Strong Refutation Heuristics for Random k-SAT

Amin Coja-Oghlan[1], Andreas Goerdt[2], and André Lanka[2]

[1] Humboldt-Universität zu Berlin, Institut für Informatik
Unter den Linden 6, 10099 Berlin, Germany
coja@informatik.hu-berlin.de
[2] Technische Universität Chemnitz, Fakultät für Informatik
Straße der Nationen 62, 09107 Chemnitz, Germany
{goerdt, lanka}@informatik.tu-chemnitz.de

Abstract. A simple first moment argument shows that in a randomly chosen k-SAT formula with m clauses over n boolean variables, the fraction of satisfiable clauses is at most $1-2^{-k}+o(1)$ as $m/n \to \infty$ almost surely. In this paper, we deal with the corresponding algorithmic *strong refutation problem*: given a random k-SAT formula, can we find a *certificate* that the fraction of satisfiable clauses is at most $1-2^{-k}+o(1)$ in polynomial time? We present heuristics based on spectral techniques that in the case $k=3$, $m \geq \ln(n)^6 n^{3/2}$ and in the case $k=4$, $m \geq Cn^2$ find such certificates almost surely. Our methods also apply to a variety of further problems such as hypergraph coloring.

1 Introduction and Results

The *k-SAT problem* – given a set of k-clauses, i.e. disjunctions of k literals over a set of boolean variables, decide whether there exists an assignment of the variables that satisfies all clauses – is *the* generic NP-complete problem. In addition to the decision version, the optimization version MAX k-SAT – given a set of k-clauses, find an assignment that satisfies the maximum number of clauses – is of fundamental interest as well. However, Håstad [17] has shown that there is no polynomial time algorithm that approximates MAX k-SAT within a factor better than $1-2^{-k}$, unless $P=NP$. Hence, it is NP-hard to distinguish between instances of k-SAT in which a $(1-\varepsilon)$-fraction of the clauses can be satisfied, and instances in which every truth assignment satisfies at most a $(1-2^{-k}+\varepsilon)$-share of the clauses for any $\varepsilon>0$. Indeed, Håstad's NP-hardness result is best possible, as by picking a random assignment, we can satisfy a $(1-2^{-k})$-fraction of the clauses in polynomial time.

These hardness results motivate the study of *heuristics* for k-SAT or MAX k-SAT that are successful at least on a large class of instances. From this point of view, the satisfiability problem is interesting in two respects. First, one could ask for heuristics for *finding* a satisfying assignment (in the case of k-SAT) or a "good" assignment (in the case of MAX k-SAT). This problem has been studied e.g. by Flaxman [10], who has shown that in a rather general model of random satisfiable formulas a satisfying assignment can be found in polynomial time almost surely (cf. also [21] for an extension to semirandom formulas). Secondly, one can ask for heuristics that can *refute* a k-SAT instance, i.e. find a certificate that no satisfying assignment exists; of course, in the worst-case this problem is coNP-complete. In this paper, we deal with the second problem. More precisely, we present *strong refutation heuristics*, i.e. heuristics that certify that no assignment satisfying considerably more than the trivial $(1-2^{-k})$-fraction of the clauses exists. One motivation for studying this problem is the relationship between the existence of strong refutation heuristics and approximation complexity pointed out by Feige [7].

In order to analyze a heuristic rigorously, we need to specify on which type of instances the heuristic is supposed to work properly. In this paper, we consider a standard model of *random* instances of MAX k-SAT. Let $V=\{x_1,\ldots,x_n\}$ be a set of n boolean variables. Then, there are $(2n)^k$ possible k-clauses over the variables V. If $0<p<1$, then we let $\mathrm{Form}_{n,k,p}$ be a random set of k-clauses obtained by including each of the $(2n)^k$ possible clauses with probability p independently. Hence, the expected number of clauses in $\mathrm{Form}_{n,k,p}$ is

$m = (2n)^k p$. (Thus, in this paper, clauses are order k-tuples, and we allow for multiple occurrences of literals in a clause. Several slightly different models exist, but the differences are only of technical relevance.)

The combinatorial structure of random k-SAT formulas has attracted considerable attention. Friedgut [11] has shown that $\mathrm{Form}_{n,k,p}$ exhibits a *sharp threshold behavior*: there exist numbers $c_k = c_k(n)$ such that $\mathrm{Form}_{n,k,p}$ is satisfiable almost surely if $m < (1-\varepsilon)c_k n$, whereas $\mathrm{Form}_{n,k,p}$ is unsatisfiable almost surely if $m > (1+\varepsilon)c_k n$. The asymptotic behavior of c_k as $k \to \infty$ has been determined by Achlioptas and Peres [2]. Moreover, a simple first moment argument shows that the maximum number of clauses of $\mathrm{Form}_{n,k,p}$ that can be satisfied by any assignment is at most $(1 - 2^k + o(1))m$ as $m/n \to \infty$. More precise results have been obtained by Achlioptas, Naor, and Peres [1].

With respect to proof complexity, various types of resolution proofs for the non-existence of satisfying assignments in $\mathrm{Form}_{n,k,p}$ have been investigated. Ben-Sasson [4] has shown that tree-like resolution proofs to refute $\mathrm{Form}_{n,k,p}$ almost surely have size $\exp(\Omega(n/\Delta^{1/(k-2)+\varepsilon}))$, where $\Delta = n^{k-1}p$ and $0 < \varepsilon < 1/2$ is an arbitrary constant. Hence, tree-like resolution proofs are of exponential length even if the expected number of clauses is $n^{k-1-\varepsilon}$ (i.e. $p = n^{-\varepsilon-k/2}$). Furthermore, [4, Theorem 2.24] shows that general resolution proofs for the nonexistence of satisfying assignments of $\mathrm{Form}_{n,k,p}$ almost surely have super polynomial size if $p \leq n^{-k/2-\delta}$ ($\delta > 0$ constant).

Goerdt and Krivelevich [16] have suggested a heuristic that uses spectral techniques for refuting $\mathrm{Form}_{n,4,p}$ with $p = \ln(n)^7 n^{-2}$ (i.e. the expected number of clauses is $m = \ln(n)^7 n^2$). No efficient resolution-based refutation heuristic is known for this range of p; in fact, tree-like resolution proofs are of exponential length by the aforementioned results. Removing the polylogarithmic factor, Feige and Ofek [8] and (independently) Coja-Oghlan, Goerdt, Lanka, and Schädlich [6] have shown that spectral techniques can be used to refute $\mathrm{Form}_{n,4,p}$ if $p \geq Cn^{-2}$ for a sufficiently large constant $C > 0$. Moreover, Feige and Ofek [9] have shown that a heuristic that combines spectral techniques with extracting and refuting a XOR formula from $\mathrm{Form}_{n,3,p}$ can refute $\mathrm{Form}_{n,3,p}$ for $p \geq Cn^{-3/2}$ (i.e. $m = Cn^{3/2}$). This result improves on previous work by Friedman and Goerdt [12], and Goerdt and Lanka [15]. We emphasize that in all of the above cases, the values of p to which the refutation heuristics apply exceed the threshold when $\mathrm{Form}_{n,k,p}$ actually becomes unsatisfiable almost surely by at least a factor of $n^{(k-2)/2}$.

The new aspect in the present paper is that we deal with *strong* refutation heuristics. That is, our aim are heuristics that on input $\mathrm{Form}_{n,k,p}$ almost surely certify that not more than a $(1 - 2^{-k} + \varepsilon)$-fraction of the clauses can be satisfied, for any $\varepsilon > 0$. This aspect has not (at least not explicitly) been studied in the aforementioned references. For instance, resolution proofs cannot provide strong refutation. Moreover, the spectral heuristics studied so far [8, 9, 6, 15, 12, 16] only certify that every assignment leaves a $o(1)$-fraction of the clauses unsatisfied. With respect to MAX 3-SAT, we have the following result.

Theorem 1. *Suppose that $p \geq \ln(n)^6 n^{-3/2}$. Let $\varepsilon > 0$ be an arbitrarily small constant. There is a polynomial time algorithm 3-`Refute` that satisfies the following conditions.*

- Correctness: *For any MAX 3-SAT instance φ, the output of 3-`Refute`(φ) is an upper bound on the number of satisfiable clauses.*
- Completeness: *If $\varphi = \mathrm{Form}_{n,3,p}$, then almost surely 3-`Refute`$(\varphi) \leq (7+\varepsilon)n^3 p$.*

Since the number of clauses of $\mathrm{Form}_{n,3,p}$ is $(8 + o(1))n^3 p$ almost surely, `3-Refute` does indeed certify almost surely that not more than a $\frac{7}{8} + \varepsilon$ fraction of the clauses can be satisfied by any assignment. Note that the value of p required for Theorem 1 is by a factor of $\ln(n)^6$ larger than that required by the heuristic of Feige and Ofek [9] (which does not provide strong refutation). Moreover, the following result addresses MAX 4-SAT.

