Improved Approximation Lower Bounds on Small Occurrence Optimization

Piotr Berman *   Marek Karpinski †

Abstract

We improve a number of approximation lower bounds for bounded occurrence optimization problems like MAX-2SAT, E2-LIN-2, Maximum Independent Set and Maximum-3D-Matching.

*Dept. of Computer Science and Engineering, The Pennsylvania State University. Research done in part while visiting Dept. of Comuter Science, University of Bonn. Work partially supported by NSF grant CCR-9700053, NIH grant 9R01HG02238-12 and DFG grant Bo 56/157-1. E-mail `berman@cse.psu.edu`.

†Dept. of Computer Science, University of Bonn. Research done in part while vising Dept. of Computer Science, Yale University. Work partially supported by DFG grants, DIMACS, and IST grant 14036 (RAND-APX). E-mail `marek@cs.uni-bonn.de`.

Problem	Former	Improved
3-MIS and 3D-matching	140	98
4-MIS	74	50
5-MIS	68	50
E3-OCC-E2-LIN-2	152	140
E3-OCC-2-LIN-2	121	112
E3-OCC-MAX-2-SAT	788	460
E3-OCC-MAX-E2-SAT	788	464
E4-OCC-MAX-E2-SAT	588	268
E4-OCC-MAX-2-SAT	588	262
4-OCC-MAX-2-SAT	588	252
E6-OCC-MAX-E2-SAT	308	164
E6-OCC-MAX-2-SAT	246	160

Figure 1: Summary of the results: values of k.

1 Introduction

We refer to [BK99] and [BK01] for a general background and notations. We define Ed-OCC-Ek-LIN-2 as a problem of constructing an assignment that maximizes the number of satisfied equations for a given system of linear equations modulo 2 (hence LIN-2), where each equation has exactly k variables (hence Ek) and each variable occurs exactly d times. If we drop an E in the acronym of the problem than we have "at most d occurrences" or "at most k variables". We replace Ek-LIN-2 with MAX-Ek-SAT if we maximize the number of satisfied disjunctive clauses. d-MIS problem is the problem of maximizing the size of an independent set in a d-regular graph.

Each result of this paper (for a Problem X considered) is of the following form: *if* $0 < \varepsilon < 1/(k-1)$, *it is NP hard to approximate a Problem X to within a factor* $k/(k-1) - \varepsilon$.

The challenge is to obtain as small k as possible for every problem. Fig. 1 summarizes the progress of this paper as compared with the previous results [BK99], [BK01], and [CC02].

2 Amplifiers

The notion of an amplifier generalizes the concept of a specific variety of expanders that are used in proving inapproximability results. This notion was introduced by Papadimitriou in [P94] (for directed graphs) and it formalizes the construction of

Papadimitriou and Yannakakis of [PY91], see also [AL97].

Consider an undirected graph $G = (V, E)$. We define

$\mathrm{Cut}(U) = \{e \in E : e \not\subset U \text{ and } e \not\subset V - U\}$ and $\mathrm{cut}(U) = |\mathrm{Cut}(U)|$.

We say that G is a *strong expander* if for every $U \subset V$ we have $\mathrm{cut}(U) \geq \min(|U|, |V-U|)$.

We say that G is an *amplifier for* $X \subset V$ if it contains no *bad sets* for X. A set $A \subset V$ is bad for X if $\mathrm{cut}(A) < \min(|X \cap A|, |X - A|)$.

An amplifier for X is B-*regular* if each node in X has $B-1$ neighbors and each node in $V - X$ has B neighbors.

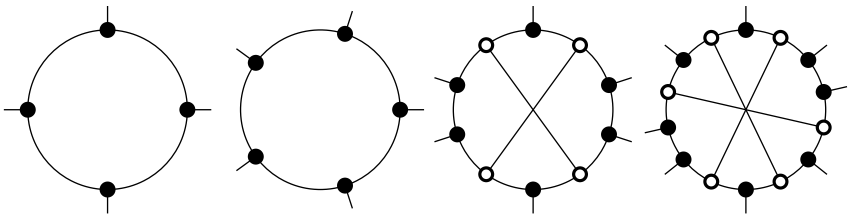


Figure 2: 3-regular amplifiers for $|X| = 4, 5, 6, 8$, $\bullet$'s are the elements of X.

We have the following results on constructibility of regular expanders.

Theorem 1. *For a set with n nodes, in random linear time one can construct a 3-regular amplifier with $7n$ nodes and $10n$ edges.*

Proof. Berman and Karpinski [BK99]. ❒

Theorem 2. *For a set with n nodes, in random linear time one can construct a 7-regular amplifier with n nodes and $3n$ edges.*

Proof. Bolobás [Bo88]. ❒

We are going to prove the following theorem.

Theorem 3. *For a set with n nodes, in random linear time one can construct a 5-regular amplifier with $1.8n$ nodes and $4n$ edges.*

Proof. The construction is as follows: start with two sets $\breve{X} = \{\breve{x}_0, \ldots, \breve{x}_{4n-1}\}$ and $\breve{U} = \{\breve{u}_0, \ldots, \breve{u}_{4n-1}\}$. Pick a random matching between $\breve{X}$ and $\breve{U}$. For $i < n$ collapse $\breve{x}_{4i}, \ldots, \breve{x}_{4i+3}$ into one node of X, and for $i < 0.8n$ collapse $\breve{u}_{5i}, \ldots, \breve{u}_{5i+4}$ into one node of $U = V - X$. Repeat until the resulting graph has exactly $4n$ edges.

A bad set $A \subset V$ can be normalized. We first define $B = A \cap X$ and U_i as the set of elements of U with exactly i neighbors in B. We can replace A with $B \cup U_3 \cup U_4 \cup U_5$: $A \cap X$ is unchanged and $\mathrm{cut}(A)$ does not increase, thus this new A is still bad.

Let $k = |B|$ and $a_i = |U_i|$. Clearly, $\mathrm{cut}(A) = a_1 + a_4 + 2(a_2 + a_3) < k$.

We need to show that the probability that our graph is not an amplifier converges to 0 as n increases. In particular, the probability that a fixed set $B \subset X$ defines a bad set is much smaller than $C(n,k)^{-1}$, where $C(n,k)$ is the binomial coefficient. Because we can choose either B or $X - B$ for our discussion, we assume that $k \leq n/2$.

