1, S_2, \ldots, S_n\}$ be a cover of $U$ (i.e. $U = \bigcup S_j$), and $\mathbf{c} = \{c_1, c_2, \ldots, c_n\}$ be positive real costs of each set in the cover. We identify the cover $\mathbf{S}$ with the set of indices $\mathbf{J} = \{1, 2, \ldots, n\}$ and any subset $\mathbf{S}^*$ of $\mathbf{S}$ with the corresponding set $\mathbf{J}^* \subset \mathbf{J}$. Define $m = |U|$. Let $0 < p \leq 1$. We say that $\mathbf{J}^*$ (or alternatively $\mathbf{S}^*$) is a $p$-partial cover of $U$ (or simply a $p$-cover of $U$) if*

$$|\bigcup_{j \in \mathbf{J}^*} S_j| = |\bigcup_{S_j \in \mathbf{S}^*} S_j| \geq pm.$$

*We call $c(\mathbf{J}^*) = c(\mathbf{S}^*) = \sum_{j \in \mathbf{J}^*} c_j$ the cost of the $p$-cover $\mathbf{J}^*$ ($\mathbf{S}^*$).*

This definition leads immediately to the following optimization problem:

Minimum Partial Cover
**Instance:** Finite set $U$ ($|U| = m$), finite cover $\mathbf{S}$, positive costs $\mathbf{c}$, $0 < p \leq 1$.
**Question:** What is the minimum cost of a $p$-cover $\mathbf{J}^*$ of $U$?

### Greedy Algorithm

We will use a minor modification of the greedy algorithm for *minimum partial cover* given by Kearns in [7].

**Step 1.** Set $\mathbf{J}^* = \emptyset$.

**Step 2.** Set $r = pm - |\bigcup_{j \in \mathbf{J}^*} S_j|$, i.e. $r$ is the number of elements of $U$ yet to be covered in order to obtain a $p$-cover.

**Step 3.** If $r \leq 0$, then STOP and output $\mathbf{J}^*$.

Intuitively, the more elements a set has and the cheaper it is the more likely it is in a minimum $p$-cover. On the other hand, a set having more than $r$ elements is no better than a set with exactly $r$ elements. That motivates the following step.

**Step 4.** Find $i \in \mathbf{J} \setminus \mathbf{J}^*$ that minimizes the quotient $\frac{c_j}{\min(r, |S_j|)}$, for $j \in \mathbf{J} \setminus \mathbf{J}^*$ and $S_j \neq \emptyset$. In case of a tie, take the smallest such $i$.

**Step 5.** Add $i$ to $\mathbf{J}^*$. For each $j \in \mathbf{J} \setminus \mathbf{J}^*$ set $S_j = S_j \setminus S_i$. Set $U = U \setminus S_i$. Return to Step 2.

## 2 Covers with General Costs

Denote by $k$ the number of sets in a $p$-cover obtained by the greedy algorithm. We can assume without loss of generality that the sets in the cover $\mathbf{S}$ are ordered in such a way that the greedy algorithm chooses the index $i$ in its $i$-th iteration, i.e. after every iteration $i$ of the greedy algorithm, $\mathbf{J}^* = \{1, 2, \ldots, i\}$.

Let $\{j_1, j_2, \ldots, j_l\}$ be a minimum $p$-cover of $U$ with the cost $c_{min}$. To simplify the notation, denote the sets of this $p$-cover by $A_1, A_2, \ldots, A_l$ and their costs by $\{\alpha_1, \alpha_2, \ldots, \alpha_l\}$, i.e. $\{A_1, A_2, \ldots, A_l\} = \{S_{j_1}, S_{j_2}, \ldots, S_{j_l}\}$ and $c_{min} = \alpha_1 + \alpha_2 + \cdots + \alpha_l$. Also denote by $A_s^{(i)}$, $S_j^{(i)}$, and $U^{(i)}$ the sets $A_s$, $S_j$, and $U$ after the $i$-th iteration of the greedy method. Finally, denote by $d_i$ the number of elements of $U$ after the $i$-th iteration yet to be covered to obtain a $p$-cover and define $u_j^{(i)}$ to be $\min(d_i, |S_j^{(i)}|)$.

**Lemma 1** *Let $\mathbf{S} = \{S_1, \ldots, S_n\}$ be a cover of $U$, $\mathbf{A} \subset \mathbf{S}$ a $p$-cover with minimum cost, and $\mathbf{S}^*$ a $p$-cover output by the greedy algorithm. Then there is a cover $\mathbf{T} \supset \mathbf{S}$ (enrichment of the original cover) such that*

*(i) the cost of a minimum $p$-cover $\mathbf{B} \subset \mathbf{T}$ is at most the cost of $\mathbf{A}$, i.e. $c(\mathbf{B}) \leq c(\mathbf{A})$, and*

*(ii) the greedy algorithm applied on $\mathbf{T}$ outputs a $p$-cover $\mathbf{T}^*$ consisting of one-element sets that is at least as expensive as the $p$-cover $\mathbf{S}^*$, i.e. $c(\mathbf{T}^*) \geq c(\mathbf{S}^*)$.*

This Lemma (proved in the Appendix) shows that the greedy $p$-cover can be made (possibly) more expensive by replacing each set in the greedy $p$-cover by an appropriate number of singletons with appropriate costs while (possibly) lowering the cost of a minimum $p$-cover. Hence the worst possible case for the quotient $c_{greedy}/c_{min}$ can be obtained by considering only those greedy $p$-covers of $U$ consisting of singletons with appropriate costs.