Theorem 2. *Suppose that $p \geq c_0 n^{-2}$ for a sufficiently large constant $c_0 > 0$. There is a polynomial time algorithm 4-`Refute` that satisfies the following conditions.*

– Correctness: *For any MAX* 4*-SAT instance* φ*, the output of* 4-`Refute`(φ) *is an upper bound on the number of satisfiable clauses.*
– Completeness: *If* $\varphi = \mathrm{Form}_{n,4,p}$*, then almost surely* 4-`Refute`$(\varphi) \leq 15n^4p + c_1n^3\sqrt{p}$*, where* $c_1 > 0$ *is a constant.*

4-`Refute` almost surely provides a certificate that not more than a $\frac{15}{16} + O(\frac{1}{n\sqrt{p}})$ fraction of the clauses can be satisfied. The second order term $O(\frac{1}{n\sqrt{p}})$ gets arbitrarily small as n^2p grows. Theorem 2 applies to the same range of p as the best previously known refutation heuristics [6, 8] for 4-SAT, but provides strong refutation.

The algorithms for Theorems 1 and 2 build on and extend the techniques proposed in [6, 14]. For instance, 4-`Refute` constructs several graphs from the input formula $\varphi = \mathrm{Form}_{n,4,p}$. To each of these graphs, 4-`Refute` applies a subroutine that tries to certify that the graph has "low discrepancy"; i.e. every set of vertices spans approximately the expected number of edges. This subroutine in turn relies on computing the eigenvalues of a certain auxiliary matrix. Finally, if all graphs have passed the discrepancy check, then we conclude that the input formula φ does not admit an assignment that satisfies more than $15n^4p + c_1n^3\sqrt{p}$ clauses. The MAX 3-SAT algorithm for Theorem 1 proceeds similarly, but is a bit more involved. Though in contrast to [6, 14] we obtain strong refutation heuristics, the algorithms and the proofs in the present paper are considerably simpler.

The techniques that the algorithms 3-`Refute` and 4-`Refute` rely on yield heuristics for a variety of further hard computational problems, e.g. for hypergraph problems. Recall that a *k-uniform hypergraph* H consists of a set $V(H)$ of vertices and a set $E(H)$ of edges. The edges are subsets of $V(H)$ of cardinality k. An *independent set* in H is a set $S \subset V(H)$ such that there is no edge $e \in E(H)$ with $e \subset S$. The *independence number* $\alpha(H)$ is the number of vertices in a maximum independent set. Moreover, H is called κ-colorable, if there exists κ independent sets $S_1, \ldots, S_\kappa$ in H such that $S_1 \cup \cdots \cup S_\kappa = V(H)$. The *chromatic number* $\chi(H)$ is the least integer $\kappa \geq 1$ such that H is κ-colorable.

In analogy with the $\mathrm{Form}_{n,k,p}$ model of random k-SAT instances, there is the $H_{n,k,p}$-model of random k-uniform hypergraphs: the vertex set of $H_{n,k,p}$ is $V = \{1, \ldots, n\}$, and each of the $\binom{n}{k}$ possible edges is present with probability $0 < p < 1$ independently. Krivelevich and Sudakov [20] have solved the combinatorial problem of determining the probable value of the independence number and of the chromatic number of random hypergraphs. The following two theorems deal with the *algorithmic* problem of refuting that a 3-uniform hypergraph has a large independent set, or that a 4-uniform hypergraph is κ-colorable.

Theorem 3. *Let* $\varepsilon > 0$ *be arbitrarily small but fixed. Suppose that* $p = f/n^{3/2}$*, where* $\ln^6 n \leq f = o(n^{1/2})$*. There is a polynomial time algorithm* 3-`RefuteInd` *that satisfies the following conditions.*

– Correctness: *If* H *is a* 3*-uniform hypergraph, then* 3-`RefuteInd`(H) *either outputs "*α *is small" or "fail". If* 3-`RefuteInd`(H) *answers "*α *is small", then* $\alpha(H) < \varepsilon n$*.*
– Completeness: *On input* $H = H_{n,3,p}$*,* 3-`RefuteInd`(H) *outputs "*α *is small" almost surely.*

Theorem 4. *Let* $\kappa \geq 2$ *be an integer. Suppose that* $p \geq c_0\kappa^4n^{-2}$ *for some sufficiently large constant* $c_0 > 0$*. There is a polynomial time algorithm* 4-`RefuteCol` *that satisfies the following conditions.*

– Correctness: *If* H *is a* 4*-uniform hypergraph, then* 4-`RefuteCol`(H) *either outputs "not* κ*-colorable" or "fail". If* 4-`RefuteCol`(H) *answers "not* κ*-colorable", then* $\chi(H) > \kappa$*.*
– Completeness: *On input* $H = H_{n,4,p}$*,* 4-`RefuteCol`(H) *outputs "not* κ*-colorable" almost surely.*

Organization of the paper. We start with the algorithm 4-`Refute` for Theorem 2 in Section 2. 4-`Refute` is a bit simpler than the algorithm 3-`Refute` for Theorem 1, which comes in Section 3. We sketch the heuristics for Theorems 3 and 4 in Section 4.

2 Random MAX 4-SAT

In Section 2.2 we peresent the heuristic for Theorems 2. The main tool is a procedure for certifying that a random bipartite graph is of low discrepancy. This procedure is the content of Section 2.1.

2.1 Discrepancy in Random Bipartite Graphs

Throughout, we let $V_1 = \{v_1, \ldots, v_n\}$ and $V_2 = \{w_1, \ldots, w_n\}$ be two disjoint sets consisting of n labeled vertices each. We consider bipartite graphs G with bipartition (V_1, V_2), i.e. the vertex set of G is $V_1 \cup V_2$, and all edges of G have one endpoint in V_1, and one in V_2. If $S_1 \subset V_1$ and $S_2 \subset V_2$, then we let $E_G(S_1, S_2)$ denote the set of edges in G that connect a vertex in S_1 with a vertex in S_2. Furthermore, $B_{n,p}$ denotes a random bipartite graph obtained by including each possible edge $\{v_i, w_j\}$ with probability p independently. The aim in this section is to prove the following proposition.

Proposition 5. *Suppose that $np \geq c_0$ for some sufficiently large constant $c_0 > 0$. There is a polynomial time algorithm `BipDisc` and a constant $c_1 > 0$ such that the following two conditions hold.*

1. *Let G be a bipartite graph with bipartition (V_1, V_2). On input G, `BipDisc` either outputs "low discrepancy" or "fail". If `BipDisc`(G) outputs "low discrepancy", then for any two sets $S_i \subset V_i$, $i = 1, 2$, we have*
$$||S_1||S_2|p - |E_B(S_1, S_2)|| \leq c_1\sqrt{|S_1||S_2|np} + n\exp(-np/c_1). \quad (1)$$
2. *`BipDisc`$(B_{n,p})$ outputs "low discrepancy" almost surely.*

If $|S_1|, |S_2| = \Omega(n)$, then Eq. (1) entails that the number $|E_G(S_1, S_2)|$ of edges from S_1 to S_2 in G deviates from its expectation $|S_1||S_2|p$ "not too much". The crucial point is that `BipDisc` certifies that Eq. (1) holds for *all* sets S_1, S_2.

`BipDisc` is based on computing the eigenvalues of a certain auxiliary matrix. Given a graph B with bipartition (V_1, V_2), we let $A = A(B) = (a_{ij})_{i,j=1,\ldots,n}$ be the matrix with entries $a_{ij} = 1$ if $\{v_i, w_j\} \in E(B)$, and $a_{ij} = 0$ if $\{v_i, w_j\} \notin E(B)$. Let J denote an $n \times n$ matrix with all entries equal to 1. Then, we let $M = M(B) = pJ - A(B)$. Furthermore, let $\|M\| = \sup\{\|M\xi\| : \xi \in \mathbf{R}^n,\ \|\xi\| = 1\}$ denote the norm of M. On input B, $\|M\|$ can be computed in polynomial time up to an arbitrarily small additive error (e.g. by computing the largest eigenvalue of the positive semidefinite matrix $M^T M$). The next lemma shows what $\|M\|$ has to do with discrepancy certification.

