On the Positive-Negative Partial Set Cover Problem

Pauli Miettinen

Helsinki Institute for Information Technology, Department of Computer Science, PO Box 68 (Gustaf Hällströmin katu 2b), FI-00014 University of Helsinki, Finland

Abstract

The Positive-Negative Partial Set Cover problem is introduced and its complexity, especially the hardness-of-approximation, is studied. The problem generalizes the Set Cover problem, and it naturally arises in certain data mining applications.

Key words: Approximation algorithms, Combinatorial problems, Hardness of approximation, Set cover

1 Introduction

The Positive-Negative Partial Set Cover (±PSC) problem is a generalization of the Red-Blue Set Cover (RBSC) problem [2], which, for one, is a generalization of the classical Set Cover (SC) problem. The RBSC problem is, however, much harder than SC admitting the *strong inapproximability* property [6]. In this paper we will prove the strong inapproximability of ±PSC. The reductions

Email address: pamietti@cs.helsinki.fi (Pauli Miettinen).

used will also lead to an approximation algorithm for ±PSC, and to results about its parameterized complexity.

1.1 Notation and Problem Definitions

In RBSC, we are given disjoint sets R and B of red and blue elements, respectively, and a collection $\mathcal{S} = \{S_1, \ldots, S_n\} \subseteq 2^{R \cup B}$. The goal is to find a collection $\mathcal{C} \subseteq \mathcal{S}$ that covers all blue elements, i.e., $B \subseteq \cup\mathcal{C}$, while minimizing the number of covered red elements. The cost of a solution $\mathcal{C}$ is defined as $\mathsf{cost}_{\text{RBSC}}(R, \mathcal{C}) = |R \cap (\cup\mathcal{C})|$, where $\cup\mathcal{C}$ is the union over $\mathcal{C}$'s sets (i.e., $\cup\mathcal{C} = \bigcup_{C \in \mathcal{C}} C$); a shorthand we use throughout this paper. We will use $\mathsf{cost}_{\text{RBSC}}(\mathcal{C})$ when R is clear from the context. Finally, let $\rho = |R|$ and $\beta = |B|$.

In ±PSC, the requirement of covering all blue elements is relaxed; instead, the goal is to find the best balance between covering the blue elements and not covering the red ones. In the context of ±PSC, the red and blue elements are called *negative* and *positive* elements, respectively.

An instance of ±PSC is a triplet $(N, P, \mathcal{Q})$ with $|N| = \nu$, $|P| = \pi$, and $\mathcal{Q} = \{Q_1, \ldots, Q_m\} \subseteq 2^{P \cup N}$. A solution of a ±PSC instance is a collection $\mathcal{D} \subseteq \mathcal{Q}$, and its cost is defined to be

$$\mathsf{cost}_{\pm\text{PSC}}(N, P, \mathcal{D}) = \Big|P \setminus \Big(\cup\,\mathcal{D}\Big)\Big| + \Big|N \cap \Big(\cup\,\mathcal{D}\Big)\Big|, \tag{1}$$

namely the number of uncovered positive elements plus the number of covered negative elements. Again we will omit N and P when they are clear from the context.

1.2 Related Work

The RBSC problem was presented by Carr et al. [2] who also gave two hardness-of-approximation results to it: (*i*) unless NP $\subseteq$ DTIME($n^{\mathrm{polylog}(n)}$), there exist no polynomial-time approximation algorithms to approximate RBSC to within a factor of $2^{(4\log n)^{1-\varepsilon}}$ for any $\varepsilon > 0$, and (*ii*) there are no polynomial-time approximation algorithms to approximate RBSC to within $2^{\log^{1-(\log\log\beta)^{-c}}\beta}$ for any constant $c < 1/2$ unless P $=$ NP. The first result was independently proved by Elkin and Peleg [4], and the latter result was based upon a result by Dinur and Safra [3]. The best upper bound for RBSC is due to Peleg [6], who recently presented a $2\sqrt{n\log\beta}$-approximation algorithm for it.

To the best of the author's knowledge, there are no previous hardness results for the ±PSC problem, nor any approximation algorithms for it. The problem itself appears in some data mining applications (e.g., [1]), but its complexity and the existence of efficient approximation algorithms for it have not been studied previously.

2 Results

The main result of this paper relates the upper and lower bounds for the ±PSC's approximability to the respective bounds for RBSC.

Theorem 1 RBSC *is approximable to within a factor of* $f(\rho, \beta, n)$ *if* ±PSC *is approximable to within a factor of* $f(\rho, \beta/\rho_{\max}, n)$*, where* $\rho_{\max}$ *is the maximum number of red elements in any set of the* RBSC *instance. Vice versa,* ±PSC *is approximable to within a factor of* $g(\nu + \pi, \pi, m + \pi)$ *if* RBSC *can be* *approximated to within a factor of* $g(\nu, \pi, m)$.

Theorem 1 and the results from Section 1.2 provide the following corollaries.