**Note:** The Lemma does not claim that the worst case greedy $p$-covering cannot be achieved by bigger sets. It simply says that a greedy $p$-covering by bigger sets cannot be worse than a greedy $p$-covering by singletons with appropriate costs.

In our quest for the worst possible $p$-cover output by the greedy algorithm, Lemma 1 enables us to restrict ourselves to single-element-set greedy $p$-covers. Let $\mathbf{J}^* = \{1, \ldots .k\}$ be such a $p$-cover. Then clearly $k = \lceil pm \rceil$ and the greedy condition implies that

$$c_{i+1} \leq \frac{\alpha_s}{\min(|A_s^{(i)}|, d_i)}$$

for all $s = 1, \ldots, l$ for which $A_s^{(i)} \neq \emptyset$.

Without loss of generality, we can assume that the sets $A_s$ in the minimum $p$-cover are disjoint and that

$$\sum_{s=1}^{l} |A_s| = \lceil pm \rceil. \tag{7}$$

This can be accomplished by possibly deleting some elements of the sets $A_s$, hence (possibly) increasing the worst case cost of the greedy $p$-cover.

Consider all fractions of the form

$$\frac{\alpha_s}{k_s}, \qquad s = 1, \ldots, l, \;\; k_s = 1, \ldots, |A_s|.$$

Note that we have $k = \lceil pm \rceil$-many fractions. Let us rearrange these fractions into a nondecreasing sequence $e_1 \leq e_2 \leq \cdots \leq e_k$. Then the following inequality (proved in the Appendix) holds.

**Lemma 2** *For each $i = 1, \ldots, k$ we must have*

$$c_i \leq e_i.$$

As a straightforward consequence of Lemma 2, we easily obtain:

**Lemma 3**

$$c_1 + \cdots + c_k \leq e_1 + \cdots + e_k = \alpha_1 H(|A_1|) + \cdots + \alpha_l H(|A_l|).$$

Now we can state (and prove) the generalization of (1).

**Theorem 1** *Let $U$ be a finite set of size $m$, $\mathbf{S} = \{S_1, \ldots, S_n\}$ be a cover of $U$, $\mathbf{c} = \{c_1, \ldots, c_n\}$ be the costs of the sets in the cover, and $0 < p \leq 1$. Let $d \in \mathbf{N}$ be such that $|S_j| \leq d$ for all $j = 1, \ldots, n$. Then*

$$c_{greedy} \leq c_{min} H(d). \tag{8}$$

**Proof:** The theorem is a straightforward consequence of Lemmas 1, 2, and 3 and the fact that $H(|A_s|) \leq H(d)$ for all $s = 1, \ldots, l$. $\square$

The equality (7) implies that $|A_s| \leq \lceil pm \rceil$ for all $s = 1, \ldots, l$. This and Lemma 3 prove the following theorem.

**Theorem 2** *Let $U$, $\mathbf{S}$, $\mathbf{c}$, and $p$ be as above. Then*

$$c_{greedy} \leq c_{min} H(\lceil pm \rceil). \tag{9}$$

Theorem 2 might seem to be a weaker version of Theorem 1. Of course, this is the case when $p = 1$, since $d \leq m$. On the other hand, for $p < 1$, it might be that $d > \lceil pm \rceil$, and in this case, Theorem 2 is stronger than Theorem 1.

## 3 Covers without Cost

Let us now consider the case, where all the sets in the cover $\mathbf{S} = \{S_1, S_2, \ldots, S_n\}$ have the same cost, i.e. $c_1 = c_2 = \cdots = c_n$.

It is easy to show that even in this special case, the bound on the greedy algorithm performance (8) cannot generally be improved. On the other hand, we can obtain some improvements of (9).

We consider general covers, that is, covers where we do not know beforehand a bound on the size of the sets in the cover. For simplicity, set $u = \lceil pm \rceil$. Since we are estimating the quotient $c_{greedy}/c_{min}$, we can assume without loss of generality that

$$c_1 = c_2 = \cdots = c_n = 1. \tag{10}$$

Then, if the number of sets in a minimum cover is $l$ and the number of sets in the cover output by the greedy algorithm is $k$, we have $c_{greedy} = k$ and $c_{min} = l$.

The case $c_{min} = 1$, that is the case where $U$ can be $p$-covered by one set, is not interesting, since the greedy algorithm will also output a single set, hence $c_{greedy} = c_{min}$. Therefore in what follows, we will consider only covers for which $c_{min} = l \geq 2$.