Lemma 6. *Let B be a graph with bipartition (V_1, V_2). Then, for any two sets $S_i \subset V_i$, $i = 1, 2$, the inequality $|\,|E_B(S_1, S_2)| - |S_1||S_2|p\,| \leq \sqrt{|S_1||S_2|} \cdot \|M(B)\|$ holds.*

Sketch of proof. Let ξ_i be the characteristic vector of S_i, i.e. the j'th entry of ξ_1 (resp. ξ_2) is 1 if $v_j \in S_1$ (resp. $w_j \in S_2$), and 0 otherwise. Then $\|\xi_i\| = \sqrt{|S_i|}$. Hence, $|\langle M\xi_2, \xi_1\rangle| \leq \sqrt{|S_1||S_2|}\|M\|$.Moreover, a direct computation shows that $\langle M\xi_2, \xi_1\rangle = |S_1||S_2|p - |E_B(S_1, S_2)|$. □

In the case $np \geq \ln(n)^7/n$, one can show that $\|M\| \leq O(\sqrt{np})$ almost surely (via the "trace method" from [13]). Hence, in this case, by Lemma 6 we could certify that (1) holds almost surely just by computing $\|M(B_{n,p})\|$. In the case $np = O(1)$, however, we almost surely have that $\|M(B_{n,p})\| = \Theta(\ln n)$, i.e. $\|M(B_{n,p})\|$ is much too large to give the bound (1). The reason is that in this case, there will be vertices of degree up to $\Theta(\ln n)$ in $B = B_{n,p}$ (cf. [19] for a more detailed discussion). Following an idea of Alon and Kahale [3], we avoid this problem by removing all edges that are incident with vertices whose degree is too high (at least $10np$, say). This leads to the following algorithm.

Algorithm 7. `BipDisc`(G)
Input: A bipartite graph $G = (V_1, V_2, E)$. *Output:* Either "low discrepancy" or "fail".

1. If the number of vertices in G that have degree $> 10np$ is $> n\exp(-c_2np)$, then output "fail" and halt. Here $c_2 > 0$ is a sufficiently small constant (cf. Lemma 8 below).
2. If the number of edges in G that are incident with vertices of degree $> 10np$ is larger than $c_3n^2p\exp(-c_2np)$, where $c_3 > 0$ is a sufficiently large constant, then halt with output "fail".
3. Let G' be the graph obtained from G by deleting all edges that are incident with vertices of degree $> 10np$. Let $M = M(G')$. If $\|M\| > c_4\sqrt{np}$ for a certain constant c_4, then output "fail" and halt.
4. Output "G has low discrepancy".

The analysis of `BipDisc` is based on two lemmas.

Lemma 8. *There are constants $c_2, c_3 > 0$ such that whp. $B = B_{n,p}$ has the following properties.*

1. *Let S be the set of all vertices that have degree $> 10np$ in B. Then $|S| \leq n\exp(-c_2np)$.*
2. *The number of edges in B that are incident with at least one vertex in S is $\leq c_3n^2p\exp(-c_2np)$.*

Lemma 9. *There is a constant $c_4 > 0$ such that whp. the random bipartite graph $B = B_{n,p}$ enjoys the following property. Let B' be the graph obtained from B by deleting all edges that are incident with vertices of degree $> 10np$ in B. Then, $\|M(B')\| \leq c_4\sqrt{np}$.*

Lemma 8 follows from a standard computation. The proof of Lemma 9 is based on estimates on the eigenvalues of random matrices from [3] (cf. Appendix A).

Proof of Proposition 5. Let $G = B_{n,p}$, and let $S_i \subset V_i$ for $i = 1, 2$. Moreover, let S be the set of vertices of degree $> 10np$ in G. Suppose that `BipDisc`(G) answers "low discrepancy". Then, by Lemma 6,

$$|\,E_G(S_1 \setminus S, S_2 \setminus S)| - |S_1 \setminus S||S_2 \setminus S|p\,| \leq c_4\sqrt{|S_1||S_2|np}.$$

Moreover, because of Step 2 of `BipDisc`, we have $|E_G(S_1, S_2)| - |E_G(S_1 \setminus S, S_2 \setminus S)| \leq c_3n^2p\exp(-c_2np)$. Finally, $|S_1||S_2|p - |S_1 \setminus S||S_2 \setminus S|p \leq 3np|S| \leq n\exp(-c_2np/2)$,as otherwise Step 1 would have failed. Thus, (1) holds for S_1, S_2. Finally, Lemmas 8 and 9 imply that `BipDisc`$(B_{n,p})$ outputs "low discrepancy" almost surely. □

2.2 The Refutation Heuristic for 4-SAT

Throughout this section, we let $V = \{x_1, \ldots, x_n\}$ be a set of n propositional variables. Moreover, we assume that $n^2p \geq c_0$ for a sufficiently large constant c_0.

Let φ be a set of 4-clauses over V. To employ the procedure `BipDisc` from Section 2.1, we construct 16 bipartite graphs $G^{(1)}, \ldots, G^{(16)}$ from φ. Each $G^{(i)}$ is a graph with bipartition (V_1, V_2), where $V_i = V \times V \times \{i\}$ (i.e. V_1, V_2 are disjoint copies of $V \times V$). Each graph $G^{(i)}$ corresponds to one of the 16 possible ways to place the negation signs in a 4-clause: in $G^{(i)}$, the edge $\{(x_{i_1}, x_{i_2}, 1), (x_{i_3}, x_{i_4}, 2)\}$ is present iff the clause $l_{i_1} \vee l_{i_2} \vee l_{i_3} \vee l_{i_4}$ is contained in φ, where l_{i_j} is either x_{i_j} or $\bar{x}_{i_j}$, according to the negation signs in Table 1. For instance, the edge $\{(x_{i_1}, x_{i_2}, 1), (x_{i_3}, x_{i_4}, 2)\}$ is in $G^{(7)}$ iff the clause $x_{i_1} \vee \bar{x}_{i_2} \vee \bar{x}_{i_3} \vee x_{i_4}$ occurs in φ. Thus, each clause of φ induces an edge in one of the graphs $G^{(i)}$, and each edge results from a unique clause. The algorithm for Theorem 2 is as follows.

Algorithm 10. `4-Refute`(φ)
Input: A set φ of 4-clauses over V. *Output:* An upper bound on the number of satisfiable clauses.

i	type	$A_i \subset V_1$	$B_i \subset V_2$
1	$x_1 \vee x_2 \vee x_3 \vee x_4$	$F \times F$	$F \times F$
2	$x_1 \vee x_2 \vee x_3 \vee \bar{x}_4$	$F \times F$	$F \times T$
3	$x_1 \vee x_2 \vee \bar{x}_3 \vee x_4$	$F \times F$	$T \times F$
4	$x_1 \vee x_2 \vee \bar{x}_3 \vee \bar{x}_4$	$F \times F$	$T \times T$
5	$x_1 \vee \bar{x}_2 \vee x_3 \vee x_4$	$F \times T$	$F \times F$
6	$x_1 \vee \bar{x}_2 \vee x_3 \vee \bar{x}_4$	$F \times T$	$F \times T$
7	$x_1 \vee \bar{x}_2 \vee \bar{x}_3 \vee x_4$	$F \times T$	$T \times F$
8	$x_1 \vee \bar{x}_2 \vee \bar{x}_3 \vee \bar{x}_4$	$F \times T$	$T \times T$

i	type	$A_i \subset V_1$	$B_i \subset V_2$
9	$\bar{x}_1 \vee x_2 \vee x_3 \vee x_4$	$T \times F$	$F \times F$
10	$\bar{x}_1 \vee x_2 \vee x_3 \vee \bar{x}_4$	$T \times F$	$F \times T$
11	$\bar{x}_1 \vee x_2 \vee \bar{x}_3 \vee x_4$	$T \times F$	$T \times F$
12	$\bar{x}_1 \vee x_2 \vee \bar{x}_3 \vee \bar{x}_4$	$T \times F$	$T \times T$
13	$\bar{x}_1 \vee \bar{x}_2 \vee x_3 \vee x_4$	$T \times T$	$F \times F$
14	$\bar{x}_1 \vee \bar{x}_2 \vee x_3 \vee \bar{x}_4$	$T \times T$	$F \times T$
15	$\bar{x}_1 \vee \bar{x}_2 \vee \bar{x}_3 \vee x_4$	$T \times T$	$T \times F$
16	$\bar{x}_1 \vee \bar{x}_2 \vee \bar{x}_3 \vee \bar{x}_4$	$T \times T$	$T \times T$

Table 1. Clause types and unsatisfied clauses in the case of 4-SAT.