Let P' be the probability that B defines a bad set, and let $P = P'C(n,k)$. P is an upper bound on the probability that there exists a bad set. We establish probability P' as follows: B contains $4k$ edge ends. The other ends of these edges are in set $\check{U}$, and every set of $4k$ such ends is equally probable, hence term $C(4n,4k)^{-1}$ in P'. Then we count the number of such sets that are consistent with parameters $a_1, \ldots, a_5$. We have $C(0.8n; a_0, \ldots, a_5)$ partitions of U into $U_0, \ldots, U_5$. For each element of U_i we select in $C(5,i)$ ways the edge ends that can connect to set B. Thus we get the following formula for P:

$$P = \frac{0.8n! \qquad 5^{a_1+a_4}\ 10^{a_2+a_3}\ 4k!\ 4(n-k)! \qquad n!}{a_0!\ a_1!\ a_2!\ a_3!\ a_4!\ a_5! \qquad\qquad 4n! \qquad k!\ (n-k)!}.$$

We will find the parameters that yield the maximum probability. We have the following system for $\alpha_i = a_i/n$ and $\kappa = k/n$:

$$\begin{aligned}
\alpha_0 + \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 + \alpha_5 &= 0.8 && (1)\\
\alpha_1 + 2\alpha_2 + 2\alpha_3 + \alpha_4 &= \kappa' \leq \kappa && (2)\\
\alpha_1 + 2\alpha_2 + 3\alpha_3 + 4\alpha_4 + 5\alpha_5 &= 4\kappa && (3)\\
\alpha_1\alpha_3^3 &= \alpha_2^3\alpha_4 && (4)\\
10\alpha_1^2 &= 25\alpha_0\alpha_2 && (5)\\
10\alpha_4^2 &= 25\alpha_5\alpha_3 && (6)
\end{aligned}$$

Equation (1) says that the union of U_i's forms U, (2) says that $\mathrm{cut}(A) \leq k$, (3) says that $\mathrm{cut}(B) = 4k$, and equations (4-6) say that we cannot increase P by little changes in the values of α's.

We consider three ways that change a's without changing $\mathrm{cut}(A)$ and $\mathrm{cut}(B)$. First, we can add $(1. - 3, 3, -1)$ to (a_1, a_2, a_3, a_4), i.e. increment a_1, decrement a_2 by 3 etc. This multiplies P with

$$\frac{a_2(a_2-1)(a_2-2)a_4}{(a_1+1)(a_3+1)(a_3+2)+(a_3+3)} \approx \frac{a_2^3 a_4}{a_1 a_3^3}.$$

If we assume that neither this change not its opposite increase P we obtain (4). (5) and (6) are similar.

It is easy to see that (4-6) hold iff for some α, β, γ we have

$$\begin{aligned}
\alpha_0 &= \alpha & \alpha_1 &= 5\alpha\beta & \alpha_2 &= 10\alpha\beta^2\\
\alpha_5 &= \alpha\gamma^5 & \alpha_4 &= 5\alpha\beta\gamma^3 & \alpha_3 &= 10\alpha\beta^2\gamma
\end{aligned}$$

Suppose that $\gamma > 1$, then $5\alpha_5 > 2.5(\alpha_0 + \alpha_5)$, $\alpha_1 + 4\alpha_4 > 2.5(\alpha_1 + \alpha_4)$, $2\alpha_2 + 3\alpha_4 > 2.5(\alpha_2 + \alpha_3)$, which with (1) and (3) imply that $4\kappa > 2$, hence $k > n/2$, a contradiction. Therefore $\gamma \leq 1$.

Because $4\,\mathrm{cut}(A) < \mathrm{cut}(B)$, we have

$$20\alpha\beta + 80\alpha\beta^2 + 80\alpha\beta^2\gamma + 20\alpha\beta\gamma^3 \le 5\alpha\beta + 20\beta^2 + 30\alpha\beta^2\gamma + 20\alpha\beta\gamma^3 + 5\alpha\gamma^5 \iff$$

$$4\beta + 16\beta^2 + 16\beta^2\gamma \le \beta + 4\beta^2 + 6\beta^2\gamma + \gamma^5 \iff 3\beta + 12\beta^2 + 10\beta^2\gamma \le \gamma^5.$$

If $6\beta > \gamma$ and $\gamma \le 1$ then $3\beta + 12\beta^2 + 10\beta^2\gamma \ge (\frac{3}{6} + \frac{12}{36} + \frac{10}{36})\gamma^5$, a contradiction. Therefore $6\beta < \gamma$.

Suppose that $\kappa' < \kappa$. Then we can increase $\mathrm{cut}(A)$, i.e. by adding $(-1, 1, 1, -1)$ to (a_0, a_1, a_4, a_5). This changes P roughly by a factor of $25\alpha_0\alpha_5\alpha_1^{-1}\alpha_4^{-1} = \gamma^2\beta^{-2} > 36$. Therefore $\kappa' = \kappa$.

Suppose that $\kappa < 0.5$. Then we can decrease n, say be removing 5 nodes from X and 4 nodes from U, and since we do not want to change $\mathrm{cut}(A)$ or $\mathrm{cut}(B)$, we remove these 4 nodes from U_0. The resulting change in P is the factor

$$\frac{\alpha_0^4}{(0.8n)^4}\frac{1}{(1-\kappa)^{15}}$$

To show that this factor is larger then 1, it suffices to show that $a_0/0.8 \ge (1-\kappa)^3.75$. Because $(1-x)^{3.75}$ is convex, $(1-0)^{3.75} = 1 - 2 \times 0$ and $(1-0.4)^{3.75} \approx 0.147$, $(1-\kappa)^{3.75} \le \max(0.2, 1-2\kappa)$. Thus it suffices to show that $a_0/0.8 \ge \max(0.2, 1-2\kappa)$ and this is pretty easy.

We conclude that $\kappa' = \kappa = 0.5$. We can now repeat the reasoning that showed $\gamma \le 1$ to show that in this case $\gamma = 1$. Thus

$$\alpha_0 = \alpha_5 = \alpha \qquad \alpha_1 = \alpha_4 = 5\alpha\beta \qquad \alpha_2 = \alpha_3 = 10\alpha\beta^2$$

and equalities (2-3) translate into

$$\begin{aligned} 10\alpha\beta + 40\alpha\beta^2 &= 0.5 \\ 5\alpha + 25\alpha\beta + 50\alpha\beta^2 &= 2 \end{aligned}$$

Thus $5 + 25\beta + 50\beta^2 = 2/\alpha = 40\beta + 160\beta^2 \iff 22\beta^2 + 3\beta - 1 = 0$. This gives $\beta \approx 0.155656$, $\alpha = 0.197964$, and thus

$$\alpha_0 = \alpha_5 = 0.197964 \qquad \alpha_1 = \alpha_4 = 0.154072 \qquad \alpha_2 = \alpha_3 = 0.047964$$

Stirling formula and $\phi(x) = x^x$ allow to approximate $P^{1/n}$ as

$$\frac{\Phi(0.8)\ 5^{0.30814}\ 10^{0.9592}\ 2}{\Phi(0.19796)^2\ \Phi(0.15407)^2\ \Phi(0.04796)^2\ 2^4} < 0.969.$$

❐

3 Eq-Reductions

The following notion of a *gap property* was introduced in [BK01]:

> $(f(n), g(n))$ gap property of an optimization problem A means that for every sufficiently small positive ε it is NP-hard to distinguish between two groups of instances of A of size n: those that have no solutions with score above $f(n) + \varepsilon n$ and those that have solutions with score at least $g(n) - \varepsilon n$.