At each step, $i$, of the greedy method, we delete $q_i$ elements. We have $k$ steps, hence

$$\sum_{i=1}^{k} q_i = u. \tag{11}$$

We know that $U$ can be $p$-covered by $l$ sets. By the pigeon hole principle, at least one of the sets in the minimum $p$-cover contains at least $\lceil \frac{u}{l} \rceil$ elements. Hence

$$q_1 \geq \lceil \frac{u}{l} \rceil.$$

Similarly,

$$q_2 \geq \lceil \frac{u - q_1}{l} \rceil$$

and, in general,

$$q_i \geq \lceil \frac{u - (q_1 + \cdots + q_{i-1})}{l} \rceil \tag{12}$$

for $i = 2, \ldots, k$. Using (11), we can rewrite (12) in the form

$$q_i \geq \lceil \frac{q_i + \cdots + q_k}{l} \rceil. \tag{13}$$

Solving for $q_i$ gives

$$q_i \geq \lceil \frac{q_{i+1} + \cdots + q_k}{l - 1} \rceil \quad \text{for } i = 1, \ldots, k. \tag{14}$$

For given $l \geq 2$, set

$$a_1 = 1$$

and

$$a_i = \lceil \frac{a_1 + \cdots + a_{i-i}}{l - 1} \rceil \tag{15}$$

for $i = 2, 3, \ldots$. Then, clearly, $a_1 \leq q_k$, $a_2 \leq q_{k-1}$, ..., $a_k \leq q_1$, hence

$$\sum_{i=1}^{k} a_i \leq \sum_{i=1}^{k} q_i. \tag{16}$$

Define

$$N(k,l) = \sum_{i=1}^{k} a_i \quad \text{for k=1,2,...} \tag{17}$$

From the discussion above, it is fairly immediate that for $k \geq l$, $N(k,l)$ represents the smallest $u$ for which there is a cover $\mathbf{S}$ of some set $U$ with $u = \lceil pm \rceil$ such that $c_{min} = l$ and $c_{greedy} = k$. Indeed, for $u < N(k,l)$ and $c_{min} = l$, inequality (16) implies that $c_{greedy} < k$, whereas if $u \geq N(k,l)$ one can easily construct a cover $\mathbf{S}$ such that $c_{min} = l$ and $c_{greedy} = k$. As far as this later point is concerned, see the following example.

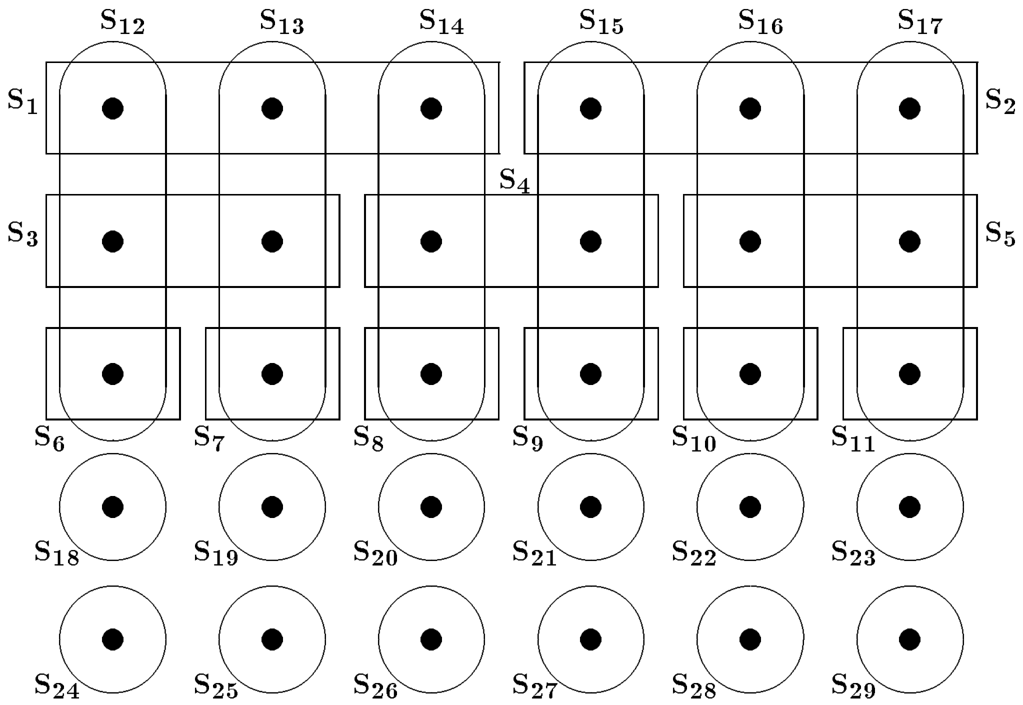


Figure 1: Example of a cover for which $c_{greedy} = k$, $c_{min} = l$, and $u = N(k,l)$.