Corollary 2 *For any* $\varepsilon > 0$, *(i) there exists no polynomial-time approximation algorithm for* $\pm$PSC *with an approximation factor of* $\Omega(2^{\log^{1-\varepsilon} m^4})$ *unless* $\mathrm{NP} \subseteq \mathrm{DTIME}(n^{\mathrm{polylog}(n)})$, *and (ii) there exists no polynomial-time approximation algorithm for* $\pm$PSC *with an approximation factor of* $\Omega(2^{\log^{1-\varepsilon} \pi})$ *unless* $\mathrm{P} = \mathrm{NP}$.

Corollary 3 *There exists a polynomial-time approximation algorithm for* $\pm$PSC *that achieves an approximation factor of* $2\sqrt{(m+\pi)\log \pi}$.

The first part of Corollary 2 follows from the result by Carr et al. [2], and Corollary 3 follows from Peleg's algorithm [6]. The second part of Corollary 2 follows from a result by Dinur and Safra [3] applied to RBSC: there exists an instance of RBSC where $\rho_{\max} = O\left(2^{\log^{1-(\log\log\beta)^{-c'}}\beta}(\log\log\beta)^{c'}\right)$ for some constant $c' < 1/2$, and unless $\mathrm{P} = \mathrm{NP}$ there are no polynomial-time approximation algorithms for it with an approximation factor of $2^{\log^{1-(\log\log\beta)^{-c}}\beta}$ for any constant $c < 1/2$. Thus, if we let $g_c(x) = 2^{\log^{1-(\log\log x)^{-c}}x}$ for all $c < 1/2$, then assuming that $\mathrm{P} \neq \mathrm{NP}$, there exists no polynomial-time approximation algorithm to $\pm$PSC achieving an approximation factor of $g_c\left(\frac{\pi}{O(g_{c'}(\pi)(\log\log\pi)^{c'})}\right)$, which is $\Omega(2^{\log^{1-\varepsilon}\pi})$ for all $\varepsilon > 0$.

Theorem 1 is proved in the following two subsections, while Section 2.3 studies the parameterized complexity of $\pm$PSC. Notice that both RBSC and $\pm$PSC have instances that have an optimal solution with zero cost. However, there are trivial polynomial-time algorithms to identify such instances and to find their optimal solutions. It is thus to be understood that henceforth all instances are such that the cost of their optimal solution is at least 1.

2.1 *From* RBSC *to* ±PSC

Consider an instance of RBSC, i.e., a triplet $(R, B, \mathcal{S})$. We map this instance to an instance of ±PSC. Let the negative elements be exactly the red elements, $N = R$. For each blue element b_i, create $\rho_{\max} = \max_{S \in \mathcal{S}} |R \cap S|$ positive elements in P. Create the set collection $\mathcal{Q}$ so that all negative elements belong to the same subsets Q_j as their corresponding red elements, and all positive elements corresponding to a blue element b_i belong to the same subsets as b_i.

Let $\mathcal{D}$ be a solution of this instance of ±PSC. If $\mathcal{D}$ covers all positive elements, then the same subsets also cover all blue elements in RBSC, and $\mathcal{D}$ is a feasible solution of $(R, B, \mathcal{S})$. Moreover, $\mathsf{cost}_{\pm\text{PSC}}(\mathcal{D}) = \mathsf{cost}_{\text{RBSC}}(\mathcal{D})$, i.e., $\mathcal{D}$ induces same costs in both problems. If, on the other hand, there exists a positive element p not covered by $\mathcal{D}$, then there must be at least $\rho_{\max}$ positive elements not covered by $\mathcal{D}$. Thus we can add any set S with $p \in S$ to $\mathcal{D}$ without increasing the cost of the solution, as we cannot cover more than $\rho_{\max}$ negative elements with any S. If $\mathcal{C}$ is the (possibly extended) solution to RBSC induced by $\mathcal{D}$, we see that $\mathsf{cost}_{\pm\text{PSC}}(\mathcal{D}) \geq \mathsf{cost}_{\text{RBSC}}(\mathcal{C})$.

Finally, it is clear that the optimal solution of a ±PSC instance will cover exactly the negative elements corresponding to the red elements covered by the optimal solution of RBSC, i.e., the costs of the optimal solutions are equal. Denoting the optimal solutions to the instances of ±PSC and RBSC by $\mathcal{D}^*$ and $\mathcal{C}^*$, respectively, we see that $\frac{\mathsf{cost}_{\pm\text{PSC}}(\mathcal{D})}{\mathsf{cost}_{\pm\text{PSC}}(\mathcal{D}^*)} \geq \frac{\mathsf{cost}_{\text{RBSC}}(\mathcal{C})}{\mathsf{cost}_{\text{RBSC}}(\mathcal{C}^*)}$, and thus if we can approximate ±PSC to within a factor of $f(\rho, \beta/\rho_{\max}, n)$, then we can approximate RBSC to within a factor of $f(\rho, \beta, n)$. This concludes the proof of the first part of Theorem 1.