While not formalized in exactly that fashion, gap properties were widely used in proving lower bounds on approximation ratios that can be attained by polynomial time algorithms.

For example, Håstad [H97] has shown that if $0 < \varepsilon < 0.5$ then for systems of n linear equations modulo 2 with 3 variables per equation it is NP-hard to distinguish between instances where a solution may satisfy $n - \varepsilon$ equations and instances where no solution satisfies more then $n/2 + \varepsilon$ equations. Thus the problem E3-LIN-2 has instances with even number of equations modulo 2 with 3 variables each, n the number of equations in an instance and this problem has $(n/2 + \varepsilon, n - \varepsilon)$ gap property. We will be omitting ε terms, so we can say that this problem has $(n/2, n)$ gap property.

We define the Eq-reductions as tools to prove gap properties.

Consider two maximization problem, A and B with objective functions a and b. An Eq-reduction from A to B has 5 randomized polynomial time computable functions, τ, t, ν, ρ and r, in its description:

- instance translation τ and parameter translation t; if x is an instance of A with parameter n then $\tau(x)$ is an instance of B with parameter $t(n)$;
- solution normalization ν; if y is a solution of $\tau(X)$, then $\nu(u)$ is another solution of $\tau(X)$ such that $b(\nu(y)) \geq b(y)$;
- solution equivalence ρ and value equivalence r; let $S_P(x)$ be the set of solutions of an instance x of problem P, ρ is 1-1 onto function from $S_A(x)$ to $\nu(S_B(\tau(x)))$ such that $b(\rho(s)) = r(a(s), n)$.

Observation 4. *Assume that problem* A *has* $(f(n), g(n))$ *gap property and that there exists an Eq-reduction from* A *to* B *with the parameters described above. Then problem* B *has* $(r(f(n), n), r(g(n), n))$ *gap property.*

3.1 Reducing E3-LIN-2 to E2-LIN-2

We refer to [BK01] for the corresponding discussion on standard reductions for linear equations, and describe a reduction from E3-LIN-2 into 2-LIN-2. Consider a system E of n equations modulo 2 with 3 variables per equation. We define $\tau(E)$ by replacing, one by one, each equation in E. Given an equation $w + x + y = b$, we replace it with $S(w, x, y, b)$. Because b is actually a constant, we have 12 equations with 2 variables and 4 equations with 1 variable (which must be an auxiliary one), thus $t(n) = 16(n)$. If the parameter of an instance of $E2-LIN-2)$ is (the number of equations with one variable, the number of equations with two variables), then $t(n) = (4n, 12n)$.

Let x be the vector of the variables of E and a be the vector of the auxiliary variables of $\tau(E)$. Given a value of (x, a) we can compute $\nu(x, a)$ by changing each a_i in such a way that a maximal number of equation is satisfied, if the two choices are equally good, we set $a_i = 0$. Because no equation involves two auxiliary variables, these value selections cannot conflict and they can be performed independently.

The solution equivalence is $\rho(x) = \nu(x, a')$, observe that $\nu(x, a')$ does not depend in a'. It is easy to see that the value equivalence is $r(k, n) = 10n + 2k$.

Value equivalence $10n + 2k$ translates $(n/2, n)$ gap property of E3-LIN-2 into $(10n + n, 10n + 2n) = (11n, 12n)$ gap property of 2-LIN-2; of we wish n to refer to the size of the new instance, i.e. $16n$, we got $(11/16\ n, 12/16\ n)$ gap.

Remark 1, The system $\tau(E)$ consists of equations that have 1 or 2 variables. We can define a similar reduction where we introduce a new variable z, and we first replace each equation $w+x+y = b$ with $w+x+y+z = b$ and then replace the new equation with a system of 16 equations as described above. We will use $\tau'(E)$ to denote the resulting system of equations with 2 variables each. This was the original reduction of Håstad [H97].

Remark 2, In the subsequent reductions we will assume that each variable in an instance of 2-LIN-2 or E3-LIN-2 has a sufficient number of occurences, For example, we can replicate all equations n times, so in terms of new number of equations each variable occurs at least $n^{1/2}$ many times.

3.2 Hardness of E3-OCC-3-LIN-2

Given an instance of E3-LIN-2 where each variable occurs sufficiently often, we can replace it with an instance in which each variable occurs exactly 3 times. Suppose that we start with $2n$ equations, so we have $6n$ variable occurences. We make each variable occurence a separate variable; given m *contact* occurences of a variable, we add $6m$ *checker* occurences. We connect these occurences with a graph that is a 3-regular amplifier for the contact occurences and then we replace each edge $\{x, y\}$ with equality $x = y \equiv x + y = 0$. As analyzed in [BK99], this creates an instance of E3-OCC-3-LIN-2 with $2n$ equations of length 3, $60n$ equations of length 2 and for which it is difficult to tell if we can satisfy almost all equations or at most $(61 + \varepsilon)n$ of them. We will call such an instance a *Hybrid instance.*

Chlebík and Chlebíková [CC02] showed that we can reduce the number of equations in the amplifiers by $0.9n$, which automatically improves some of the results discussed in this note.

4 Hardness of k-MIS

We are going to discuss now small degree instances of MIS problem.

4.1 Hardness of 4-MIS

Theorem 5. *For every* $\varepsilon \in (0, 1/49)$*, it is NP hard to approximate 4-MIS to within a factor* $50/49 - \varepsilon$*.*

Proof. Given an instance of E3-LIN-2 and the corresponding Hybrid instance construct an instance of 4-MIS as follows.

- For every variable x that is an amplifier node we create an edge $P_x = \{x_0, x_1\}$.
- For every amplifier edge $\{x, y\}$ we connect P_x and P_y with edges $\{x_0, y_1\}$ and $\{x_1, y_0\}$.
- For every variable x that is a *contact occurence* we create a *direct contact*, an edge $D_x = \{\breve{x}_0, \ \breve{x}_1\}$ that is connected with P_x with edges $\{x_0, \breve{x}_1\}$ and $\{x_1, \breve{x}_0\}$,
- For an equation $e \equiv x + y + z = b$ we create *equation quadruple* Q_e of nodes of the form $\boxed{\alpha\beta\gamma}$ where nodes of the form $\boxed{\alpha\beta\gamma}$ where $\alpha + \beta + \gamma = b$ **mod** 2. We connect $\boxed{\alpha\beta\gamma}$ with $\breve{x}_{\neg\alpha}$, $\breve{y}_{\neg\beta}$ and $\breve{z}_{\neg\gamma}$. We connect the nodes of Q_e with arbitrary two disjoint edges.

A Hybrid instance created from an E3-LIN-2 instance with $2n$ equations is translated into a graph G that has $8 \times 6n$ pairs (for nodes in the amplifiers and the direct contacts) and $2n$ Q-quadruples.