1. If the number of clauses in φ is larger than $16n^4p + n^3\sqrt{p}$, then return the total number of clauses in φ as an upper bound and halt.
2. Compute the graphs $G^{(i)}$ for $i = 1, \ldots, 16$ and run $\texttt{BipDisc}(G^{(i)})$ for $i = 1, \ldots, 16$. If $\texttt{BipDisc}(G^{(i)})$ answers "fail" for at least one i, then return the total number of clauses in φ and halt.
3. Return $15n^4p + c_1 n^3\sqrt{p}$, where c_1 is a sufficiently large consant.

Let us first prove that 4-$\texttt{Refute}$ outputs an upper bound on the number of clauses that can be satisfied.

Lemma 11. *There is a constant $c_2 > 0$ such that the following holds. Let φ be a set of 4-clauses such that $\texttt{BipDisc}(G^{(i)})$ answers "low discrepancy" for all i. Then there is no assignment that satisfies more than $|\varphi| - n^4p + c_2 n^3\sqrt{p}$ clauses of φ.*

Proof. Consider an assignment that sets the variables $T \subset V$ to true, and $F = V \setminus T$ to false. We shall bound the number of edges in the graphs $G^{(i)}$ that correspond to unsatisfied clauses. Let $A_i \subset V_1$ and $B_i \subset V_2$ be the sets defined in Table 1 for $i = 1, \ldots, 16$. Then, in the graph $G^{(i)}$, the edges corresponding to unsatisfied clauses are precisely the A_i-B_i-edges. Thus, invoking Proposition 5, we have the following bound on the number of unsatisfied clauses:

$$\sum_{i=1}^{16} |E_{G_i}(A_i, B_i)| \geq \sum_{i=0}^{4} \binom{4}{i} \left(|F|^i |T|^{4-i} p - c_3 n^3\sqrt{p}\right) \geq (|F| + |T|)^4 p - c_2 n^3\sqrt{p} = n^4 p - c_2 n^3\sqrt{p},$$

where c_2, c_3 are suitable constants. □

Proof of Theorem 2. The correctness claimed in the theorem follows from Lemma 11. Since by Chernoff bounds (cf. [18, p. 26]) the total number of clauses in $\mathrm{Form}_{n,4,p}$ is at most $16n^4p + o(n^3\sqrt{p})$ almost surely, the completeness follows from Proposition 5. □

Remark 12. Though this issue is not addressed explicitly in that paper, a strong refutation heuristic could also be obtained from the techniques presented in [6]. However, the approach in the present paper has some advantages. First of all, the algorithm is much simpler both to execute and to analyze. Secondly, the bound on the number of satisfiable clauses that could be obtained using the techniques in [6] is not as precise as those obtained in Theorem 2. Indeed, the approach in [6] can only be used to show that the fraction of satisfiable clauses is $\leq \frac{15}{16} + \varepsilon$ for an arbitrarily small but fixed $\varepsilon > 0$. By contrast, the Theorem 2 gives a the bound $\frac{15}{16} + O(\frac{1}{n\sqrt{p}})$, where the second order term tends to 0 as n^2p grows.

3 Random MAX 3-SAT

While our refutation heuristic for 4-SAT is based on certifying that certain (bipartite) graphs are of low discrepancy, the heuristic for 3-SAT needs to certify that a couple of triple systems are of low discrepancy. In Section 3.1, we describe the procedure for certifying low discrepancy in triple systems. Then, in Section 3.2, we show how to employ this procedure in order to refute MAX 3-SAT strongly.

3.1 Discrepancy in Triple Systems

Let $V = \{x_1, \ldots, x_n\}$ be a fixed set of cardinality n. In this section, we consider *triple systems* over V, i.e. subsets $S \subset V \times V \times V$. If $V_1, V_2, V_3 \subset V$, then we let $(V_1, V_2, V_3) = (V_1, V_2, V_3)_S$ signify the set of triples $(v_1, v_2, v_3) \in S$ with $v_i \in V_i$ for $i = 1, 2, 3$. Let $\varepsilon > 0$ be a constant. We say that S *has low discrepancy with respect to* ε if the following holds for all $X \subseteq V$ with $\varepsilon n \leq |X| \leq (1-\varepsilon)n$: letting $Y = V \setminus X$ and $\alpha = |X|/n$, we have

$$\begin{aligned} |(X,X,X)| &= (1+o(1)) \cdot \alpha^3 \cdot |S|, \\ |(X,X,Y)|, |(X,Y,X)|, |(Y,X,X)| &= (1+o(1)) \cdot \alpha^2(1-\alpha) \cdot |S|, \\ |(X,Y,Y)|, |(Y,X,Y)|, |(Y,Y,X)| &= (1+o(1)) \cdot \alpha(1-\alpha)^2 \cdot |S|, \\ |(Y,Y,Y)| &= (1+o(1)) \cdot (1-\alpha)^3 \cdot |S|. \end{aligned}$$

For $0 < p < 1$, we obtain the random triple system $S_{n,p}$ by including each triple in V^3 with probability p independently. The aim of this section is to prove the following propostion.

Proposition 13. *For each $\varepsilon > 0$ there is a polynomial time algorithm $\texttt{TripleDisc}_\varepsilon$ that satisfies the following conditions.*

- *For each triple system $S \subset V^3$ the output of $\texttt{TripleDisc}_\varepsilon(S)$ is either "low discrepancy" or "fail". If the output is "low discrepancy", then S has low discrepancy w.r.t. ε.*
- *If $p \geq \ln(n)^6 n^{-3/2}$, then the output of $\texttt{TripleDisc}_\varepsilon(S_{n,p})$ is "low discrepancy" almost surely.*

To certify that the triple system $S \subset V^3$ is of low discrepancy, the algorithm `TripleDisc` constructs three *projection graphs* G_{ij}, $1 \leq i < j \leq 3$. The vertex set of G_{ij} is V, and the edge $\{x, y\}$ is present in G_{ij} iff there is a triple $(z_1, z_2, z_3) \in S$ with $x = z_i$ and $y = z_j$, or $x = z_j$ and $y = z_i$. Thus, if $S = S_{n,p}$, then the edge $\{x, y\}$ is present in G_{ij} with probability $p' \sim 2np$ independently of all other edges, so that G_{ij} is distributed as a binomial random graph $G_{n,p'}$.

We say that a graph $G = (V, E)$ has *low discrepancy w.r.t.* ε if for all $X \subset V$ of cardinality $\varepsilon n \leq |X| \leq (1-\varepsilon)n$ we have

$$||E_G(X)| - |X|^2 n^{-2}|E|| \leq \varepsilon|E| \text{ and } ||E_G(X, V \setminus X)| - 2|X|(n-|X|)n^{-2}|E|| \leq \varepsilon|E|,$$

where $E_G(X)$ is the set of edges in G with both endpoints in X, and $E_G(X, Y)$ is the set of edges in G with one endpoint in X and the other in Y. One ingredient to the algorithm `TripleDisc` for Proposition 13 is to certify that the graphs G_{ij} are of low discrepancy. The following lemma provides us with a polynomial time algorithm for this problem.

Lemma 14. *Let $\varepsilon > 0$. Suppose that $p' \geq 1/n^{1/2}$. There is a polynomial time algorithm $\mathcal{A}$ that satisfies the following conditions.*

– Correctness: *For any graph $G = (V, E)$, the output of $\mathcal{A}(G)$ is either "low discrepancy" or "fail". If the output is "low discrepancy", then G has low discrepancy w.r.t. ε.*
– Completeness: *If $G = G_{n,p'}$, then the output of $\mathcal{A}(G)$ is "low discrepancy" almost surely.*

The proof of Lemma 14 is based on the relationship between graph discrepancy and eigenvalues (cf. [5]) and results on the eigenvalues of random symmetric matrices [13].

In order to certify that the triple system S has low discrepancy, it is, however, *not* sufficient to check that the projection graphs G_{ij} are of low discrepancy. Therefore, in addition to the projection graphs, one could consider the *product graph* $G_\pi = (V \times V, E_\pi)$, which is defined as follows: an edge $\{(a_1, b_1), (a_2, b_2)\}$ is in E_π iff there exists a $z \in V$ such that there are two different triples $(a_1, a_2, z), (b_1, b_2, z) \in S$. Note that in contrast to the projection graphs G_{ij}, the product graph G_π is not distributed as a binomial random graph (the edges do not occur independently). If the projection graphs G_{ij} and the product graph G_π all have low discrepancy, then S is of low discrepancy as well.

