A better approximation ratio for the Vertex Cover problem

George Karakostas
Dept. of Computing and Software
McMaster University

September 27, 2004

Abstract

We reduce the approximation factor for Vertex Cover to $2-\Theta(\frac{1}{\sqrt{\log n}})$ (instead of the previous $2-\Theta(\frac{\log\log n}{\log n})$, obtained by Bar-Yehuda and Even [2], and by Monien and Speckenmeyer [10]). The improvement of the vanishing factor comes as an application of the recent results of Arora, Rao, and Vazirani [1] that improved the approximation factor of the sparsest cut and balanced cut problems. In particular, we use the existence of two big and well-separated sets of nodes in the solution of the semidefinite relaxation for balanced cut, proven in [1]. We observe that a solution of the semidefinite relaxation for vertex cover, when strengthened with the triangle inequalities, can be transformed into a solution of a balanced cut problem, and therefore the existence of big well-separated sets in the sense of [1] translates into the existence of a big independent set.

1 Introduction

One of the most well-studied problems in combinatorial optimization is the vertex cover (VC) problem: given a graph $G=(V,E)$, we look for a minimum size subset of vertices such that for every $(u,v)\in E$, at least one of u,v belongs to this subset. In the *weighted* version of VC, each vertex has an integral weight, and we are looking for the minimum total weight subset of vertices with the property above.

Since the complexity of VC has been heavily studied since Karp's original proof of its NP-completeness [8], the related bibliography is vast and cannot be covered, of course, in this introductory note. We mention here that VC is known to be APX-complete [11], and moreover it cannot be approximated within a factor of 1.36 [5], unless P=NP. A 2-approximation on the other hand can be trivially obtained by taking all the vertices of a maximal matching in the graph.

Improving this simple 2-approximation algorithm has been a quite non-trivila task. The best approximation algorithms known before this work were published 20 years ago by Bar-Yehuda and Even [2], and by Monien and Speckenmeyer [10]. They achieved an approximation factor of $2-\frac{\ln\ln n}{2\ln n}$, where n is the number of vertices. If Δ is the maximum degree of the graph, Halperin [7] showed that a factor of $2-(1-o(1))\frac{2\ln\ln\Delta}{\ln\Delta}$ can be achieved by using the semidefinite programming (SPD) relaxation of VC.

In this work we use a stronger SDP relaxation to improve the approximation factor achieved in polynomial time to $2-\Theta(\frac{1}{\sqrt{\log n}})$. We observe that the introduction of all the so-called *triangle inequalities* to the standard SDP relaxation of VC is, in fact, very similar to the balanced cut SDP relaxation used by Arora, Rao, and Vazirani [1]. Then we use one of the main results of [1], which asserts that in the solution of this SDP, there are two big and well-separated vertex subsets. At the same time, we show that edges that were not covered by a trivial initial rounding are too big to have both of their endpoints in either of these two sets. Hence, one of these two big subsets has to be a

big independent set, which can be excluded. We show this process first for the unweighted VC, and then we show how it can be extended to the weighted case in a straight-forward manner. We end with some open problems.

2 The unweighted case

The following is a semidefinite-programming relaxation of unweighted Vertex Cover (VC) for a graph $G = (V, E)$ with n nodes:

$$\begin{aligned}
&\min \sum_{i=1}^{n} \frac{1 + v_0 v_i}{2} \text{ s.t.} && && \text{(SDP)}\\
&(v_0 - v_i)(v_0 - v_j) = 0, && \forall (i,j) \in E && (1)\\
&(v_i - v_j)(v_i - v_k) \geq 0, && \forall i,j,k \in V \cup \{0\} && (2)\\
&v_i^2 = 1, && \forall i \in V \cup \{0\} && (3)
\end{aligned}$$

where $v_i \in \mathbb{R}^{n+1}$. Constraints (2) are *triangular inequalities*, which must be satisfied by the vertex cover. In an 'integral' solution of (SDP) (which would correspond to a vertex cover of G), vertices that are picked coincide with v_0, while vertices that are not picked coincide with $-v_0$. In general though, an optimal solution of (SDP) will not be 'integral'.