**Example 1** Let $k \geq l$, $0 < p \leq 1$, and $u \geq N(k,l)$ be given. Set $m = \lfloor u/p \rfloor$ (hence $u = \lceil pm \rceil$) and define $U = \{1, 2, \ldots, m\}$. Since $u \geq N(k,l)$ there are positive integers $q_1, \ldots, q_k$ satisfying (14) and (11). Define a cover $\mathbf{S}$ of $U$ in the following way.

(i) For $i = 1, \ldots, k$ set

$$S_i = \{q_1 + \cdots + q_{i-1} + 1, q_1 + \cdots + q_{i-1} + 2, \ldots, q_1 + \cdots + q_i\},$$

i.e. each $S_i$ contains exactly $q_i$ elements, the sets are disjoint, and $\bigcup_{i=1}^{k} S_i = \{1, \ldots, u\}$.

(ii) Set $d = \lceil u/l \rceil$. Then one can write $u = l_1 d + l_2(d-1)$ for some $l_1$, $l_2$ such that $l_1 + l_2 = l$. Define

$$S_{k+i} = \{i, i+l, \ldots, i+l(d-1)\} \quad \text{for } i = 1, \ldots, l_1$$

and

$$S_{k+i} = \{i, i+l, \ldots, i+l(d-2)\} \quad \text{for } i = l_1 + 1, \ldots, l.$$

Then, clearly, the first $l_1$ sets contain $d$ elements each, the next $l_2$ sets contain $d-1$ elements each, the sets are disjoint, and $\bigcup_{i=1}^{l} S_{k+i} = \{1, \ldots, u\}$.

(iii) Finally, define

$$S_{k+l+i} = \{u+i\} \quad \text{for } i = 1, \ldots, m-u.$$

Figure 1 shows this construction for $m = 30$, $u = \lceil pm \rceil = 18$, $d = 3$, $l = 6$, $k = 11$.

We claim that the greedy algorithm outputs a $p$-cover $\mathbf{J}^* = \{1, \ldots, k\}$. Indeed, $q_1 \geq d$, hence in the first iteration $S_1$ is chosen. Assume that after the $i$-th step, $\mathbf{J}^* = \{1, \ldots, i\}$, leaving $q_{i+1} + \cdots + q_k$ elements to cover. Set $r = \max\{|S_{k+j}^{(i)}| : j = 1, \ldots, l\}$. Because of the construction of the sets in the cover, we have $r - 1 \leq |S_{k+j}^{(i)}| \leq r$ for all $j = 1, \ldots, l$. Hence $q_{i+1} + \cdots + q_k = \sum_j |S_{k+j}^{(i)}| > (r-1)l$ and thus $q_{i+1} \geq r$ by inequality (13). As a result, the greedy algorithm will choose the set $S_{i+1}$ in its $(i+1)$-st step. $\square$

It is now clear that, for fixed $c_{min} = l$, $u < N(k,l)$ implies $c_{greedy} < k$. On the other hand, if $u \geq N(k,l)$, there are covers for which $c_{min} = l$ and $c_{greedy} = k$. This proves the following Lemma:

**Lemma 4** *For any set $U$ and any $0 < p \leq 1$ such that $u = \lceil pm \rceil \geq 2$, and for any cover $\mathbf{S}$ of $U$,*

$$\frac{c_{greedy}}{c_{min}} \leq \max\{\frac{k}{l} \mid N(k,l) \leq u\}. \tag{18}$$

*Moreover, for any $u \geq 2$, there are covers for which the equality is attained.*

Lemma 4 establishes a tight bound on the quotient $c_{greedy}/c_{min}$. Unfortunately, it is of little practical use since we know almost nothing about the numbers $N(k,l)$. The only immediate facts we have (as a consequence of (17) and (15)) are that, for any $2 \leq l \leq k$, $N(l,l) = l$ and

$$N(k+1,l) = N(k,l) + \lceil \frac{N(k,l)}{l-1} \rceil = \lceil \frac{l}{l-1} N(k,l) \rceil. \tag{19}$$

Hence we can generate $N(k,l)$ for any $k \geq l \geq 2$. This allows us to evaluate the bound for the quotient $c_{greedy}/c_{min}$ for small $u$, but it does not say much about asymptotic behavior. Let us now establish some asymptotic properties of $N(k,l)$.

Using (19), one can show that

$$N(k,l) \geq \left(\frac{l}{l-1}\right)^{k-l} N(l,l) \geq e^{\frac{k-l}{l}} l \tag{20}$$

which easily gives

$$\ln N(k,l) \geq k/l + \ln l - 1. \tag{21}$$

The following lemma improves the above inequality. The proof uses more involved estimates and is in the Appendix.

**Lemma 5** *For any $k \geq l \geq 2$*

$$\ln N(k,l) \geq \ln l + (k-l)\frac{2}{2l-1}. \tag{22}$$

Rearranging (22) gives immediately

$$\frac{k}{l} \leq \frac{2l-1}{2l}(\ln N(k,l) - \ln l) + 1 \tag{23}$$

for any $k \geq l \geq 2$.