However, for the values of p in Proposition 13, we do not know a direct way to derive bounds on the eigenvalues of the adjacency matrix of the product graph (e.g. it seems difficult to apply the methods in [3, 8, 14]). Therefore, instead of dealing with the product graph and its adjacency matrix, we consider the matrix $\mathbf{A} = \mathbf{A}(S, p)$ defined as follows. For $0 < p < 1$ and $b_1, b_2, z \in V$ we let $B_{b_1b_2z} = B_{b_1b_2z}(S, p) = -1$ if $(b_1, b_2, z) \in S$, and $B_{b_1b_2z} = B_{b_1b_2z}(S, p) = p/(1-p)$, otherwise. Then, the $n^2 \times n^2$-matrix $\mathbf{A} = \mathbf{A}(S, p) = (\mathbf{a}_{b_1c_1,b_2c_2})_{(b_1,c_1),(b_2,c_2)\in V^2}$ is given by

$$\mathbf{a}_{b_1c_1,b_2c_2} = \sum_{z\in V} (B_{b_1b_2z} \cdot B_{c_1c_2z} + B_{b_2b_1z} \cdot B_{c_2c_1z}) \text{ if } (b_1, b_2) \neq (c_1, c_2),$$

and $\mathbf{a}_{b_1c_1,b_2c_2} = 0$ if $(b_1, b_2) = (c_1, c_2)$. Since $\mathbf{A}$ is symmetric and real-valued, the matrix has n^2 real eigenvalues $\lambda_1 \geq \cdots \geq \lambda_{n^2}$. We let $\|\mathbf{A}\| = \max\{\lambda_1, -\lambda_{n^2}\}$ signify the norm of $\mathbf{A}$.

If $S \subset V^3$, $x \in V$, and $i \in \{1, 2, 3\}$, then the *degree of x in slot i* is $d_{x,i} = |\{(z_1, z_2, z_3) \in S : z_i = x\}|$. We say that S is *asymptotically regular* if $d_{x,i} = (1 + o(1))n^{-1}|S|$ for all x, i. Equipped with these definitions, we can state the following sufficient condition for S being of low discrepancy.

Lemma 15. *Let $f = pn^{3/2}$, and suppose that $\ln^6 n \leq f = o(n^{1/2})$. If S is a triple system that satisfies the following four conditions, then S is of low discrepancy w.r.t. $\varepsilon > 0$.*

1. *$s = |S| = f \cdot n^{3/2} \cdot (1 + o(1))$.*
2. *S is asymptotically regular.*
3. *The three projection graphs of S are of low discrepancy with respect to $\varepsilon > 0$.*
4. *We have $\|\mathbf{A}(S, p)\| \leq \ln^5 n \cdot f$.*

The proof can be found in Appendix B. As by Lemma 14 we can check in polynomial time whether the conditions in Lemma 15 hold, we obtain the following algorithm.

Algorithm 16. $\texttt{TripDisc}_\varepsilon(S)$
Input: A set $S \subset V^3$. *Output:* Either "low discrepancy" or "fail".

1. Check whether Conditions 1–4 in Lemma 15 hold.
2. If so, output "low discrepancy". If not, return "fail".

In order to prove Proposition 13, it remains to establish that the algorithm is complete. A standard application of Chernoff bounds (cf. [18, p. 26]) shows that the random triple system $S = S_{n,p}$ with p as in Proposition 13 satisfies Conditions 1–2 in Lemma 15 almost surely. Moreover, the third condition holds almost surely by Lemma 14. Thus, it suffices to show that Condition 4 holds almost surely. The rather technical proof of the following lemma is based on the trace method from [13] (cf. Appendix C).

Lemma 17. *Let $f \geq \ln(n)^6$, and let $p = fn^{-3/2}$. If $S = S_{n,p}$, then $\|\mathbf{A}\| = \|\mathbf{A}(S,p)\| \leq \ln^5 n \cdot f$ almost surely.*

3.2 The Refutation Heuristic for 3-SAT

i	type	$U_i \subset V \times V \times V$
1	$x_1 \vee x_2 \vee x_3$	$F \times F \times F$
2	$x_1 \vee x_2 \vee \bar{x}_3$	$F \times F \times T$
3	$x_1 \vee \bar{x}_2 \vee x_3$	$F \times T \times F$
4	$x_1 \vee \bar{x}_2 \vee \bar{x}_3$	$F \times T \times T$

i	type	$U_i \subset V \times V \times V$
5	$\bar{x}_1 \vee x_2 \vee x_3$	$T \times F \times F$
6	$\bar{x}_1 \vee x_2 \vee \bar{x}_3$	$T \times F \times T$
7	$\bar{x}_1 \vee \bar{x}_2 \vee x_3$	$T \times T \times F$
8	$\bar{x}_1 \vee \bar{x}_2 \vee \bar{x}_3$	$T \times T \times T$

Table 2. Clause types and unsatisfied clauses in the case of 3-SAT.

Let φ be a set of 3-clauses over the variable set $V = \{x_1, \ldots, x_n\}$. To apply the procedure `TripDisc` from Section 3.1, we construct 8 triple systems $S^{(1)}, \ldots, S^{(8)} \subset V^3$ from φ, each corresponding to one of the 8 possible ways to set the negation signs in a 3-clause. In the triple system $S^{(i)}$, the triple $(x_{i_1}, x_{i_2}, x_{i_3}) \in V^3$ is present iff the clause $l_{i_1} \vee l_{i_2} \vee l_{i_3}$ occurs in φ, where either $l_{i_j} = x_{i_j}$ or $l_{i_j} = \bar{x}_{i_j}$, according to the negation signs for the clause types in Table 2. For instance, in $S^{(3)}$ the triple $(x_{i_1}, x_{i_2}, x_{i_3})$ is present iff $x_{i_1} \vee \bar{x}_{i_2} \vee x_{i_3} \in \varphi$. Thus, the clauses in φ and the triples in $S^{(1)}, \ldots, S^{(8)}$ are in one-to-one correspondence.

Algorithm 18. 3-`Refute`(φ, ε)
Input: A set φ of 3-clauses over V. *Output:* An upper bound on the number of satisfiable clauses.

1. Compute the triple systems $S^{(i)}$ and run $\texttt{TripDisc}_{\varepsilon/8}(S^{(i)})$ for $i = 1, \ldots, 8$. If the output is "fail" for at least one i, then return the total number of clauses in φ as an upper bound and halt.
2. Return $(7 + \varepsilon)n^3 p$.

Finally, considering Table 2 instead of Table 1, we can prove Theorem 1 using a similar argument as in the proof of Lemma 11.

Remark 19. Though it is not stated explicitly in that paper, the approach in [14] can be used to obtain a strong refutation heuristic that certifies that at most a $(\frac{7}{8} + \varepsilon)$-fraction of the clauses can be satisfied almost surely. However, the methods in [14] only apply to somewhat bigger values of the clause probability p (namely, $p \geq n^{-3/2+\delta}$, $\delta > 0$ fixed) than those addressed in Theorem 1. Furthermore, the algorithm and the analysis that we have presented in the present paper are considerably simpler.

4 Hypergraph Problems

Let $H = (V, E) = H_{n,4,p}$ be a random 4-uniform hypergraph with vertex set $V = \{1, \ldots, n\}$. Let κ be an integer, and suppose that $p \geq c_0 \kappa^4 n^{-2}$ for a sufficiently large constant c_0. The algorithm 4-`RefuteCol` for Theorem 4 is randomized. On input H, the algorithm obtains a set $S \subset V^4$ of *ordered* 4-tuples as follows (recall that the edges E are not ordered). If $e = \{x_1, x_2, x_3, x_4\} \in E$, then there are $4! = 24$ possibilities to order the vertices x_1, x_2, x_3, x_4. Let $T(e)$ be the set of the 24 possible ordered tuples. Letting $p_0 = 1 - (1-p)^{1/24}$, we choose the set $\emptyset \neq X_e \subset T(e)$ of tuples that we include into S to represent e according to the distribution

$$\mathrm{P}(X_e) = p_0^{|X_e|}(1-p_0)^{24-|X_e|}p^{-1}.$$

Thus, each edge $e \in E$ gives rise to at least one tuple in S. The choice of the sets X_e is independent for all $e \in E$. Furthermore, we include each tuple $(x_1, x_2, x_3, x_4) \in V^4$ such that $|\{x_1, x_2, x_3, x_4\}| < 4$ into S with probability p_0 independently. A trite computation shows that if $H = H_{n,4,p}$, then the resulting set $S = S(H)$ of 4-tuples is distributed so that every possible 4-tuple in V^4 is present with probability p_0 independently.