In fact one can strengthen this SDP relaxation for VC by adding all so called triangle inequalities:

$$\begin{aligned}
&\min \sum_{i=1}^{n} \frac{1 + v_0 v_i}{2} \text{ s.t.}\\
&(v_0 - v_i)(v_0 - v_j) = 0, && \forall (i,j) \in E\\
&(v_i - v_j)(v_i - v_k) \geq 0, && \forall i,j,k \in V \cup \{0\}\\
&(v_i + v_j)(v_i - v_k) \geq 0, && \forall i,j,k \in V \cup \{0\}\\
&(v_i + v_j)(v_i + v_k) \geq 0, && \forall i,j,k \in V \cup \{0\}\\
&v_i^2 = 1, && \forall i \in V \cup \{0\}
\end{aligned}$$

This relaxation is in fact equivalent to the following relaxation: We add n more 'shadow' points to (SDP) so that for every unit vector $v_i,\ i = 1, \ldots, n$ we add unit vector v_i' which is the antipodal of v_i, i.e., $v_i v_i' = -1,\ \forall i$. Let V' be the set of shadow points. Note that in an integral solution of (SDP), exactly half (n) of the points in $V \cup V'$ coincide with v_0 and the other half coincide with $-v_0$. Therefore the following must hold

$$\sum_{i,j \in V \cup V'} |v_i - v_j|^2 = 4n^2$$

where every pair (i, j) appears only once in the sum. (Hence the set $V \cup V'$ is 1/2-spread in the terminology of [1]). In addition, the triangular inequalities (2) must also hold when we extend V with V'. Hence we have the following strengthened SDP:

$$\min \sum_{i=1}^{n} \frac{1+v_0 v_i}{2} \text{ s.t.} \tag{SDP'}$$

$$(v_0 - v_i)(v_0 - v_j) = 0, \qquad \forall (i,j) \in E \tag{4}$$

$$(v_i - v_j)(v_i - v_k) \geq 0, \qquad \forall i,j,k \in V \cup V' \cup \{0\} \tag{5}$$

$$v_i^2 = 1, \qquad \forall i \in V \cup V' \cup \{0\} \tag{6}$$

$$v_i v_i' = -1, \qquad \forall i \in V \tag{7}$$

$$\sum_{i,j \in V \cup V'} |v_i - v_j|^2 = 4n^2 \tag{8}$$

where $v_i, v_i' \in \mathbb{R}^d$ for some $d \gg \log n$. Constraint (8) is in fact unnecessary since it is always satisfied by a set of points and their antipodals, but we include it in order to point out that this relaxation defines a *spread metric* as defined in [1]. Now we can use results of [1] to find an approximate VC.

For any $\varepsilon > 0$, we define the following two sets of graph vertices:

$$\begin{aligned} S_1 &:= \{v \in V : v_0 v > \varepsilon\} \\ S_2 &:= \{v \in V \cup V' : -\varepsilon \leq v_0 v \leq \varepsilon\} \end{aligned}$$

For now, we concentrate our attention on S_2. Note that in S_2 we have included also shadow points. In fact, note that if $v_i \in V$ belongs to S_2 then its shadow $v_i' \in V'$ belongs to S_2 as well, and vice-versa. In other words, S_2 contains both original points and their shadows.

Lemma 1

$$\sum_{i,j \in S_2} |v_i - v_j|^2 = 4|S_2|^2$$

Proof: Note that for a particular pair $i, j \in S_2 \cap V$ we have $v_i v_j' = v_i' v_j = -v_i v_j$. So if we group the summation terms according to pairs of vertices $i, j \in S_2 \cap V$, we get the lemma, due to cancellation of terms. □

Let $\Delta, \sigma > 0$ be two parameters to be determined later. Let u be a random unit vector, and let

$$\begin{aligned} S_u &:= \{v \in S_2 : uv \geq \frac{\sigma}{\sqrt{d}}\} \\ T_u &:= \{v \in S_2 : uv \leq -\frac{\sigma}{\sqrt{d}}.\} \end{aligned}$$

Since $v_i = -v_i'$, it is easy to prove the following

Lemma 2 *If $v_i \in S_u$ for some $v_i \in V$, then $v_i' \in T_u$, and vice-versa, if $v_i' \in T_u$, then $v_i \in S_u$. The same holds with the roles of S_u, T_u interchanged.*

As a result of Lemma 2, $S_u \cup T_u$ contains only pairs of points in V with their shadow points, and each such pair is separated between S_u, T_u, and $|S_u| = |T_u|$. Moreover, the following easy fact also holds:

Lemma 3

$$\textit{If} \quad \left.\begin{array}{l} v_i \in S_u,\ v_j \in T_u,\ |v_i - v_j|^2 \leq \Delta \\ v_i' \in S_u,\ v_j \in T_u,\ |v_i' - v_j|^2 \leq \Delta \\ v_i \in S_u,\ v_j' \in T_u,\ |v_i - v_j'|^2 \leq \Delta \\ v_i' \in S_u,\ v_j' \in T_u,\ |v_i' - v_j'|^2 \leq \Delta \end{array}\right\} \Longrightarrow \left\{\begin{array}{l} v_j' \in S_u,\ v_i' \in T_u,\ |v_i' - v_j'|^2 \leq \Delta \\ v_j' \in S_u,\ v_i \in T_u,\ |v_i - v_j'|^2 \leq \Delta \\ v_j \in S_u,\ v_i' \in T_u,\ |v_i' - v_j|^2 \leq \Delta \\ v_j \in S_u,\ v_i \in T_u,\ |v_i - v_j|^2 \leq \Delta \end{array}\right.$$

Let $c' > 0$ be another parameter which will be defined later. We modify the procedure SET-FIND of [1] as follows:

- If $|S_u| < 2c'|S_2|$ or $|T_u| < 2c'|S_2|$ then we HALT (just like in [1]).
- Otherwise, pick any $x \in S_u, y \in T_u$ such that $|x-y|^2 \leq \Delta$. Then, because of Lemma 3, the corresponding pair of antipodal points $y' \in S_u, x' \in T_u$ also satisfy $|x'-y'|^2 \leq \Delta$. Delete x, x', y, y'. Repeat until no such x, y can be found.

Note that initially T_u contains the antipodal points of S_u (Lemma 2), and every deletion eliminates two points from each of S_u, T_u, and these four actually form two (a point in V, its shadow point in V') pairs. Therefore, in the end, the remaining points in S_u are *exactly* the antipodal points of T_u (or both S_u, T_u are empty). As in [1], $|x-y|^2 > \Delta$, $\forall x \in S_u, y \in T_u$. One can define the parameters c', σ so that, initially, S_u, T_u are big with high probability:

Lemma 4 [Lemma 4 in [1]] *For every positive $c < 1/3$, there are $c', \sigma > 0$ such that the probability (over the choice of u) is at least $c/8$ that the initial sets S_u, T_u defined above have size at least $2c'|S_2|$.*

Proof: From Lemma 1 and application of Lemma 4 of [1]. □

In fact, since S_u initially (and throughout the running of the algorithm) contains the antipodal points of T_u, $|S_u| = |T_u| = |S_2|/2$ before the algorithm starts running no matter which u we choose, therefore $c' = 1/4$.

One of the main results of [1] is to show that, with high probability over u, not many points are deleted before SET-FIND terminates. Note that the points removed form a matching (at every step, x is matched to y, and x' is matched to y'). Theorem 5 in [1] shows that, with $\Delta = O(\log^{-2/3} n)$, the probability that SET-FIND removes a matching of size $c'|S_2|$ is $o(1)$. Hence the final S_u, T_u of SET-FIND have size $\geq c'|S_2|$ with probability $\Omega(1)$, and $|S_u| = |T_u|$. In what follows, we assume that S_u, T_u are the big final sets we get with high probability from SET-FIND.

Lemma 5 *If $\varepsilon \leq \Delta/4$, then there is no edge $(i,j) \in E$ such that $v_i, v_j \in V$ belong both to S_u or both to T_u.*

Proof: W.l.o.g. suppose that there is $(i,j) \in E$ such that $v_i, v_j \in S_u$. Then their shadow (antipodal) points belong to T_u, i.e., $v_i', v_j' \in T_u$. Since $v_i, v_j \in S_2$ and constraint (4) holds, we have that

$$v_i v_j = v_0 v_i + v_0 v_j - 1 \leq -(1-2\varepsilon). \tag{9}$$

Since $v_i' \in T_u$ and $v_j \in S_u$ are not deleted in SET-FIND, $|v_i' - v_j|^2 > \Delta$, or, equivalently, $|v_i + v_j|^2 > \Delta$. This implies that

$$v_i v_j > -1 + \frac{\Delta}{2}. \tag{10}$$

But (9) and (10) together imply that $\varepsilon > \Delta/4$ which contradicts the hypothesis.