Let $J_{\alpha\beta\gamma} = \{\breve{x}_\alpha, \breve{y}_\beta, \breve{z}_\gamma\}$ and let $I_{\alpha\beta\gamma} = J_{\alpha\beta\gamma} \cup \{\boxed{\alpha\beta\gamma}\}$.

Lemma 6. *Given equation* $e \equiv x + y + z = b$*, the gadget of* e *is* $A_e = Q_e \cup D_x \cup D_y \cup D_z$*. Every maximum independent set in* A_e *is of the form* $I_{\alpha\beta\gamma}$*.*

Proof. Consider an independent set $J \subset A_e$. Because the four nodes of Q_e are connected with a matching, $|J \cap Q_e| \leq 2$. Note that every two nodes in Q_e have exactly one common neighbor in $A_e - Q_e$ and thus together they have 5 such neighbors. Therefore if $|J \cap Q_e| = 2$ then $|J - Q_e| \leq 2$ and $|J| \leq 3$. Because $|J \cap D_v| \leq 1$, if $|J \cap Q_e| = 0$ then $|J| \leq 3$. Finally, if $|J \cap Q_e| = 1$, then for some α, β, γ we have $J \cap Q_e = \{\boxed{\alpha\beta\gamma}\}$, and $\breve{x}_\alpha, \breve{y}_\beta, \breve{z}_\gamma$ are the only nodes of $A_e = Q_e$ that are not connected to $\boxed{\alpha\beta\gamma}$. Therefore if $|J| = 4$ then $J = I_{\alpha\beta\gamma}$. ❐

We will describe a normalization of an independent set I in G in stages. Apply each rule in turn as long as possible.

(i) Normalization of equation gadgets. Consider equation $x + y + z = b$ and its gadget A_e. One of the cases below must apply.

Case a: there exist α, β, γ are such that $\alpha + \beta + \gamma = b$ **mod** 2 and $x_{\neg\alpha} \notin I$, $y_{\neg\alpha} \notin I$, $z_{\neg\alpha} \notin I$. We set $I \cap A_e$ to be $I_{\alpha\beta\gamma}$.

Case b: case **a** does not hold, i.e. $P_x \cap I = \{x_\alpha\}$, $P_y \cap I = \{y_\beta\}$, $P_z \cap I = \{z_\gamma\}$, and $\alpha + \beta + \gamma \neq b$. We set $I \cap A_e$ to be $J_{\alpha\beta\gamma}$.

(ii) First assignment of Boolean values. To every D- or P- pair that contains a node of I with subscript α give value α, note that after the normalization of equation gadgets every D-pair has a value.

(iii) Second assignment of Boolean values. If P_x has no Boolean value assigned and no neighbor with value $\neg\alpha$, insert x_α to I and assign value α to P_x.

(iv) Third assignment of Boolean values. Let W be the graph where nodes are pairs without Boolean values, Because a pair in W has a neighbor with value 0 and a neighbor with value 1, it has at most one neighbor in W. Thus W has connected components of size 1 or 2.

Consider a connected component of size 1: its pair has two neigbors with value b and one with value $\neg b$, we give this pair value b.

Consider a connected component of size 2: it has four neighbors, two with value 0 and two with value 1; we give value 0 to the elements of this component.

Now every element of W is adjacent to a different edge between pairs with different Boolean values.

(v) Normalization of Boolean values. After this assignment of values, the connections between pairs with value 0 and pairs with value 1 form a matching. If the size of this matching within an amplifier A is a, and this amplifier contains some $7m$ P-pairs, then $|I \cap A| = 7m - a$. If A has $k \leq m/2$ direct contact pairs with value $\neg b$ then the property of the amplifiers says that $a \geq b$. We convert entire A and the adjacent direct contacts to value b, thus assuring that $|I \cap A| = 7m$; to keep I as the independent set, we may need to remove its elements from b variable quadruples.

After normalization, every pair in amplifiers and every contact pair has exactly one element in I, and we have exactly one value given to all occurences of a variable of the original instance of E3-LIN-2. One can also see that an equation quadruple contains an element of I if and only if this equation is satisified by the values given to the variables. Given $2n$ equations, we had $48n$ pairs and $2n$ quadruples, thus n vs $2n$ question is translated into $49n$ vs $50n$ question. ❒

In terms of τ defined in [CC02] we can improve the above bound from 50 to $6\tau+8$.

4.2 Hardness of 3-MIS

Theorem 7. *For every $\varepsilon \in (0, 1/97)$, it is NP hard to approximate 3-MIS to within a factor $98/97 - \varepsilon$.*

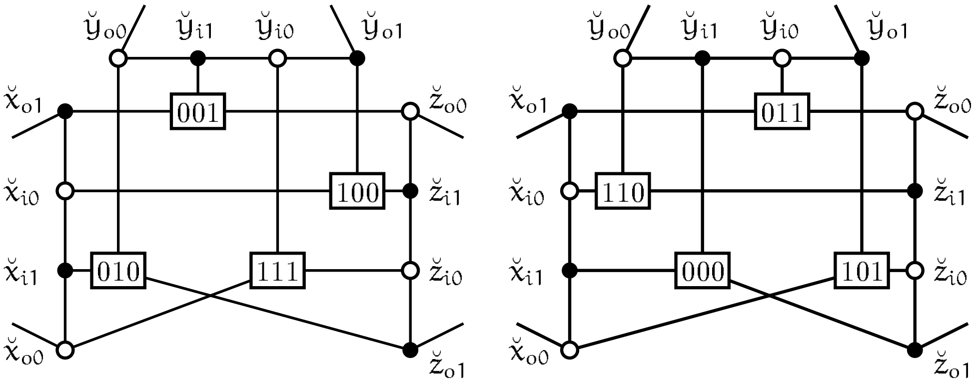


Figure 3: Gadgets of equations $x + y + z = 1 \textbf{ mod } 2$ and $x + y + z = 0 \textbf{ mod } 2$.

Proof. Given an instance of E3-LIN-2 and the corresponding Hybrid instance construct an instance of 3-MIS as follows. We use the fact that amplifiers of the Hybrid instance have the following structure: all nodes are on a single cycle, which we will view as directed for the sake of the construction, and the nodes that are *checker occurences* are connected with an additional matching.

- For every variable x that is an amplifier node we create a path $P_x = (x_{o0}, x_{i1}, x_{i0}, x_{o1})$.
- For every amplifier cycle edge (x, y) we connect P_x and P_y with edges $\{x_{i0}, y_{o1}\}$ and $\{x_{i1}, y_{o0}\}$.
- For every amplifier matching edge (x, y) we connect P_x and P_y with edges $\{x_{o0}, y_{o1}\}$ and $\{x_{o1}, y_{o0}\}$.
- For every variable x that is a *contact occurence* we create a *direct contact*, a path $D_x = (\check{x}_{o0}, \check{x}_{i1}, \check{x}_{i0}, \check{x}_{o1})$. We connect P_x with D_x with edges edges $\{x_{o0}, \check{x}_{o1}\}$ and $\{x_{o1}, \check{x}_{o0}\}$.
- For an equation $e \equiv x + y + z = b$ we create *equation quadruple* Q_e of nodes of the form $\boxed{\alpha\beta\gamma}$ where $\alpha + \beta + \gamma = b \textbf{ mod } 2$. We connect Q_e, D_x, D_y and D_z as shown in Fig. 3.