Let $V_1 = V \times V \times \{1\}$, $V_2 = V \times V \times \{2\}$ be two disjoint copies of V. Having computed $S = S(H)$, 4-`RefuteCol` constructs a graph G with bipartition (V_1, V_2) in which the edge $\{(x_1, x_2, 1), (x_3, x_4, 2)\}$ is present iff $(x_1, x_2, x_3, x_4) \in S$. If $H = H_{n,4,p}$, then G is a random bipartite graph B_{n^2,p_0}. To this graph G, 4-`RefuteCol` applies the procedure `BipDisc`. If `BipDisc` answers "low discrepancy", then 4-`RefuteCol` answers "H is not κ-colorable". Otherwise, the output is "fail".

To prove the correctness of the algorithm, consider an independent set I of H, and let $I_i = I \times I \times \{i\} \subset V_i$ for $i = 1, 2$. Then, $E_G(I_1, I_2) = \emptyset$. Hence, if `BipDisc`$(G)$ outputs "low discrepancy", then (1) implies that $\#I < n/\kappa$ (provided that c_0 is large enough), so that $\chi(H) > \kappa$. The completeness follows from Prop. 5.

The heuristic 3-`RefuteInd`(H) for Theorem 3 transforms the hypergraph H into a triple system using in a similar manner as 4-`RefuteCol` (cf. Section 4). Then, 3-`RefuteInd`(H) applies the procedure `TripleDisc`.

Acknowledgment. We are grateful to Uri Feige for helpful discussions.

References

1. Achlioptas, D., Naor, A., Peres, Y.: The fraction of satisfiable clauses in a typical formula. Proc. 44th FOCS (2003) 362–370.
2. Achlioptas, D., Peres, Y.: The threshold for random k-SAT is $2^k \ln 2 - O(k)$. Proc. 35th STOC (2003) 223–231
3. Alon, N., Kahale, N.: A spectral technique for colouring random 3-colourable graphs. SIAM J. Comput. **26** (1997) 1733–1748.
4. Ben-Sasson, E.: Expansion in Proof Complexity. Ph.D. thesis, (http://www.eecs.harvard.edu/ eli/papers/thesis.ps.gz)
5. Chung, F.K.R.: Spectral Graph Theory. American Mathematical Society 1997
6. Coja-Oghlan, A., Goerdt, A., Lanka, A., and Schädlich, F.: Certifying unsatisfiability of random $2k$-Sat instances using approximation techniques. Proc. 14th FCT (2003) 15–26
7. Feige, U.: Relations between average case complexity and approximation complexity. Proc. 24th STOC (2002) 534–543
8. Feige, U., Ofek, E.: Spectral techniques applied to sparse random graphs. Report MCS03-01, Weizmann Institute of Science (2003) (http://www.wisdom.weizmann.ac.il/math/research.shtml)
9. Feige, U., Ofek, E.,: Easily refutable subformulas of large random 3CNF formulas. (http://www.wisdom.weizmann.ac.il/˜erano/)
10. Flaxman, A.: A spectral technique for random satisfiable 3CNF formulas. Proc. 14th SODA (2003) 357–363
11. Friedgut, E.: Necessary and sufficient conditions for sharp thresholds of graph properties and the k-SAT problem. Journal of the American Mathematical Society **12** (1999) 1017–1054
12. Friedman, J., Goerdt, A.: Recognizing more unsatisfiable random 3-Sat instances efficiently. Proc. 28th ICALP (2001) 310–321.
13. Füredi, Z., Komloś, J.: The eigenvalues of random symmetric matrices. Combinatorica **1** (1981) 233–241
14. Goerdt, A., Jurdzinski, T.: Some results on random unsatisfiable k-SAT instances and approximation algorithms applied to random structures. Combinatorics, Probability and Computing **12** (2003) 245 – 267
15. Goerdt, A., Lanka, A.: Recognizing more random unsatisfiable 3-SAT instances efficiently. Proc. Typical Case Complexity and Phase Transitions, Satellite Workshop of Logic in Computer Science 2003 (Ottawa).
16. Goerdt, A., Krivelevich., M.: Efficient recognition of random unsatisfiable k-SAT instances by spectral methods. Proc. 18th STACS (2001) 294–304
17. Håstad, J.: Some optimal inapproximability results. Journal of the ACM **48** (2001) 798–859
18. Janson, S., Łuczak, T., Ruciński, A.: Random graphs. John Wiley and Sons 2000
19. Krivelevich, M., Sudakov, B.: The largest eigenvalue of sparse random graphs. Combinatorics, Probability and Computing **12** (2003) 61–72
20. Krivelevich, M., Sudakov, B.: The chromatic numbers of random hypergraphs. Random Structures & Algorithms **12** (1998) 381–403
21. Vilenchik, D.: Finding a satisfying assignment for random satisfiable 3CNF formulas. M.Sc. thesis, Weizmann Institute of Science (2004)

A Proof of Lemma 9

Given a graph G with bipartition (V_1, V_2), we let G' be the graph obtained from G by deleting all edges that are incident with vertices of degree $> 10np$ in G. Furthermore, we let $A' = A(G')$ be the $n \times n$-matrix whose ij'th entry is 1 if v_i, w_j are adjacent in G', and 0 otherwise. We need the following lemma from [3] (Lemma 3.3 in that paper).

Lemma 20. *If $G = B_{n,p}$, then almost surely $|\langle A'\xi, \eta\rangle| = O(\sqrt{np})$ for all unit vectors $\xi, \eta \perp \mathbf{1}$.*

The next lemma shows that $e = \|\mathbf{1}\|^{-1}\mathbf{1}$ is "almost" an eigenvector of A' almost surely.

Lemma 21. *Let $G = B_{n,p}$. Then, $\|A'e - npe\| = O(\sqrt{np})$ almost surely.*

Proof. Letting d'_v denote be the degree of $v \in V_1$ in the graph G' and d_v the degree of v in G, we have

$$\|A'\mathbf{1} - np\mathbf{1}\|^2 = \sum_{v\in V_1}(d'_v - np)^2 \leq \sum_{v\in V_1}(d_v - np)^2.$$

Set $X = \sum_{v\in V_1}(d_v - np)^2$. Since d_v is binomially distributed with parameters n and p, we conclude that

$$\mathrm{E}\left(\|A'\mathbf{1} - np\mathbf{1}\|^2\right) \leq \mathrm{E}(X) = n\mathrm{Var}(d_v) \leq n^2 p.$$

Furthermore, as the random variables d_v are mutually independent, the variance of X is $\sum_{v\in V_1}\mathrm{Var}((d_v - np)^2)$. A trite computation shows that $\mathrm{Var}((d_v - np)^2) = O(np)^2$, whence $\mathrm{Var}(X) = O(n^3p^2)$. Therefore, Chebyshev's inequality entails that

$$\mathrm{P}(\|A'\mathbf{1} - np\mathbf{1}\|^2 > 2n^2p) \leq \mathrm{P}(X - \mathrm{E}(X) > n^2p) \leq O\left(\frac{n^3p^2}{n^4p^2}\right) = O(n^{-1}) = o(1),$$

thereby proving the lemma. □

Proof of Lemma 9. Let $G = B_{n,p}$. By Lemma 20 and Lemma 21 we may assume that $|\langle A'\xi, \eta\rangle| = O(\sqrt{np})$ for all unit vectors $\xi, \eta \perp \mathbf{1}$, that $\|A'e - npe\| = O(\sqrt{np})$, and that $\|A'^T e - npe\| = O(\sqrt{np})$. Let $M = M(G') = pJ - A'$. To bound

$$\alpha = \max\{M'\xi : \xi \perp \mathbf{1},\ \|\xi\| = 1\},$$

let $\xi, \eta \perp \mathbf{1}$ be unit vectors. Then,

$$|\langle M'\xi, \eta\rangle| = |\langle A'\xi, \eta\rangle| = O(\sqrt{np}), \tag{2}$$

because $J\xi = 0$. Further,

$$|\langle M\xi, e\rangle| = |\langle A'\xi, e\rangle| \leq |\langle A'^T e, \xi\rangle| \leq \|A'^T e - npe\| = O(\sqrt{np}). \tag{3}$$