□

From now on we set $\varepsilon := \Delta/4 > 0$. Since $|S_u| = |T_u| \geq c'|S_2|$, and the two sets contain antipodal points, one of them (w.l.o.g. let's assume that this is S_u), contains at least $\frac{c'|S_2|}{2}$ points from V. Let I be this set of points from V. Lemma 5 implies that I is an *independent set* of G of size at least $c_0|S_2|$, where $c_0 := c'/2 > 0$. We return the set $S := S_1 \cup (S_2 \setminus (I \cup V'))$ as our vertex cover.

Lemma 6 *S is a vertex cover of G.*

Proof: If there is $(i,j) \in E$ with $v_i, v_j \in V \setminus (S_1 \cup S_2)$, we have (by the definition of S_1, S_2) that $v_0 v_i < -\varepsilon$ and $v_0 v_j < -\varepsilon$, which implies that $v_0 v_i + v_0 v_j - 1 < -1 - 2\varepsilon$. Then constraint (4) implies that $v_i v_j < -1 - 2\varepsilon$, a contradiction. Also, since I is an independent set, not both of v_i, v_j can belong to it. If $v_i \in I$ and $v_j \in V \setminus (S_1 \cup S_2)$, then $v_0 v_i \leq \varepsilon$ and $v_0 v_j < -\varepsilon$, therefore constraint (4) implies that $v_i v_j < -1$, a contradiction. We conclude that every edge must have at least one of its endpoints in S.

□

Our main result is the following

Theorem 1 $|S| \leq (2 - \Theta(\frac{1}{\log^{2/3} n}))VC(G)$.

Proof: We follow the analysis of Halperin [7]. From (SDP') and the definition of S_1, S_2 we have that

$$VC(G) \geq |S_1| \frac{1+\varepsilon}{2} + |S_2 \setminus V'| \frac{1-\varepsilon}{2}$$

or, equivalently,

$$|S_1| \leq \frac{2 \cdot VC(G)}{1+\varepsilon} - |S_2 \setminus V'| \frac{1-\varepsilon}{1+\varepsilon}. \tag{11}$$

Hence

$$|S| = |S_1| + |S_2 \setminus V'| - |I| \overset{(11)}{\leq} \frac{2}{1+\varepsilon} VC(G) + |S_2 \setminus V'| (\frac{2\varepsilon}{1+\varepsilon} - c_0).$$

Note that for $\Delta = \Theta(\log^{-2/3} n)$, $\frac{2\varepsilon}{1+\varepsilon} = \Theta(\log^{-2/3} n) < c_0$, for big enough n. Therefore,

$$|S| \leq \frac{2}{1+\varepsilon} VC(G) = (2 - \Theta(\log^{-2/3} n)) \cdot VC(G).$$

□

Very recently, J. Lee proved that the SET-FIND algorithm of [1] can also be used to obtain their stronger result [9], i.e., Δ can be as big as $\Theta(1/\sqrt{\log n})$. Therefore we can get the following strengthening of Theorem 1:

Theorem 2 $|S| \leq (2 - \Theta(\frac{1}{\sqrt{\log n}}))VC(G)$.

Theorem 2 can be somewhat strengthened by noticing that in the proof of Theorem 1 we just need to pick Δ so that $\frac{2\varepsilon}{1+\varepsilon} < c_0$, and therefore [1] and [9] imply that if $x := 1/\Delta^2$, it is enough for x to be the solution of equation

$$\frac{x}{\log x} = c \log n$$

where $c > 0$ is a constant (cf. [4] for more details on solving this equation through Lambert's W function).

3 The weighted case

The following is a semidefinite-programming relaxation of weighted Vertex Cover (VC) for a graph $G = (V, E)$ with n nodes:

$$\min \sum_{i=1}^{n} w_i \cdot \frac{1+v_0 v_i}{2} \text{ s.t.} \tag{WSDP}$$
$$\begin{aligned} (v_0 - v_i)(v_0 - v_j) &= 0, & \forall (i,j) \in E \\ (v_i - v_j)(v_i - v_k) &\geq 0, & \forall i,j,k \in V \cup \{0\} \\ v_i^2 &= 1, & \forall i \in V \cup \{0\} \end{aligned}$$

where w_i is the *integral* weight of node i. Let $W := \sum_{i=1}^{n} w_i$.