This and Lemma 4 prove the following.

**Proposition 1** *For any set $U$ such that $u = \lceil pm \rceil \geq 1$ and any cover* $\mathbf{S}$ *of $U$, the $p$-cover output by the greedy algorithm satisfies*

$$\frac{c_{greedy}}{c_{min}} \leq \frac{2c_{min} - 1}{2c_{min}} (\ln u - \ln c_{min}) + 1.$$

The bound of Proposition 1 is not very useful - we need a bound without specific dependence on $c_{min}$. Tedious analysis (details can be found in the Appendix) proves

**Theorem 3** *Define $M(u)$ to be* $\max\{k/l \mid N(k,l) \leq u\}$ = *"the worst case for $c_{greedy}/c_{min}$". Then there is a function $f(u) = \Theta(\ln \ln u)$ such that*

$$M(u) \leq \ln u - f(u).$$

*Thus*

$$\frac{c_{greedy}}{c_{min}} \leq \ln u - \Omega(\ln \ln u).$$

**Note:** One can show analytically that $f(u) \geq \ln \ln u - a$ for $u \geq u_o$ where $a > \ln 2$ and $u_o$ depends on $a$. Actual evaluation of $M(u)$ for $u \leq u_o$ shows that $\ln u - M(u) \geq \ln \ln u - a$ for any $a \geq a_o = 3 + \ln \ln 32 - \ln 32 \approx .78$ and $2 \leq u \leq u_o$. Therefore

$$\frac{c_{greedy}}{c_{min}} \leq M(u) \leq \ln u - \ln \ln u + a_o$$

for all $u \geq 2$.

# 4 Further Improvements

The bounds in the previous section are incomparable with those of [12]. In this section we further improve our estimates for covers with equal costs and show that the performance guarantee of the greedy algorithm is better than that of the algorithms based on randomized rounding technique.

Generalizing [8] we define a fractional $p$-cover $\mathbf{T}$ of $U$ to be a system of weights $\mathbf{T} = \{t_1, \ldots, t_n\}$ such that for at least $u = \lceil pm \rceil$ points $x \in U$ we have

$$\sum_{\{j \mid x \in S_j\}} t_j \geq 1.$$

Denote by $c^*(\mathbf{T})$ the cost of the fractional $p$-cover $\mathbf{T}$, i.e.

$$c^*(\mathbf{T}) = \sum_{j=1}^{n} t_j c_j = \sum_{j=1}^{n} t_j$$

and let

$$c^*_{min} = \min_{\mathbf{T}} c^*(\mathbf{T}).$$

This formulation is equivalent to the LP relaxation of the set cover problem considered by Srinivasan in [12]. Obviously, $c^*_{min} \leq c_{min}$.

Let us follow the steps in Section 3. Set $l^* = c^*_{min}$. A simple argument shows that $c^*_{min} = 1$ implies $c_{min} = 1$, hence by considering only those covers for which $c_{min} = l \geq 2$, we actually consider covers for which $c^*_{min} = l^* > 1$. Now,

$$q_1 l^* = q_1 \sum_{j=1}^{n} t_j \geq \sum_{j=1}^{n} |S_j| t_j = \sum_{j=1}^{n} \Big( \sum_{x \in S_j} t_j \Big) = \sum_{x \in U} \Big( \sum_{\{j \mid x \in S_j\}} t_j \Big) \geq u,$$

hence $q_1 \geq \lceil \frac{u}{l^*} \rceil$. Similarly, $q_i \geq \lceil \frac{u-(q_1+\cdots+q_{i-1})}{l^*} \rceil$.

Next, we can define $a_1^* = 1$ and

$$a_i^* = \lceil \frac{a_1^* + \cdots + a_{i-i}^*}{l^* - 1} \rceil,$$

and obtain as before $\sum_{i=1}^k a_i^* \leq \sum_{i=1}^k q_i$. Let us define

$$N^*(k, l^*) = \sum_{i=1}^{k} a_i^* \quad \text{for k=1,2,}\ldots. \tag{24}$$

From the discussion above, it is clear that $u < N^*(k, c_{min}^*)$ implies $c_{greedy} < k$ hence the following counterpart of Lemma 4 holds.

**Lemma 6** *For any set $U$ and any $0 < p \leq 1$ such that $u = \lceil pm \rceil \geq 2$, and for any cover $\mathbf{S}$ of $U$,*

$$c_{greedy} \leq \max\{k \mid N(k, c_{min}^*) \leq u\}.$$

We have again that

$$N^*(k+1, l^*) = N^*(k, l^*) + \lceil \frac{N^*(k, l^*)}{l^* - 1} \rceil = \lceil \frac{l^*}{l^* - 1} N^*(k, l^*) \rceil, \tag{25}$$

and, with a small adjustment, $N^*(\lfloor l^* \rfloor, l^*) = \lfloor l^* \rfloor$ for any $l^* > 1$. Careful analysis shows that

$$N^*(k, l^*) \geq \left( \frac{l^*}{l^* - 1} \right)^{k - \lfloor l^* \rfloor} N^*(\lfloor l^* \rfloor, l^*) \geq e^{\frac{k - l^*}{l^*}} l^* \tag{26}$$

for $k \geq l^*$, hence

$$\ln N^*(k, l^*) \geq k/l^* + \ln l^* - 1. \tag{27}$$

Thus we have shown the following.