A Hybrid instance created from an E3-LIN-2 instance with $2n$ equations is translated into a graph G that has $7 \times 6n$ paths of length 4 (for nodes in the amplifiers and the direct contacts) and $2n$ Q-quadruples.

Given equation $e \equiv x + y + z = b$, the gadget of e is $A_e = Q_e \cup D_x \cup D_y \cup D_z$. We define $J_{\alpha\beta\gamma} = \{\check{x}_{o\alpha}, \check{x}_{i\alpha}, \check{y}_{o\beta}, \check{y}_{i\beta}, \check{z}_{o\gamma}, \check{z}_{i\gamma}\}$ and $I_{\alpha\beta\gamma} = J_{\alpha\beta\gamma} \cup \{\boxed{\alpha\beta\gamma}\}$.

Lemma 8. *If* $J \subset A_e$ *is an independent set, then* $|J| \leq 7$.

Proof. We present the proof for the case when $b = 1$, using the left part of Fig. 3. We cover A_e with a cycle of length 7: $(\check{x}_{o0}, \check{x}_{i1}, \boxed{010}, \check{y}_{o0}, \check{y}_{i1}, \check{y}_{i0}, \boxed{111})$ and a path of length 9: $(\check{z}_{o1}, \check{z}_{i0}, \check{z}_{i1}, \check{z}_{o0}, \boxed{001}, \check{x}_{o1}, \check{x}_{i0}, \boxed{100}, \check{y}_{o1})$. Clearly, if $|J| > 7$ then J must have the maximum number of nodes in the cycle and on the path, i.e. 3 and 5 nodes respectively. Thus J contains these nodes from the path: $\check{z}_{o1}, \check{z}_{i1}, \boxed{001}, \check{x}_{i0}$ and $\check{y}_{o1}$, and only 3 nodes on the cycle are not their neighbors: $\check{x}_{o0}, \boxed{111}$ and $\check{y}_{o0}$. However, J cannot contain both $\boxed{111}$ and $\check{x}_{o0}$. ❐

Lemma 9. *Assume that* $e \equiv x+y+z = b \ \mathbf{mod}\ 2$*, I is an independent set,* $J = I \cap A_e$*,* $x_{o\alpha} \in P_x \cap I$*,* $y_{o\beta} \in P_y \cap I$*,* $z_{o\gamma} \in P_z \cap I$*, and* $\alpha+\beta+\gamma \neq b$*. Then* $|J| \leq 6$*.*

Proof. In the proof of Lemma 8 we argued that if $|J \cap Q_e| > 2$ then $|J| \leq 6$. Thus we can assume that $|J \cap Q_e| \leq 2$. Assume by the way of contradiction that $|J| > 6$, then $|J \cap D_v| > 1$ for two v's among x, y, z, say x and y. This implies that $J \cap (D_x \cup D_y) = \{\check{x}_{i\alpha}, \check{x}_{o\alpha}, \check{y}_{i\beta}, \check{y}_{o\beta}\}$ and the only element of Q_e that may belong to J is $\boxed{\alpha\beta\neg\gamma}$; consequently $|J \cap D_z| = 2$. Because $\check{z}_{o\neg\gamma}$ is adjacent to $z_{o\gamma}$, $J \cap D_z = \{z_{i\gamma}, z_{o\gamma}\}$. This is a contradiction because one of these two nodes must be a neighbor of $\boxed{\alpha\beta\neg\gamma}$. ❐

We will describe a normalization of an independent set I in G in stages. Apply each rule in turn as long as possible.

(i) Normalization of equation gadgets. Consider equation $x+y+z=b$ and its gadget A_e. One of the cases below must apply.

  Case a: there exist α, β, γ are such that $\alpha+\beta+\gamma = b \ \mathbf{mod}\ 2$ and $x_{o\neg\alpha} \notin I$, $y_{o\neg\alpha} \notin I$, $z_{o\neg\alpha} \notin I$. We set $I \cap A_e$ to be $I_{\alpha\beta\gamma}$.

  Case b: case **a** does not hold, i.e. $P_x \cap I = \{x_\alpha\}$, $P_y \cap I = \{y_\beta\}$, $P_z \cap I = \{z_\gamma\}$, and $\alpha+\beta+\gamma \neq b$. We set $I \cap A_e$ to be $J_{\alpha\beta\gamma}$.

(ii) Elimination of ambigous paths. We say that a amplifier variable v is ambigous if $\{v_{o0}, v_{o1}\} \subset I$. Suppose that (x, y) is a cycle edge of an amplifier, y is ambigous and x is not. If $x_{o0} \notin I$ we remove y_{oO} from I and replace it with x_{i1}, and if $x_{o1} \notin I$, we replace y_{o1} with x_{i0}. Note that a matching edge of an amplifier cannot connect ambigous variables; therefore this rule eliminates all ambigous variables.

(iii) Temporary removal of nodes. If $|P_x \cap I| \leq 1$, I becomes $I - P_x$. Let a be the number of such paths.

(iv) First assignment of Boolean values. To every D- or P- path that contains two nodes of I with subscript α give value α. Because we have normalized the equation gadgets, every D_x has a value assigned.

(v) Second assignment of Bolean values. If $P_x \cap I = \emptyset$ and no neighbor of P_x has assigned value $\neg\alpha$, we assign value α to P_x and insert $x_{i\alpha}$ and $x_{o\alpha}$ to I. Remaining P_x's without assigned values have one neighbor with value 0 and one with value 1.

(vi) Putting back the removed node. For every path P_x such that $P_x \cap I = \emptyset$, consider the edge (x, y) of an amplifier cycle. If $y_{o0} \notin I$, insert x_{i0}, otherwise insert x_{i1}.

(vii) Perform the third assignment of Boolean values and the normalization of Boolean values as in the proof of Theorem 5.

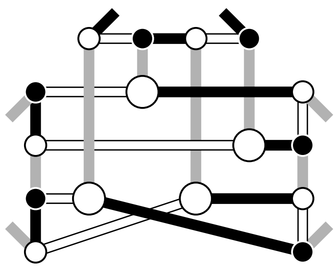
After normalization, every P_x and every D_x contains exactly two elements in I, and we have exactly one value given to all occurences of a variable of the original instance of E3-LIN-2. One can also see that an equation quadruple contains an element of I if and only if this equation is satisified by the values given to the variables. Given $2n$ equations, we had $48n$ paths and $2n$ quadruples, thus n vs $2n$ satisfied equations translates into $(48\times2+1)n$ vs $(48\times2)+2n$ nodes in an independent set. ❒

In terms of τ defined by [CC02] we can improve the above bound from 98 to $12\tau + 14$.