In order to apply the methods of Section 2, we solve (SDP') with the weights incorporated in the objective function, and replace every v_i by w_i copies of v_i (v_i' is also replaced by w_i copies of v_i'). In fact we don't need to do this replacement in practice, but this mental experiment is helpful in order to see how the unweighted case applies here, too. Note that this new set of vectors still satisfies the triangular inequalities, and Lemmata 4 through 6 in Section 2 apply here as well with $n := W$. Note that SET-FIND can be made to run in polynomial time in this case (recall that we don't really do the replacement of v_i with w_i, all we need to do is to keep track of how much weight remains for each node after each matching). Now Theorem 1 (and hence Theorem 2) can be proven in the same way as before, if we replace the cardinality of sets $|\cdot|$ with their weights $w(\cdot)$.

4 Open problems

Obviously one of the biggest open problems in theoretical computer science is the exact determination of the approximability of VC. There is a big gap between the hardness and the approximability results. In particular, we point out that any improvement of the sparsest cut results of [1] that improves the SET-FIND routine leads automatically to an improvement of the vanishing factor for VC. Maybe this is an indication of the power of the techniques of [1], in the sense that whether one believes that these techniques can lead to a constant factor approximation of the sparsest cut depends on his belief that 2 is the correct approximation factor for VC.

We couldn't extend our techniques to other problems related to VC, for example the maximum independent set problem (IS), and we don't know whether this is possible (Halperin's [7] techniques, on the contrary, can be applied to IS). Another extension of VC is the vertex cover problem in hypergraphs. We don't know how to extend our techniques to this problem as well. Therefore we leave the application of the results above to these and other problems as an open question.

Finally, we point out that we don't know what the integrality gap of the strengthened SDP relaxation (SDP') used above is. A weaker formulation, that doesn't contain all the triangle inequalities but is equivalent to Schrijver's θ' function [6], was proven to have an integrality gap of $2 - \varepsilon$ for any constant $\varepsilon > 0$ by Charikar [3]. It would be interesting to show the same result for the stronger SDP.

Acknowledgements I am grateful to Sanjeev Arora for reading an earlier draft of this work and for bringing [9] to my attention.

References

[1] S. Arora, S. Rao, and U. Vazirani. Expander flows, geometric embeddings and graph partitioning. In Proc. of 36th STOC, pp. 222–231, 2004.

[2] R. Bar-Yehuda and S. Even. A local-ratio theorem for approximating the weighted vertex cover problem. *Annals of Discrete Mathematics*, **25**, pp. 27–45, 1985.

[3] M. Charikar. On semidefinite programming relaxations for graph coloring and vertex cover. In Proc. of 13th SODA, pp. 616–620, 2002.

[4] R. M. Corless, G. H. Gonnet, D. E. G. Hare, D. J. Jeffrey, and D. E. Knuth. On the Lambert W Function. *Advances in Computational Mathematics*, vol. 5, pp. 329–359, 1996.

[5] I. Dinur and S. Safra. On the importance of being biased (1.36 hardness of approximating Vertex-Cover). *Annals of Mathematics*, to appear. Also in Proc. of 34th STOC, 2002.

[6] M. Goemans and J. Kleinberg. The Lovász Theta Function and a Semidefinite Programming Relaxation of Vertex Cover. *SIAM Journal on Discrete Mathematics* **11**(2), pp. 196–204, 1998.

[7] E. Halperin. Improved approximation algorithms for the vertex cover problem in graphs and hypergraphs. *SIAM J. on Computing*, **31**(5), pp. 1608–1623, 2002. Also in Proc. of 11th SODA, pp. 329–337, 2000.

[8] R. Karp. Reducibility among combinatorial problems. in R. E. Miller and J. W. Thatcher (eds.) *Complexity of Computer Computations*, Plenum Press, NY, pp. 85–103.

[9] J. R. Lee. Scale-based embeddings, Euclidean snowflakes, and the local theory of L_p spaces. Manuscript, see `http://www.eecs.berkeley.edu/~jrl`.

[10] B. Monien and E. Speckenmeyer. Ramsey numbers and an approximation algorithm for the vertex cover problem. *Acta Informatica*, **22**, pp. 115–123, 1985.

[11] C. E. Papadimitriou and M. Yannakakis. Optimization, approximation, and complexity classes. *JCSS*, **43**(3), pp. 425–440, 1991.