A Note on Set Cover Inapproximability Independent of Universe Size

Jelani Nelson*
MIT CSAIL
minilek@mit.edu

September 21, 2007

Abstract

In the set cover problem we are given a collection of \( m \) sets whose union covers \( \{1, \ldots, n\} \) and must find a minimum-sized subcollection whose union still covers \( \{n\} \). We investigate the approximability of set cover by an approximation ratio that depends only on \( m \) and observe that, for any constant \( c < 1/2 \), set cover cannot be approximated to within \( O(2^{\log^{1-1/(\log \log m)} m}) \) unless SAT can be decided in slightly subexponential time. We conjecture that polynomial time \( m^{1-\epsilon} \)-approximation is impossible for any \( \epsilon > 0 \) unless SAT can be decided in subexponential time.

1 Introduction

Set cover is one of the oldest known NP-complete problems, being listed as one of Karp’s “original 21 NP-complete problems” [6]. In the set cover problem we are given a collection of \( m \) sets whose union covers \( \{n\} = \{1, \ldots, n\} \) and must find a subcollection of minimum size still covering \( \{n\} \). Its approximability in terms of \( n \) is well-understood. It is known that a simple greedy algorithm [5, 12] and a randomized rounding scheme of an LP relaxation [7, 13] both achieve an approximation ratio of \((1 - o(1)) \ln n\). On the negative side, it is known that set cover cannot be approximated to within \( c \ln n \) for some constant \( c < 1 \) unless \( P = NP \) [11], or to within better than \((1 - o(1)) \ln n \) unless \( NP \subseteq \text{TIME}(n^{O(\log \log n)}) \) [4].

This work makes a first attempt at understanding the (in)approximability of set cover to within an approximation ratio that only depends on \( m \), independent of \( n \). The only known result concerning approximability of set cover in terms of \( m \) that the author could find is a result of [2], which gives a non-trivial improvement over \( O(m) \)-approximation only when the VC-dimension of the set system is sufficiently smaller than \( \lg m - \lg \lg m \).\(^1\) The original hardness proof for set cover by Lund and Yannakakis [1, 8] gives a reduction where \( n \) and \( m \) are polynomially related, as do subsequent proofs. Such results thus imply \( \Omega(m) \)-hardness of approximation for set cover. The question then becomes whether polynomial-time \( O(m) \)-approximation is possible regardless of the relationship between \( m \) and \( n \). An instance with \( m = O(n) \) can be solved exactly in polynomial time by brute force since set cover can be solved in time \( O(n \cdot 2^{O(m)}) \), but what about \( m \) slightly superlogarithmic? By combining the proof of Lund and Yannakakis [8] with a result of Dinur and Safra [3], we observe that set cover cannot be approximated to within \( 2^{\log^{1-\delta_c(m)} m} \) in polynomial time for any constant \( c < 1/2 \) unless SAT can be decided in time \( 2^{O(2^{\log^{1-\delta_c(m)} n})} \), where \( \delta_c(n) = 1/(\log \log n)^c \).

\(^*\)Supported by an NDSEG fellowship.

\(^1\)The abstract of [2] has a slight bug; the VC-dimension “d” they mention actually refers to the VC-dimension of the dual set system.
2 A Motivating Example

Here we give an example of why one may want an approximation algorithm for set cover whose approximation ratio depends only on \( m \). Consider the following taxonomy labeling problem introduced by Rabani, Schulman, and Swamy \([10]\). In this problem there is some finite alphabet \( \Sigma \) and a tree with \( n \) nodes. Each node is labeled with a string in \( (\Sigma \cup \{0, 1\})^m \), and we must replace all occurrences of \( \ast^n \) in all labels with elements of \( \Sigma \) so as to minimize the maximum Hamming distance of labels of adjacent nodes. The authors show a reduction from taxonomy labeling to what they call the multicut packing problem on trees. In the multicut packing problem on trees we are given a tree on \( n \) nodes and a set of \( m \) multicut instances \( M_i = \{(s_1, t_1), \ldots, (s_{m_i}, t_{m_i})\} \). For each \( i \) we must output a multicut, i.e. a set of edges whose removal disconnects \( s_i \) from \( t_i \) for all \( i \), and the objective is to minimize the maximum number of times any edge is used in a multicut. Rabani, Schulman, and Swamy obtain a \( O(\log^2 m) \)-approximation for multicut packing on trees, independent of \( n \).

Now one may observe that multicut packing on trees is actually a special case of the following generalization of set cover. We are given \( N \) collections \( C_i = \{S_1^i, \ldots, S_{m_i}^i\} \) of subsets of \([n]\). We must choose a subcollection \( C_i' \) from each \( C_i \) so that \( \bigcup_{i \in I} S = [n] \), and the objective is to minimize \( \max_i |C_i'| \). In the case of multicut packing on trees, for each edge \( e \) we have a collection \( C_e \). The universe to be covered consists of all commodities in all multicut instances. Each \( m_e \) equals \( m \), the number of multicut instances, and \( S_e^i \) is the set of commodities \((s, t) \in M_i\) such that \( e \) is on the unique path from \( s \) to \( t \), i.e. the removal of edge \( e \) cuts \((s, t)\) (recall the graph is a tree). As the approximation ratio for multicut packing on trees obtained in \([10]\) is \( O(\log^2 m) \), one may wonder whether such a result could be extended to all instances of this generalized set cover problem. Our observation implies that such a result is impossible unless SAT has subexponential time algorithms since \( O(2^{\log^{1-\epsilon} n}) \)-approximation is hard even when \( N = 1 \) (the usual set cover problem).