**Proposition 2** *For any set $U$ such that $u = \lceil pm \rceil \geq 1$ and any cover $\mathbf{S}$, the p-cover output by the greedy algorithm satisfies*

$$c_{greedy} \leq c_{min}^* (\ln u - \ln c_{min}^* + 1). \tag{28}$$

Proceeding as in Section 3, we can further improve (28) and obtain the following analogy of Proposition 1. The proof is in the Appendix.

**Theorem 4** *For any set $U$ such that $u = \lceil pm \rceil \geq 1$ and any cover $\mathbf{S}$, the p-cover output by the greedy algorithm satisfies*

$$c_{greedy} \leq \left( c_{min}^* - \frac{1}{2} \right) (\ln u - \ln c_{min}^*) + c_{min}^*. \tag{29}$$

Theorem 4 shows that the performance guarantee for the greedy algorithm is substantially better than the performance guarantee (3) for the randomized rounding algorithm. Moreover, the above inequality is of the same form as the bound in Proposition 1, only $c_{min}$ is replaced by $c_{min}^*$. Thus a simple repetition of the steps in the proof of Theorem 3 proves that

$$\frac{c_{greedy}}{c_{min}^*} \leq \ln u - \Theta(\ln \ln u).$$

**Acknowledgements**

I would like to thank Eugene Kleinberg, my advisor, for many useful discussions. I am also very grateful to Jon Kleinberg for his comments, references, and suggestions. In particular, he pointed me towards establishing the result in Theorem 3.

## References

[1] V. Chvátal. A Greedy Heuristic for the Set-covering Problem. *Mathematics of Operations Research* 4(1979), pp. 233-235.

[2] P. Crescenzi and V. Kann. A Compendium of NP Optimization Problems. Technical Report SI/RR-95/02, Department of Computer Science, University of Rome "La Sapienza", 1995.

[3] M.R. Garey and D.S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness* W.H. Freeman, New York, NY (1979).

[4] M.M. Halldórsson. Approximating Discrete Collections via Local Improvements. *Proc. 6th Annual ACM-SIAM Symposium on Discrete Algorithms* (1995).

[5] D.S. Johnson. Approximation Algorithms for Combinatorial Problems. *Journal of Computer and System Sciences* 9(1974), pp. 256-278.

[6] R.M. Karp. Reducibility among Combinatorial Problems. In R.E. Miller and J.W. Thatcher, editors, *Complexity of Computer Computations*, 85-103, Plenum Press, New York, NY (1972).

[7] M.J. Kearns. *The Computational Complexity of Machine Learning.* MIT Press, Cambridge, MA (1990).

[8] L. Lovasz. On the Ratio of Optimal Integral and Fractional Covers. *Discrete Mathematics* 13(1975), pp. 383-390.

[9] C. Lund and M. Yannakakis. On the Hardness of Approximating Minimization Problems. *Journal of the ACM* 41(1994), pp. 960-982.

[10] P. Raghavan. Probabilistic Construction of Deterministic Algorithms: Approximating Packing Integer Programs. *Journal of Computer and System Sciences* 37(1988), pp. 130-143.

[11] P. Raghavan and C.D. Thompson. Randomized Rounding: a Technique for Provably Good Algorithms and Algorithmic Proofs. *Combinatorica* 7(1987), pp. 365-374.

[12] A. Srinivasan. Improved Approximation of Packing and Covering Problems. In *Proc. ACM Symposium on Theory of Computing* (1995), pp. 268-276.

# Appendix

**Proof**[Lemma 1]: In every iteration $(i+1)$ of the greedy method, we choose a subscript $j$ for which $c_j/u_j^{(i)}$ is minimum. Assume, as before, that the greedy algorithm chooses index $i$ in its $i$-th iteration. Define

$$f(U^{(i)}) = \min\left\{\frac{c_j}{u_j^{(i)}} \mid j = i+1,\ldots,n\right\} = \frac{c_{i+1}}{u_{i+1}^{(i)}}$$

for all $i = 0,\ldots,k-1$. Clearly, $f(U^{(i)}) \leq f(U^{(i+1)})$ for any $i = 0,\ldots,k-2$.