4.2.1 Hardness of 3D-Matching

In 3D-Matching problem we are given 3-partitie hypergraph with node set $V_0 \cup V_1 \cup V_2$ and hyperedge set E such that for every edge e and for $i = 0, 1, 2$ we have $e \cap V_i| = 1$. A matching is a set of pairwise disjoint hyperedges and we want to approximate a maximum matching.

We can put another restriction on the problem: every node belongs to exactly two hyperedges. Then the line dual hypegraph is a 3-regular graph. This restricted 3D-Matching is a restricted 3-MIS. The restriction is that we can color edges with 3 colors and each node belongs to edges of 3 different colors.



We can provide the 3 coloring to the instances of 3-MIS produced in the proof of Theorem 7 if we restrict them a bit. First, the amplifiers should form a bipartite graph, second, one the cycles of amplifiers some contacts should be separated by six checkers, and some by five (however few).

As we see here on the right, and equation gadgets can be colored provided that (a) every two edges that connect the gadget to an amplifier have the same color, and (b) not all edges connecting the gadget to an amplifier have the same color.

Fig. 4 shows how we can color edges inside an amplifier. In this figure every path P_x is depicted as a column, with white and black cirles indicating nodes that correspond to the two Boolean values. Edges that correspond to the matching edges of the amplifiers are the short incomplete edges that extend up and down from the columns, and edges that connect to equation gadgets are similar, except longer.

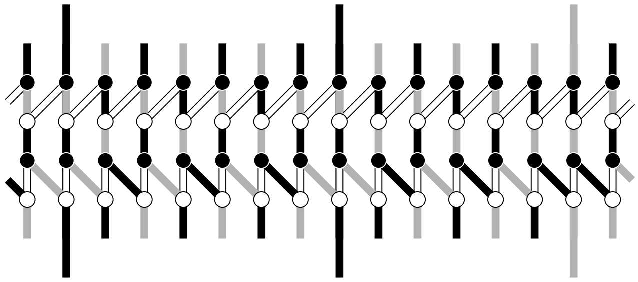


Figure 4: Coloring in the amplifier

5 Hardness of k-OCC-MAX-2-SAT

5.1 Hardness of E3-OCC-2-LIN-2 and E3-OCC-E2-LIN-2

The results of this sections are obtained by modifying the Eq-reductions that are described in the following lemma.

Lemma 10. *There exists an Eq-reduction* R *from E3-LIN-2 to E3-OCC-2-LIN-2 with value equivalence function* $110n + 2k$ *and an Eq-reduction* R' *from E3-LIN-2 to E3-OCC-E2-LIN-2 with value equivalence* $138n + 2k$.
Proof. Given a system of equations E we describe the instance transformation in five steps. Whenever we refer two an edge between variables x and y we mean their equality, i.e. the equation $x + y = 0$.

(i) For R' only: add z to each equation.

(ii) For a variable x that has m occurrences, create a 3-regular amplifier with $2m$ contacts. Every node in this amplifier is a variable.

(iii) Replace each equation of E, say $x_0 + x_1 + x_2 = b$ with 16 equations of $S(x_0, x_1, x_2, b)$. Next, replace each variable occurence in $S(x_0, x_1, x_2, b)$ with a new variable, occurences of one variable form *quadruples*.

(iv) Connect quadruples of auxiliary variables into simple cycles.

(v) To each quadruple of a variable x, say x^0, x^1, x^2, x^3 add two extra variables x^4, x^5 and connect them into a simple cycle $(x^0, x^4, x^1, x^2, x^5, x^3)$. Connect x^4 and x^5 with two contacts of the amplifier of x; make sure that each contact is used only once in this manner.

The solution normalization is described in four stages.

(i) In each amplifier and each cycle of an auxiliary variable make all values equal to the value that is the majority among the contacts, this cannot decrease the number of satisfied equations by the very definition of an amplifier. Note that a cycle of 4 nodes is an amplifier for these nodes.

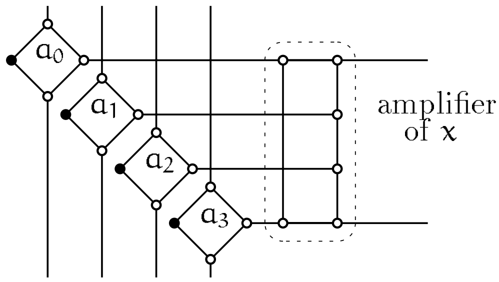


Figure 5: Part of the gadget replacing an equation with 3 variables. The other two variables also have their cycles of 6 variables. Empty circles indicate variables, solid circles indicate equations with just 1 variable, edges indicate equations. We can add variable z to the original equation to eliminate the equations with 1 variable only.

(ii) Let α be the common value of the variables in the amplifier of variable x. Consider a cycle of variable x in which not all values are equal, and let us use the above notation $x_0, \ldots, x_5$. Suppose that we have δ edges between α and $\neg\alpha$ values on the cycle, β many $\neg\alpha$ values among $x_0, \ldots, x_3$ and γ many $\neg\alpha$ values among x_4, x_5, i.e. adjacent to the amplifier. If $\delta + \gamma \geq \beta$, we convert $\neg\alpha$ values to α without decreasing the number of satisfied equations. As $\beta \leq 4$ and δ equals 2 or 4, we are done if $\delta \neq 2$, or if $\beta - \gamma \leq 2$. Moreover, if $\beta = 4$, we can convert each α value to $\neg\alpha$ and increase the number of satisfying equations and that also normalizes the cycle. Thus it remains to normalize the case when $\delta = 2$, $\beta = 3$, $\gamma = 0$. One can see that this is not possible.

(iii) Now each cycle is consistent. We normalize the values in the cycles of auxiliary variables as in the normalization of τ, to maximize the number of satisfied equations.

(iv) Suppose that a cycle of an original variable x is consistent, but with value $\neg\alpha$ while its amplifier is consistent with value α. We convert this cycle to α, and renormalize the auxiliary variables. We gain 2 equations that connect the cycle with the amplifier of x, and we loose at most 2 equations (among 16 equations if $S(x, \ldots)$ we satisfy 10 or 12, so we could drop by at most 2).

The solution equivalence is simple: the value of x is given to all replica in its amplifiers, the other variables in the new instance are set with some default and then we normalize this solution.

It remains to calculate the value equivalence.

We started with E that had n equations and $3n$ variable occurrences. In reduction R', we add z to each equations, which makes $4n$ variable occurrences.

For each equation, we made 16 equations, of which 12 are satisfied if the equation was satisfied, and otherwise only 10.

In these 16 equations, we have 16 occurrences of auxiliary variables that are connected into simple cycles, thus creating 16 satisfied equations.

An occurrence of an original variable has a cycle with 6 equations, 2 equations connecting it with its amplifier. A wheel amplifier has 10 equations for each contact, so this occurrence needs 20. The total number of equations for an occurrence is $6+2+20=28$.

