Algorithm 1 Greedy Weighted Set Cover

```
GreedySetCover(U, S_1, ..., S_m, w_1, ..., w_m)
  R ← U
  S ← ∅
  while R ≠ ∅ do
    select S' = arg min_{S_i} w_i / |S_i ∩ R|
    S ← S + S'
    R ← R − S'
  end do
  return S
```

Greedy Approximation

In the previous section, we applied a greedy approach to come up with an approximate solution to TSP. Unfortuantely, we were not able to provide any firm bounds on the quality of the approximation. In this section, we consider a greedy approximation for another NP-complete problem—the Set Cover problem—and provide a more reasonable bound.

Recall the definition of the Set Cover problem:

Definition: Let U be a set of n elements, and let $S_1, \ldots, S_m$ be a list of m subsets of U. Then $\mathcal{S} \subseteq \{S_1, \ldots, S_m\}$ is a *set cover* if the union of all the subsets in $\mathcal{S}$ equals U.

Definition: The *weighted minimal set cover problem* associates a weight w_i with each subset S_i. The solution to the problem is a set cover $\mathcal{S}^*$ of minimum weight $\sum_{S_i \in \mathcal{S}} w_i$.

The greedy approximation algorithm is very similar to the greedy approximation algorithm for TSP. Note that the function to be minimized could be based on the weights w_i, except we would also like to have the selected sets cover as many elements as possible. Therefore, a reasonable function combines these by considering the average weight of the elements not currently covered in by the set cover, $w_i/|S_i|$.

We now analyze the performance of this greedy algorithm. We first not that if $\mathcal{S}$ is the set cover obtained by GreedySetCover, then $\sum_{S_i \in \mathcal{S}} w_i = \sum_{s \in U} c_s$ where $c_s = w_i/|S_i \cap R|$. Thus we consider how much total cost any single subset S_k can account for. We know that the optimum solution must cover the cost of each of the members if U.

Recall during the "analysis" portion of this course, we discussed the Harmonic function

$$H(n) = \sum_{i=1}^{n} \frac{1}{i}$$

and showed how to find an asymptotic bound on this function by approximating with integrals. Specifically, we showed that

$$H(n) \leq 1 + \int_1^n \frac{1}{x} dx = 1 + \ln x$$

and

$$H(n) \geq \int_1^{n+1} \frac{1}{x} dx = \ln(n+1).$$

Thus we were able to show that $H(n) = \Theta(\ln n)$. We will use this fact in establishing our approximation ratio. First\, we prove the following lemma.

Lemma: For every set S_k, the sum $\sum_{s \in S_k} c_s$ is at most $H(|S_k|) \cdot w_k$.

Proof: Without loss of generality, assume the elements in S_k correspond to the first $d = |S_k|$ elements in U, i.e., $S_k = \{s_1, \ldots, s_d\}$. We will also assume the elements are labeled in the order in which they are assigned cost c_{s_j} by the greedy algorithm. Now consider the iteration where element s_j is covered

by some subset selected by the greedy algorithm for some $j \leq d$. Thus at the start of the iteration, $s_j, s_{j+1}, \ldots, s_d \in R$ because of the way we labeled the elements. Thus $|S_j \cap R| \geq d - j + 1$. So the average cost of the set S_k is at most

$$\frac{w_k}{|S_k \cap R|} \leq \frac{w_k}{d - j + 1}.$$

Suppose we select some subset S_i using the greedy algorithm. Recall that it is the average cost that gets assigned to the elements that are not covered, so we have

$$c_{s_j} = \frac{w_i}{|S_i \cap R|} \leq \frac{w_k}{|S_k \cap R|} \leq \frac{w_k}{d - j + 1}.$$

Thus,

$$\sum_{s \in S_k} c_s = \sum_{j=1}^{d} c_{s_j} \leq \sum_{j=1}^{d} \frac{w_k}{d - j + 1} = \frac{w_k}{d} + \frac{w_k}{d-1} + \cdots + \frac{w_k}{1} = H(d) \cdot w_k.$$

This prepares us to prove the main result. Here, let $d^* = \max_i |S_i|$ denote the maximum size of any set. Then we have the following:

Theorem: The set cover $\mathcal{S}$ selected by GreedySetCover has weight at most $H(d^*)$ times the optimal weight w^*.

Proof: Let $\mathcal{S}^*$ denote the optimum set cover. Then $w^* = \sum_{S_i \in \mathcal{S}} w_i$. For each $S_i \in \mathcal{S}$, the above Lemma implies

$$w_i \geq \frac{1}{H(d^*)} \sum_{s \in S_i} c_s.$$

Since we have a set cover, we also have

$$\sum_{S_i \in \mathcal{S}^*} \sum_{s \in S_i} c_s \geq \sum_{s \in U} c_s.$$

Previously, we noted that $\sum_{S_i \in \mathcal{S}} w_i = \sum_{s \in U} c_s$, so we can then conclude

$$w^* = \sum_{S_i \in \mathcal{S}^*} w_i \geq \sum_{S_i \in \mathcal{S}^*} \frac{1}{H(d^*)} \sum_{s \in S_i} c_s \geq \frac{1}{H(d^*)} \sum_{s \in U} c_s \frac{1}{H(d^*)} \sum_{S_i \in \mathcal{S}} w_i.$$

Consequently, GreedySetCover returns a solution within a factor of $O(\log d^*)$ of optimal.