3 The Main Observation

\textbf{Definition 1.} LabelCover\((c, s)\) is the promise problem where we are given a bipartite graph that is both left-regular and right-regular with bipartition \( V = V_1 \cup V_2 \ (|V| = n) \), edge set \( E \), label set \( [L] = \{1, \ldots, L\} \), and a set of functions \( f_e : [L] \rightarrow [L] \) indexed by edge (there is exactly one such function per edge in \( E \)). A labeling is a function \( \ell : V \rightarrow [L] \), and an edge \( e = (v_1, v_2) \) is said to be satisfied by \( \ell \) if \( \ell(v_2) = f_e(\ell(v_1)) \). In the promise problem we are given an instance where either a labeling exists satisfying at least \( c|E| \) edges, or no labeling satisfies more than \( s|E| \) edges. We must decide which case holds.

\textbf{Theorem 2} \([3]\). For any constant \( c < 1/2 \) deciding LabelCover\((1, 2^{-\log^{1-\epsilon} n})\) with a polynomial-size alphabet is \( NP \)-hard, where \( \delta_c(n) = 1/(\log \log n)^c \).

The work of \([3]\) actually defines LabelCover differently. In their definition one must assign a \( c \)-set of labels \( \ell(v) \) to each vertex \( v \in V \) so as to satisfy all edges. In this scenario an edge \( e = (v_1, v_2) \), where \( v_1 \in V_1 \), is said to be satisfied when for each label \( \ell_2 \in \ell(v_2) \) there is a label \( \ell_1 \in \ell(v_1) \) such that \( f_e(\ell_1) = \ell_2 \). The goal is then to minimize the \( l_p \) norm of the vector \( \langle |\ell(v)| \rangle_{v \in V} \). The work of \([3]\) shows that polynomial-time \( 2^{\log^{1-\epsilon} n} \)-approximation is \( NP \)-hard for any \( 1 \leq p \leq \infty \), which implies Theorem 2 by using a known relationship \([1]\) between the version of LabelCover defined in \([3]\) with \( p = 1 \) to the version of LabelCover in Definition 1.

\textbf{Theorem 3} \([1, 8]\). Suppose it is \( NP \)-hard to decide LabelCover\((1, \epsilon)\) with a label set of size \( L = O(f(n)) \). Then for any \( \ell \) such that \( 2/\ell^2 < \epsilon \), set cover has no polynomial-time \( O(\ell) \)-approximation unless \( SAT \) can be decided in time \( (n f(n)2^f)^O(1) \).

\textbf{Proof.} For any \( \ell \) satisfying \( 2/\ell^2 < \epsilon \), the work of \([1, 8]\) reduces a LabelCover\((1, \epsilon)\) instance \( \mathcal{I} \) with \( n \) vertices, \( m \) edges, and label size \( L \) to a set cover instance \( \mathcal{S} \) with universe size \( O(mL2^{2\ell}) \) and collection size \( nL \) such that approximating \( \mathcal{S} \) to within \( O(\ell) \) in time polynomial in \( |\mathcal{S}| \) allows one to decide \( \mathcal{I} \) in time polynomial in \( |\mathcal{S}| \). Furthermore, the reduction takes time polynomial in \( |\mathcal{S}| \). \( \Box \)
Corollary 4. Set cover has no polynomial-time $2^{\log^{1-\delta_c(m)} m}$-approximation unless SAT can be decided in time $2^{O(2^{\log^{1-\delta_c(n)} n})}$.

Proof. Combine Theorems 2 and 3 with $\ell = 2^{-\log^{1-\delta_c(m)} m}$.

4 Conclusion

In the above discussion, to get better hardness for set cover in terms of $m$ we showed $\Omega(\log n)$-hardness of approximation for set cover while decreasing $m$ as a function of $n$. While this may be the right approach for reaching the limits of hardness in terms of $m$, such an approach may not be necessary to get more immediate improvements. For example, one may imagine being able to show $\Omega(\sqrt[4]{\log n})$-hardness for set cover with a reduction where $m = O(\log^2 n)$, which would imply $\Omega(m^{1/4})$-hardness of approximation.

We conclude with the following conjecture:

Conjecture 5. Set cover cannot be approximated to within $m^{1-\epsilon}$ in polynomial time for any constant $\epsilon > 0$ unless SAT has subexponential time algorithms.

Note that this conjecture cannot be proven by reduction from LabelCover since there is a constant $c > 0$ such that LabelCover$(1, f(n))$ is not NP-hard for any $f(n) \leq c/\sqrt{n}$: David Peleg gives a randomized $O(\sqrt{n})$ approximation algorithm for LabelCover in [9] which he states can be derandomized. Approaching $m^{1-\epsilon}$-hardness would require a different kind of reduction.

Acknowledgments

I thank Erik Demaine for bringing [2] to my attention, Venkatesan Guruswami for bringing [1] to my attention, and Hervé Brönnimann, Swastik Kopparty, Yuval Rabani, Prasad Raghavendra, Madhu Sudan, and David Woodruff for useful comments and discussion.

References