In Eq-reduction R, for each original equation we created $16+16+4+3\times28+16=116$ equations. In a normalized solution that satisfies the original equation we satisfy $12+16+3\times28=112$, and otherwise we satisfy two equations less. Thus the value equivalence is $r(k,n)=110n+2kn$.

In Eq-reduction R' we have need to add 28 satisfied equations, thus we produced $144n^2$ equations and the value equivalence is $r(k,n)=138n^2+2kn$. ❒
We conclude that $(n/2,n)$ gap property of E2-LIN-2 implies $(112/116n, 111/116n)$ gap property of E3-OCC-2-LIN-2. and $(140/144n, 139/144n)$ gap property of E3-OCC-2-LIN-2.

By using the same approach as in [BK99], we can extend the result for E3-OCC-E2-LIN-2 to an identical result for 3-MAX-CUT. Thus we can formulate this conclusion as follows.

Theorem 11. *For every* $\varepsilon\in(0,1/139)$, *it is NP hard to approximate E3-OCC-E2-LIN-2 and E3-MAX-CUT to within a factor* $140/139-\varepsilon$ *and to approximate E3-OCC-2-LIN-2 to within a factor of* $112/111-\varepsilon$.

5.2 How to Modify Eq-Reductions

We will use a modification of Eq-reduction from E3-LIN-2 to E3-OCC-2-LIN-2. Suppose that we have an instance X of E3-LIN-2 with $2n$ equations. We form a system of constraints $f(X)$ where each equation of X is replaced with some L constraints, and in normalized solutions of $f(X)$ the satisfied equations of X correspond to a group of L constraints where all but 4 are satisfied, and unsatisfied equation corresponds to a similar group where all but 6 are satisfied. If we ignore εn terms, it is hard to tell if we can satisfy only n or up to $2n$ equations of X, this maps into a questions if we can satisfy only $Ln-6n+Ln-4n$ constrainst or up to $Ln-4n+Ln-4n$, which gives the hardness of the ratio $(2L-8)/(2L-10)=(L-4)/(L-5)$. In terms of the theorem schema from the introduction we have $K=L-4$.

We split the construction of our group of L constrains into several parts. In the reduction of E3-LIN-2 to E3-OCC-E2-LIN-2 such a group contained the following building blocks:

- 4 groups of equations that involve a single a, each group had $j=8$ equations (the cycle of 4 and the incident equations);
- 3 x-cycles and their connections to their amplifier, hence 3 times $k=8$ equations (6 on a cycle, 2 to connect to the amplifier);

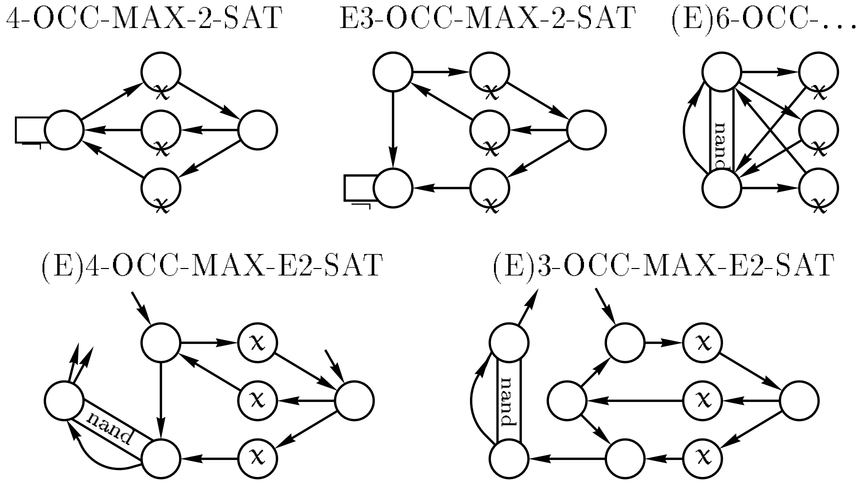


Figure 6: Replacement systems for equations with a fixed auxiliary variable.

- pieces of 3-regular amplifiers that together have 6 contacts, and the amplifiers has $l = 10$ equalites per contact.

This gives $L = 4j + 3k + 6l = 116$ and $K = 112$. We will show new versions of these building blocks of the reduction to find the respective values of j, k and l.

5.2.1 Equations with a Fixed Auxiliary Variable

Equation with a fixed auxiliary variable form a system S like that:

$$\begin{cases} x_0 + a &= 1 \\ x_1 + a &= 0 \\ x_2 + a &= 0 \\ b + a &= 0 \end{cases} \quad \Longleftrightarrow \quad \begin{cases} a &= \hat{x}_0 \\ a &= \hat{x}_1 \\ a &= \hat{x}_2 \\ a &= 0 \end{cases}$$

The universal form on the right can always be obtained if we replace some x's with their nagations. Because we choose the value of a, we view this system as a function $f_S : \{0,1\}^3 \to Z$ that returns maximum number of satisfied equations in S. Because we want to replace this system with a larger one, it is convenient to decrease the value of this function by the number of equations in the system. Thus $f_S(0,0,0) = 0, f_S(0,0,1) = -1, f_S(0,1,1) = -2$ and $f_S(1,1,1) = -1$ (note that f is symmetric).

We will construct a system T in which each x occurs twice and which otherwise satisfies limitations of a particular variation of MAX-2-SAT. As we will see, a requirement that the set of constrains should be regular increases the size of T. The next figure represents these variations of T as follows. Circles with an x inside indicate variables x_0, x_1 and x_2; empty circles indicate replicated copies of a, arrows indicate implications, [nand] indicates a clause of the form $\neg u \vee \neg v$ and [¬] indicates a clause of the form $\neg u$. When in a system some variables occur less than allowed nummber of times, we can add implications between such variables; incomplete arrows indicate where we can do it.

When we allow clauses of length one, we obtained j equal to 7 for 4-OCC-MAX-2-SAT, and to 8 for E3-OCC-MAX-2-SAT. Otherwise, we obtained 11 (12) for (E)3-OCC-MAX-E2-SAT, 9 (11) for (E)4-OCC-MAX-E2-SAT and 8 (9) for (E)6-OCC-MAX-E2-SAT.

5.2.2 Equations of an Amplifier

For amplifiers we did not notice as yet any size savings if we allow shorter clauses or a below-maximal number of occurences. Therefore we will skip E's when we discuss various versions of MAX-2-SAT.

For 3-OCC-MAX-2-SAT we adapt 5-regular amplifiers from Theorem 3. We replace such a node of degree 5 with a system of 20 implications, and an equation with a system of 4 implications; thus we replace 1.8 node and 4 equations with 52 implications, so we have $l = 52$.

For 3-OCC-MAX-2-SAT we adapt 7-regular amplifiers from Theorem 3, i.e. strong expanders. These amplifiers for each contact have one node and 3 edges. We replace such a node with a system of 21 implications and each equality edge with 2 implications, so we get $l = 27$.