Assume that $k < \lceil pm \rceil$. Then there is at least one $i$ such that $u_{i+1}^{(i)} > 1$ (i.e. at the $(i+1)$ st iteration, we delete at least two elements). For some $x \in S_{i+1}^{(i)}$, define

$$S_{n+1} = \{x\},\ S_{n+2} = S_{i+1}^{(i)} \setminus \{x\},\ c_{n+1} = f(U^{(i)}),\ \text{and } c_{n+2} = f(U^{(i)} \setminus \{x\})(u_{i+1}^{(i)} - 1).$$

Consider the cover

$$\mathbf{R} = \{R_1,\ldots,R_{n+2}\} = \{S_1,\ldots,S_i,S_{n+1},S_{n+2},S_{i+2},\ldots,S_n,S_{i+1}\}$$

consisting of the sets of the original cover $\mathbf{S}$ and the two new sets $S_{n+1}$ and $S_{n+2}$. Clearly, enriching the cover $\mathbf{S}$ by the two sets in no way increases the minimum cost of a $p$-cover - it might only decrease it. On the other hand, because of the choice of the costs $c_{n+1}$ and $c_{n+2}$ and the fact that in the case of two sets having the same cost per element, the one with smaller subscript is chosen, the greedy algorithm applied on $\mathbf{R}$ will output a $p$-cover

$$\mathbf{R}^* = \{R_1,\ldots,R_{k+1}\} = \{S_1,\ldots,S_i,S_{n+1},S_{n+2},S_{i+2},\ldots,S_k\}.$$

Since

$$\begin{aligned} c_{n+1} + c_{n+2} &= f(U^{(i)}) + f(U^{(i)} \setminus \{x\})(u_{i+1}^{(i)} - 1) \\ &\geq f(U^{(i)}) + f(U^{(i)})(u_{i+1}^{(i)} - 1) = f(U^{(i)})u_{i+1}^{(i)} = c_{i+1} \end{aligned}$$

we have

$$c(\mathbf{R}^*) \geq c(\mathbf{S}^*).$$

Hence after adding sets $S_{n+1}$ and $S_{n+2}$ (with appropriate costs) to the original cover, and appropriate rearranging, the greedy algorithm outputs a $p$-cover of $k+1$ sets that is at least as expensive as the $p$-cover output originally.

For simplicity, set $u = \lceil pm \rceil$ and $N = n + 2(\lceil pm \rceil - k)$. Applying the above method repeatedly $(u-k)$ times, we obtain a cover $\mathbf{R} = \{R_1,\ldots,R_N\}$ such that the minimum cost of a $p$-cover for $\mathbf{R}$ does not exceed $c(\mathbf{A})$ and the greedy algorithm applied on $\mathbf{R}$ outputs a $p$-cover $\mathbf{R}^* = \{R_1,\ldots,R_u\}$ such that $c(\mathbf{R}^*) \geq c(\mathbf{S}^*)$. Also $u_{i+1}^{(i)} = 1$ for all $i = 0,\ldots,u-1$, i.e. only one element of each set $R_i$ contributes to our greedy $p$-cover. Defining

$$R_{N+i} = R_i \setminus \bigcup_{j=1}^{i-1} R_j = R_i^{(i-1)}, \quad \text{for } i = 1,\ldots,u-1$$

and

$$R_{N+u} = \{x\} \quad \text{for some } x \in R_u^{(u-1)},$$

gives us

$$|R_{N+1}| = \cdots = |R_{N+u}| = 1.$$

Set $c_{N+i} = c_i$ for all $i = 1, \ldots, u$. Then the cover

$$\mathbf{T} = \{T_1, \ldots, T_{N+u}\} = \{R_{N+1}, \ldots, R_{N+u}, R_{u+1}, \ldots, R_N, R_1, \ldots, R_{u-1}, R_u\}.$$

proves the Lemma. $\square$

**Proof**[Lemma 2]: After $i$-th iteration of the greedy algorithm, $i = 0, \ldots, k-1$, there are exactly $d_i = k - i$ elements to be covered, i.e. there are at least $k - i$ elements in $\bigcup A_s$ not deleted in the previous steps. Hence, if we define $b_s^{(i)} = \min(|A_s^{(i)}|, d_i)$, we have $\sum_{s=1}^{l} b_s^{(i)} \geq d_i = k - i$. The greedy condition implies that

$$c_{i+1} \leq \frac{\alpha_s}{b_s^{(i)}}$$

for all $s = 1, \ldots, l$ for which $b_s^{(i)} > 0$. Therefore

$$c_{i+1} \leq \frac{\alpha_s}{k_s},$$

for all $s = 1, \ldots, l$ and all $k_s = 1, \ldots, b_s^{(i)}$, i.e.

$$c_{i+1} \leq e_j$$

for at least $k - i$ indices j. But $e_1 \leq \cdots \leq e_k$, hence

$$c_{i+1} \leq e_{k-(k-i)+1} = e_{i+1}$$

for any $i = 0, \ldots, k-1$. $\square$

**Proof**[Lemma 5]: Since the function $y = 1/x$ is convex (concave up), we have

$$\ln(a+b) - \ln a > \frac{b}{a + \frac{b}{2}} = \frac{2b}{2a+b}$$

for any $a, b > 0$. Also, for any $0 \leq a < b < c$, $\frac{b}{c} \geq \frac{b-a}{c-a}$, hence

$$\ln N(k+1, l) - \ln N(k, l) > \frac{2\lceil \frac{1}{l-1} N(k,l) \rceil}{2N(k,l) + \lceil \frac{1}{l-1} N(k,l) \rceil} \geq \frac{2\frac{1}{l-1} N(k,l)}{2N(k,l) + \frac{1}{l-1} N(k,l)} = \frac{2}{2l-1}.$$