Combining (2) and (3), we obtain $\alpha = O(\sqrt{np})$. Finally,

$$\|Me\| = \|npe - A'e\| = O(\sqrt{np}),$$

whence $\|M\| \leq \alpha + \|Me\| = O(\sqrt{np})$. □

B Proof of Lemma 15

Let $S = S_{n,p}$ be a random triple system with $p = f/n^{3/2}$ and $f \geq \ln^6 n$ as well as $f = o(n^{1/2})$. Let X be an arbitrary subset of V with $|X| = \alpha n$ and $\varepsilon \leq \alpha \leq 1 - \varepsilon$ and $Y = V \setminus X$. For $z \in V$ let

$$M_z \;=\; (X, X, \{z\}) \text{ and } M = \{(B, C) \,|\, B, C \;\in\; M_z \text{ for a } z \in V\,, \text{ and } B \neq C\},$$

thus a typical pair in M is $((b_1, b_2, z), (c_1, c_2, z))$ where $(b_1, b_2) \neq (c_1, c_2)$ and $b_1, b_2, c_1, c_2 \in X$. Furthermore let

$$m \;=\; |M| \;\text{ and }\; m_z \;=\; |M_z|.$$

We proceed in two steps. Step 1: We show that asymptotically $m \;=\; \alpha^4 f^2 n^2 = \alpha^4 s^2/n$. Step 2: We show that asymptotically $|(X, X, X)| \;=\; \alpha^3 s$.

From Step 2 the claim follows: By low discrepancy of the projections of S we have asymptotically that

$$|(X, X, V)| \;=\; |(X, V, X)| \;=\; |(V, X, X)| \;=\; \alpha^2 s \tag{4}$$

and we get

$$|(X, X, Y)| = |(X, X, V)| - |(X, X, X)| \;=\; (1 - \alpha)\alpha^2 s,$$

which applies in the same way to $|(X, Y, X)|, |(Y, X, X)|$. Asymptotic regularity of S implies that $|(X, V, V)| = \alpha n s/n \;=\; \alpha s$. From this we get

$$|(X, Y, Y)| \;=\; |(X, V, V)| - |(X, X, Y)| - |(X, Y, X)| - |(X, X, X)| \;=\; \alpha(1 - \alpha)^2 s.$$

We can argue in the same way for $|(Y, Y, X)|$ and $|(Y, X, Y)|$. As $1s - 3(1 - \alpha)\alpha^2 s - 3(1 - \alpha)^2 \alpha s - \alpha^3 s = (1-\alpha)^3 s$ we must have that $(Y, Y, Y) = (1-\alpha)^3 s$. As the set X is arbitrary we have that S has low discrepancy and the theorem is proved.

We first derive Step 2 from the equation proved in Step 1. Observing with (4) that

$$\sum_{z \in X} m_z \;+\; \sum_{z \in Y} m_z \;=\; \sum_{z \in V} m_z \;=\; |(X, X, V)| = \alpha^2 s(1 + o(1)), \tag{5}$$

we have that

$$m \;=\; \sum_{z \in V} m_z(m_z - 1) \;=\; \sum_{z} m_z^2 \;-\; \sum_{z} m_z = \sum_{z \in X} m_z^2 \;+ \sum_{z \in Y} m_z^2 \;- \alpha^2 s(1 + o(1)). \tag{6}$$

Now we have that

$$\sum_{z \in X} m_z^2 \;\geq\; \alpha n \left(\frac{|(X, X, X)|}{\alpha n}\right)^2 \;=\; \frac{|(X, X, X)|^2}{\alpha n}. \tag{7}$$

Estimate (7) holds because the sum $\sum_{z \in X} m_z^2$ subject to the condition $\sum_{z \in X} m_z \;=\; |(X, X, X)|$ is minimized when each term is the arithmetic mean of all αn terms $|(X, X, X)|/\alpha n$. With $|(X, X, Y)| = |(X, X, V)| - |(X, X, X)|$ we get in the same way that

$$\sum_{z \in Y} m_z^2 \;\geq\; (1 - \alpha)n \left(\frac{(|X, X, Y|}{(1 - \alpha)n}\right)^2 \;=\; \frac{(|(X, X, V)| - |(X, X, X|)^2}{(1 - \alpha)n}.$$

Now let the real δ be such that $|(X,X,X)| = (\alpha^3+\delta)s$. We show that $\delta = o(1)$. Using Step 1 we have asymptotically, that is up to $(1+o(1))$-factors, that

$$\begin{aligned}\alpha^4 f^2 n^2 &= m \\ &= \sum_{z\in X} m_z^2 + \sum_{z\in Y} m_z^2 - \alpha^2 s \qquad \text{Using (6).} \\ &\geq \frac{|(X,X,X)|^2}{\alpha n} + \frac{(|(X,X,V)| - |(X,X,X)|)^2}{(1-\alpha)n} - \alpha^2 s \qquad \text{Using (7).} \\ &= \frac{((\alpha^3+\delta)s)^2}{\alpha n} + \frac{(\alpha^2 s - (\alpha^3+\delta)s)^2}{(1-\alpha)n} - \alpha^2 s \qquad \text{Using (4).}\end{aligned}$$

Dividing both sides of the preceding estimate by $s^2/n = f^2 n^2$ we get by simple algebra

$$\begin{aligned}\alpha^4 &\geq \frac{(\alpha^3+\delta)^2}{\alpha} + \frac{(\alpha^2(1-\alpha)-\delta)^2}{1-\alpha} - o(1) \\ &= \alpha^5 + 2\delta\alpha^2 + \frac{\delta^2}{\alpha} + \alpha^4(1-\alpha) - 2\alpha^2\delta + \frac{\delta^2}{1-\alpha} - o(1) \\ &= \frac{\delta^2}{\alpha} + \alpha^4 + \frac{\delta^2}{1-\alpha} - o(1),\end{aligned}$$

and as $\varepsilon \leq \alpha \leq 1-\varepsilon$ we must have that $\delta = o(1)$ and thus $|(X,X,X)| = \alpha^3 s(1+o(1))$. This shows Step 2.

We are left to show Step 1. The Courant-Fischer characterization of Eigenvalues implies that

$$\lambda_1 = \max_{v\neq 0} \frac{v^T \mathbf{A} v}{v^T v} \qquad \text{and} \qquad \lambda_{n^2} = \min_{v\neq 0} \frac{v^T \mathbf{A} v}{v^T v},$$

where v stands for a real vector with n^2 coordinates, and v^T is the transpose of v.

Now let χ be the characteristic column vector of $X\times X$, that is χ is 1 in each coordinate corresponding to an element of $X\times X$ and 0 otherwise. We get that

$$\lambda_{n^2} \leq \frac{\chi^T \mathbf{A}\chi}{\chi^T\chi} \leq \lambda_1 \qquad \text{and therefore} \qquad \left|\frac{\chi^T\mathbf{A}\chi}{\chi^T\chi}\right| \leq \max\{\lambda_1, -\lambda_{n^2}\} = \|\mathbf{A}\|.$$

As $\chi^T\chi = |X\times X| = \alpha^2 n^2$ we have that

$$|\chi^T \mathbf{A}\chi| \leq \alpha^2 n^2 \cdot \|\mathbf{A}\|.$$

Direct linear algebra calculation and the definition of $\mathbf{A}$ shows that

$$\chi^T\mathbf{A}\chi = \sum_{(b_1,b_2)\in X\times X} \sum_{(c_1,c_2)\in X\times X} \mathbf{a}_{b_1,c_1,b_2,c_2} = \sum_{\substack{(b_1,b_2,c_1,c_2)\in X^4 \\ (b_1,b_2)\neq(c_1,c_2)}} \sum_{z\in V} (B_{b_1b_2z}\cdot B_{c_1c_2z} + B_{b_2b_1z}\cdot B_{c_2c_1z})$$

and we get that

$$\begin{aligned}2\cdot\left|\sum_{\substack{(b_1,b_2,c_1,c_2)\in X^4 \\ (b_1,b_2)\neq(c_1,c_2)}} \sum_{z\in V} B_{b_1b_2z}\cdot B_{c_1c_2z}\right| &= \left|\sum_{\substack{(b_1,b_2,c_1,c_2)\in X^4 \\ (b_1,b_2)\neq(c_1,c_2)}} \sum_{z\in V} (B_{b_1b_2z}\cdot B_{c_1c_2z} + B_{b_2b_1z}\cdot B_{c_2c_1z})\right| \\ &\leq \alpha^2 n^2 \|\mathbf{A}\|.\end{aligned}$$

We show below by lengthy algebra that asymptotically

$$\sum_{\substack{(b_1,b_2,c_1,c_2)\in X^4\\(b_1,b_2)\neq(c_1,c_2)}} \sum_{z\in V} B_{b_1b_2z}\cdot B_{c_1c_2z} \;=\; m \;-\; \alpha^4 n^2 f^2 \tag{8}$$

Equation (8) implies Step 1 as we now know that

$$2\cdot|m \;-\; \alpha^4 n^2 f^2| \le \alpha^2 n^2 \|\mathbf{A}\|$$

and as $\|\mathbf{A}\| = o(f^2)$ it must be the case that $m = \alpha^4 n^2 f^2(1+o(1))$.