For 6-OCC-MAX-2-SAT we adapt 9-regular "very strong" expanders. According to Bolobas [Bo88] a random 9-regular graph has isoperimetric number larger then 2; therefore if we have a minority among contacts of size k, the min-cut between this minority and the majority of contacts is at least $2k$. As a result we can use this contact node twice. A contact with 9 connections inside the expander and 2 connections outside can be replaced with a system of 22 implications plus 9 implications for the 9 adjacent expander edges. Thus we get $l = 31/2$.

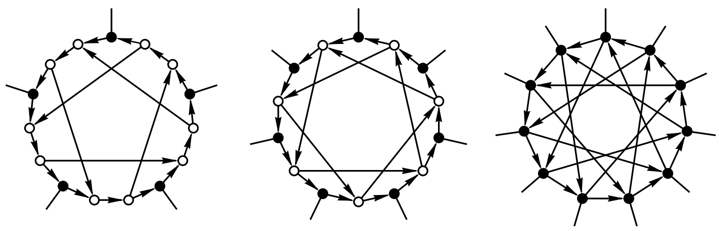


Figure 7: Replacements of of nodes of degree 5, 7 and 11 by 3-, 4- and 6-regular graphs of implications.

5.2.3 Equations of an x-Cycle

We connect copies of x as in the Fig. 8. The adjacent inequalities are attached as follows: x_0 and x_3 are connected with the amplifier (by equality gadgets), and x_1, x_2, x_4 and x_5 with the auxiliary variables. In case of 6-regular system, we have only 5 variables, and instead of variable x_3 being connected with a respective contact of the amplifier, say $\breve{x}_3$, we have implications $x_4 \to \breve{x}_3 \to x_5$.

To compute l, we count the number of implications inside the gadgets and add implications of equality gadgets that connect them with the amplifiers. For 3-regular systems, equality (equation with two variables) gadget is a cycle of 4 implications, and in other cases this is a pair of implications. One can see that for 3-OCC-, 4-OCC- and 6-OCC- problems we got $l = 36, 22, 13$.

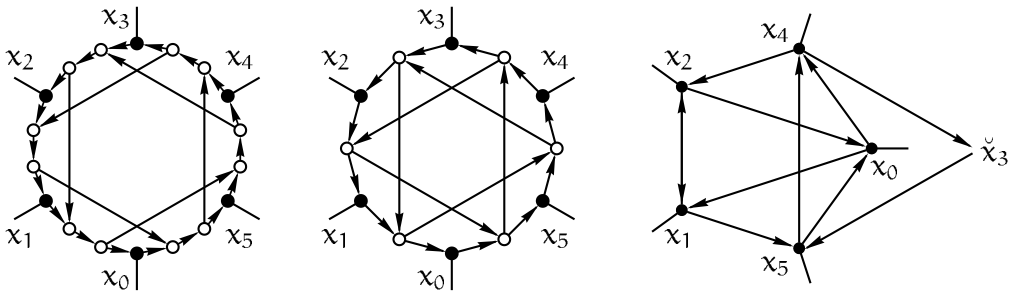


Figure 8: Cycle gadgets implemented as 3-, 4- and 6-regular graphs of implications.

5.2.4 Summary of the MAX-2SAT Results.

We summarize now the results on the small occurrence instances of MAX-2SAT.

Theorem 12. *For every* $\varepsilon \in (0, 1/(k-1)$*, it is NP hard to approximate a version of MAX-2-SAT to within a factor* $k/(k-1) - \varepsilon$*, where for*

- *E3-OCC-MAX-2-SAT we have* $k = 464 = 4 \times 12 + 3 \times 36 + 6 \times 52 - 4$;
- *E3-OCC-MAX-E2-SAT we have* $k = 460 = 4 \times 11 + 3 \times 36 + 6 \times 52 - 4$;
- *E4-OCC-MAX-E2-SAT we have* $k = 268 = 4 \times 11 + 3 \times 22 + 6 \times 27 - 4$;
- *E4-OCC-MAX-2-SAT we have* $k = 262 = 4 \times 9.5 + 3 \times 22 + 6 \times 27 - 4$;
- *4-OCC-MAX-2-SAT we have* $k = 252 = 4 \times 7 + 3 \times 22 + 6 \times 27 - 4$;
- *E6-OCC-MAX-E2-SAT we have* $k = 164 = 4 \times 9 + 3 \times 13 + 6 \times 31/2 - 4$;
- *6-OCC-MAX-E2-SAT we have* $k = 160 = 4 \times 8 + 3 \times 13 + 6 \times 31/2 - 4$;

□

6 Open Problems

Our constructions have two parts: gadgets that replace equations of E3-LIN-2, and amplifiers. It would be very interesting to investigate how the theory of optimal gadgets can be applied here. Our impression is that because of the degree bounds, we have quite large gadgets, e.g. for 4-OCC-MAX-2-SAT we have gadgets with more then 90 clauses. Exhaustive search for a better gadget does not have to be feasible, but some research is clearly needed towards that end.

The amplifiers are not fully understood either. Moreover, systems of implications and independent set problems should have their own versions of amplifier properties and a separate probabilistic analysis.

How about the explicit inapproximability bounds for *very small* occurrence instances of MAX-3SAT and MAX-4SAT? Very recently, [BKS03] established the first inapproximability results on E4-OCC-MAX-E3-SAT and E6-OCC-MAX-E4-SAT. It would be very interesting to shed some more light on the approximation hardness of such instances.

References

[AL97] S. Arora and C. Lund, *Hardness of Approxiamtions, in Approximatiojn Algorithms for NP-Hard Problems*, D. S. Hochbaum (ed.), PWS Publishing, Boston, 1997, 399–446.

[BK99] P. Berman and M. Karpinski, *On Some Tighter Inapproximability Results*, Proc. 26th ICALP, LNCS 1644, Springer, 1999, 200-209.

[BK01] P. Berman and M. Karpinski, *Efficient Amplifiers and Bounded Degree Optimization*, ECCC TR01-53 (2001).

[BKS03] P. Berman, M. Karpinski and A. D. Scott, *Approximation Hardness and Satisfiability of Bounded Occurrence Instances of SAT*, Manuscript, 2003.

[Bo88] B. Bolobás, *The Isoperimetric Number of Random Graphs*, Europ. J. Combinatorics **9** (1988), 241–244.

[CC02] M. Chlebík and J. Chlebíková, *Approximation Hardness for Small Occurrence Instances of NP-hard Problems*, ECCC TR02-073 (2002).

[H97] J. Håstad, *Some Optimal Inapproximability Results*, Proc. 29th ACM STOC, 1-10, 1997.

[P94] C. H. Papadimitriou, *Computational Complexity*, Addison-Wesley, New York, 1994, 315-319.

[PY91] C. H. Papadimitriou and M. Yannakakis, *Optimization, Approximation and Complexity Classes*, JCSS **43**, 1991, 425–440.