Applying the above inequality repeatedly yields

$$\ln N(k,l) \geq \ln N(l,l) + (k-l)\frac{2}{2l-1} = \ln l + (k-l)\frac{2}{2l-1}.$$

$\square$

**Proof**[Theorem 3]: Set $k = c_{greedy}$, $l = c_{min}$, $g(l,u) = \dfrac{\ln u - \ln l}{2l} + \ln l - 1$ and $f(u) = \min_{2 \leq l \leq u} g(l,u)$. By Proposition 1, we have

$$\ln u - M(u) \geq g(l,u) \geq f(u).$$

The partial derivative of $g(l,u)$ with respect to $l$ is

$$\frac{\partial g(l,u)}{\partial l} = \frac{1}{2l^2}(2l + \ln l - \ln u - 1),$$

hence, for fixed $1 \leq u \leq 2e^3$, $g(l,u)$ is an increasing function of $l$, and for fixed $u > 2e^3$, $g(l,u)$ is a unimodal function with both relative and absolute minimum at $l = \hat{l}$, where $\hat{l}$ satisfies $\ln u = h(\hat{l})$

with

$$\ln u = h(\hat{l}) = \ln \hat{l} + 2\hat{l} - 1. \tag{30}$$

Therefore

$$f(u) = \begin{cases} g(2,u) = \frac{\ln u}{4} + \frac{3\ln 2}{4} - 1 & \text{for } 1 \le u \le 2e^3, \\ g(\hat{l},u) = \ln \hat{l} - \frac{1}{2\hat{l}} = \ln h^{-1}(\ln u) - \frac{1}{2h^{-1}(\ln u)} & \text{for } u > 2e^3. \end{cases}$$

Here $h^{-1}$ is the inverse of $h$. Since $h$ is increasing, so is $h^{-1}$, hence $f$ is also increasing. Clearly, $\lim_{u\to\infty} f(u) = \infty$. Finally, $\ln u = \Theta(\hat{l})$ hence $\hat{l} = \Theta(\ln u)$, therefore $f(u) = \Theta(\ln \hat{l}) = \Theta(\ln\ln u)$. Hence $M(u) \le \ln u - \Theta(\ln\ln u)$. (Detailed analysis shows that $0.4\ln\ln u \le f(u) \le \ln\ln u$ for $u \ge e^{2e}$.) $\square$

**Proof**[Theorem 4]**:** In order to simplify the notation, let us omit the "*" when referring to $N^*$ and $l^*$. Hence $l > 1$ is now a rational number. As before

$$\ln N(k+1,l) - \ln N(k,l) > \frac{2}{2l-1},$$

hence

$$\ln N(k,l) \ge \ln\lfloor l\rfloor + \frac{2(k-\lfloor l\rfloor)}{2l-1} = \ln l + \frac{2(k-l)}{2l-1} + \omega,$$

where

$$\omega = \frac{2(l-\lfloor l\rfloor)}{2l-1} + \ln\lfloor l\rfloor - \ln l.$$

Set $\alpha = l - \lfloor l\rfloor$. Then

$$\omega = \frac{2\alpha}{2l-1} + \ln(l-\alpha) - logl \ge \frac{2\alpha}{2l-1} - \frac{\alpha}{l-\alpha} = \frac{\alpha(1-2\alpha)}{(2l-1)(l-\alpha)}.$$

Thus for $0 \le \alpha \le 1/2$ and any $k \ge l > 1$, we have $\omega \ge 0$, hence

$$\ln N(k,l) \ge \ln l + \frac{2(k-l)}{2l-1}. \tag{31}$$

If $\alpha \ge 1/2$, then $l \ge 3/2$, hence $N(\lceil l\rceil, l) = \lfloor l\rfloor + 2$. Therefore

$$\ln N(k,l) \ge \ln(\lfloor l\rfloor + 2) + \frac{2(k-\lfloor l\rfloor-1)}{2l-1} = \ln l + \frac{2(k-l)}{2l-1} + \epsilon,$$

where

$$\epsilon = \frac{2(l-\lfloor l\rfloor-1)}{2l-1} + \ln(\lfloor l\rfloor+2) - \ln l.$$

Similarly as above,

$$\epsilon \ge \frac{2\alpha-2}{2l-1} + \frac{2-\alpha}{l+1} = \frac{2l+3\alpha-4}{(2l-1)(l+1)} \ge 0,$$

hence (31) is valid for $\alpha \ge 1/2$ as well. Rearranging (31) completes the proof. $\square$