2.2 From ±PSC *to* RBSC

Consider an instance of ±PSC: $(N, P, \mathcal{Q})$. For each $n_i \in N$, let there be a red element $r_i^- \in R$, and for each $p_i \in P$, let there be a blue element $b_i \in B$ and a red element $r_i^+ \in R$. For each set $Q_j \in \mathcal{Q}$, let there be a set $S_j^+ \in \mathcal{S}$ and for each positive element $p_i \in P$, let there be a set $S_i^- \in \mathcal{S}$. Define these sets as

$$S_j^+ = \{r_k^- \mid n_k \in Q_j\} \cup \{b_k \mid p_k \in Q_j\} \quad \text{and}$$
$$S_i^- = \{r_i^+, b_i\}.$$

Let $\mathcal{C}$ be a solution of the thus created RBSC instance. Create $\mathcal{D}$, a solution of the ±PSC instance, by adding each Q_j to $\mathcal{D}$ if the corresponding set S_j^+ is in $\mathcal{C}$.

To show that this reduction preserves the approximability, we start by considering the cost induced by $\mathcal{D}$. First, let n_k be a negative element in $\cup\mathcal{D}$. That is, there is a set Q_j so that $n_k \in Q_j$ and $Q_j \in \mathcal{D}$. But this means that the corresponding set S_j^+ must be in $\mathcal{C}$, and therefore the red element r_k^- corresponding to n_k is in $\cup\mathcal{C}$.

Second, let p_k be a positive element that is not in $\cup\mathcal{D}$, so none of the sets Q_j that contain p_k are in $\mathcal{D}$. This means that none of the sets S_j^+ that contain b_k are in $\mathcal{C}$. But as b_k must be covered by $\mathcal{C}$, it must be that S_k^- is in $\mathcal{C}$, and so r_k^+ is in $\cup\mathcal{C}$. Hence $\mathsf{cost}_{\pm\text{PSC}}(\mathcal{D}) \leq \mathsf{cost}_{\text{RBSC}}(\mathcal{C})$.

Consider then $\mathcal{D}^*$, the optimal solution of $(N, P, \mathcal{Q})$. We show that the cost of the optimal solution of the RBSC instance created from the ±PSC instance is at most that of $\mathcal{D}^*$. Create $\mathcal{C}$ so that S_j^+ is in $\mathcal{C}$ if $Q_j \in \mathcal{D}^*$. For all blue elements b_i not yet covered by $\mathcal{C}$, add S_i^- in $\mathcal{C}$. It is straightforward to see that $\mathsf{cost}_{\pm\text{PSC}}(\mathcal{D}^*) = \mathsf{cost}_{\text{RBSC}}(\mathcal{C}) \geq \mathsf{cost}_{\text{RBSC}}(\mathcal{C}^*)$. Therefore, $\frac{\mathsf{cost}_{\text{RBSC}}(\mathcal{C})}{\mathsf{cost}_{\text{RBSC}}(\mathcal{C}^*)} \geq \frac{\mathsf{cost}_{\pm\text{PSC}}(\mathcal{D})}{\mathsf{cost}_{\pm\text{PSC}}(\mathcal{D}^*)}$, so that if we can approximate RBSC to within a factor of $g(\nu, \pi, m)$, then we can approximate ±PSC to within a factor of $g(\nu + \pi, \pi, m + \pi)$.

2.3 Parameterized Complexity

We denote the parameterized versions of ±PSC and RBSC by p-±PSC and p-RBSC. The parameter for both problems is the cost of the solution. The p-RBSC problem is W[2]-hard due to the results in [2] and [5].

In the reduction from RBSC to ±PSC (Section 2.1) the costs of the optimal solutions are equal, and in the reduction from ±PSC to RBSC (Section 2.2) the cost of the optimal solution to RBSC is at most the cost of the optimal solution to ±PSC. This proves that both reductions are indeed fpt-reductions [5], and gives rise to the following proposition.

Proposition 4 *The p-±PSC problem is equivalent to the p-RBSC problem under the fpt-reductions; especially, the p-±PSC problem is* W[2]*-hard.*

3 Conclusions

This paper studied the ±PSC problem, proving both upper and lower bounds for its approximability. In addition to being important results as such, these bounds also provide new insights into the hardness of certain data mining problems. Bounding the approximability of ±PSC (and RBSC) in terms of ν (and ρ) remains an open problem.

Acknowledgements

The author thanks Heikki Mannila and Niina Haiminen for their comments.

References

[1] F. Afrati, A. Gionis, H. Mannila, Approximating a collection of frequent sets, in: Proc. 10th KDD, 2004, pp. 12–19.

[2] R. D. Carr, S. Doddi, G. Konjevod, M. Marathe, On the red-blue set cover problem, in: Proc. 11th ACM-SIAM SODA, 2000, pp. 345–353.

[3] I. Dinur, S. Safra, On the hardness of approximating label-cover, Inf. Process. Lett. 89 (2004) 247–254.

[4] M. Elkin, D. Peleg, The hardness of approximating spanner problems, in: Proc. 17th STACS, vol. 1770 of LNCS, 2000, pp. 370–381.

[5] J. Flum, M. Grohe, Parameterized Complexity Theory, an EATCS Series, Springer-Verlag, 2006.

[6] D. Peleg, Approximation algorithms for the label-cover$_{\mathrm{MAX}}$ and red-blue set cover problems, J. Discrete Algorithms 5 (2007) 55–64.