To prove (8) we observe that $\displaystyle\sum_{\substack{(b_1,b_2,c_1,c_2)\in X^4\\(b_1,b_2)\neq(c_1,c_2)}} \sum_{z\in V} B_{b_1b_2z}\cdot B_{c_1c_2z}$ has the following terms:

(a) $\displaystyle\sum_{z\in V} m_z(m_z-1) = m$-times the term 1.
This is for those cases when $(b_1,b_2,z),(c_1,c_2,z)\in S$ and hence both B-factors above are -1.

(b) $2\cdot\displaystyle\sum_{z\in V} m_z(\alpha^2n^2-m_z)$-times the term $-p/(1-p)$.
This is for those cases when $(b_1,b_2,z)\in S$ and $(c_1,c_2,z)\notin S$ or vice versa. In this case one B-factor is -1 and the other one is $p/(1-p)$. Note that $X\times X = \alpha^2n^2$ and we have $\alpha^2n^2-m_z$ triple $(b_1,b_2,z)\notin S$ with $b_1,b_2\in X$

(c) $\displaystyle\sum_{z\in V}(\alpha^2n^2-m_z)\cdot(\alpha^2n^2-m_z-1)$-times the term $(p/(1-p))^2$.
This is for those cases when $(b_1,b_2,z),(c_1,c_2,z)\notin S$.

Observing that by assumption $s = fn^{3/2}\cdot(1+o(1))$ and by (5)

$$\sum_{z\in V} m_z = |(X,X,V)| = \alpha^2 f n^{3/2}(1+o(1))$$

we get for the terms in (b) using $1/(1-p) = 1+o(1)$ and $m_z = O(fn^{3/2}) = o(n^2)$

$$\begin{aligned}
2\cdot\sum_{z\in V} m_z(\alpha^2n^2-m_z)\cdot\frac{-p}{1-p} &= 2\cdot\sum_{z\in V} m_z(\alpha^2n^2\cdot(1+o(1)))\cdot -p\cdot(1+o(1))\\
&= -2\alpha^2n^2p\cdot(1+o(1))\cdot\sum_{z\in V} m_z\\
&= -2\alpha^2n^{1/2}f\cdot(1+o(1))\cdot\alpha^2fn^{3/2}\cdot(1+o(1))\\
&= -2\alpha^4f^2n^2(1+o(1)).
\end{aligned}$$

For the terms in (c) we get

$$\begin{aligned}
\sum_{z\in V}(\alpha^2n^2-m_z)\cdot(\alpha^2n^2-m_z-1)\left(\frac{p}{1-p}\right)^2 &= \sum_{z\in V}(\alpha^4n^4\cdot(1+o(1)))\cdot p^2\cdot(1+o(1))\\
&= \sum_{z\in V}\alpha^4n^4f^2/n^3\cdot(1+o(1)))\\
&= \alpha^4n^2f^2\cdot(1+o(1))
\end{aligned}$$

Summing all three types yields

$$m \;-\; \alpha^4f^2n^2\cdot(1+o(1))$$

which implies (8). □

C Proof of Lemma 17

Remember $\mathbf{A}$ has the following definition. For $0 < p < 1$ and $b_1, b_2, z \in V$ we let

$$B_{b_1b_2z} = B_{b_1b_2z}(S,p) = \begin{cases} -1 & \text{if } (b_1,b_2,z) \in S \\ p/(1-p) & \text{otherwise.} \end{cases}$$

Then, the $n^2 \times n^2$-matrix $\mathbf{A} = \mathbf{A}(S,p) = (\mathbf{a}_{b_1c_1,b_2c_2})_{(b_1,c_1),(b_2,c_2)\in V^2}$ is given by

$$\mathbf{a}_{b_1c_1,b_2c_2} = \begin{cases} \sum_{z\in V} (B_{b_1b_2z} \cdot B_{c_1c_2z} + B_{b_2b_1z} \cdot B_{c_2c_1z}) \text{ if } (b_1,b_2) \neq (c_1,c_2) \\ 0 \text{ if } \qquad (b_1,b_2) = (c_1,c_2) \end{cases}.$$

Each possible triple occurs in S with probability p. Therefore we have that for all $b_1, b_2, z \in V$ that

$$\mathbf{E}[B_{b_1b_2z}] = p \cdot (-1) + (1-p) \cdot \frac{p}{1-p} = -p + p = 0. \tag{9}$$

We need this later on.

Let $\lambda_1 \geq \ldots \geq \lambda_{n^2}$ be the eigenvalues of $\mathbf{A}$. Let λ denote the norm $\|\mathbf{A}\| = \max\{\lambda_1, -\lambda_{n^2}\}$ of $\mathbf{A}$. The trace of any matrix is the sum of the elements on the main diagonal of $\mathbf{A}$ and we have

$$\mathrm{Trace}[\mathbf{A}] = \sum \mathbf{a}_{b_1,c_2,b_1,c_2} = \sum_{i=1}^{n^2} \lambda_i \qquad \text{and} \qquad \mathrm{Trace}[\mathbf{A}^k] = \sum_{i=1}^{n^2} \lambda_i^k$$

for any integer $k \geq 1$. As the Eigenvalues of $\mathbf{A}$ are all real we have for even k that

$$\lambda^k \leq \sum_{i=1}^{n^2} \lambda_i^k = \mathrm{Trace}[\mathbf{A}^k]$$

and therefore especially that $\mathbf{E}[\lambda^k] \leq \mathbf{E}[\mathrm{Trace}[A^k]]$. We show below that there exists an even $k = k(n)$ that

$$\mathbf{E}[\mathrm{Trace}[\mathbf{A}^k]] \leq (\ln^4 n \cdot f)^k \tag{10}$$

in this case Markov's inequality implies that

$$\Pr[\lambda \geq \ln^5 n \cdot f] = \Pr[\lambda^k \geq (\ln^5 n \cdot f)^k] \leq \frac{\mathbf{E}[\lambda^k]}{(\ln^5 n \cdot f)^k} \leq \frac{\mathbf{E}[\mathrm{Trace}[\mathbf{A}^k]]}{(\ln^5 n \cdot f)^k} \leq \frac{(\ln^4 n \cdot f)^k}{(\ln^5 n \cdot f)^k} = o(1)$$

which is the lemma. We proceed to show (10). We have

$$\mathrm{Trace}[\mathbf{A}^k] = \sum_{b_1=1}^{n} \sum_{c_1=1}^{n} \ldots \sum_{b_k=1}^{n} \sum_{c_k=1}^{n} \mathbf{a}_{b_1c_1,b_2c2} \cdot \mathbf{a}_{b_2c_2,b_3c3} \cdot \ldots \cdot \mathbf{a}_{b_kc_k,b_1c1}$$

In case we have an $1 \leq i < k$ such that $(b_i, b_{i+1}) = (c_i, c_{i+1})$ or in case we have $(b_k, b_1) = (c_k, c_1)$ the whole product of the $\mathbf{a}$'s evaluates to 0. We ignore these cases in the sequel and assume that $(b_i, b_{i+1}) \neq (c_i, c_{i+1})$ for all $1 \leq i < k$ and $(b_k, b_1) \neq (c_k, c_1)$.

The definition of the $\mathbf{a}$'s yields

$$\begin{aligned}\mathrm{Trace}[\mathbf{A}^k] &= \sum_{b_1,\ldots,b_k}\sum_{c_1,\ldots,c_k}\left(\sum_{z_1\in V}(B_{b_1b_2z_1}\cdot B_{c_1c_2z_1}+B_{b_2b_1z_1}\cdot B_{c_2c_1z_1})\right)\cdot\ldots\\ &\qquad\cdot\left(\sum_{z_k\in V}(B_{b_kb_1z_k}\cdot B_{c_kc_1z_k}+B_{b_1b_kz_k}\cdot B_{c_1c_kz_k})\right)\\ &= \sum_{b_1,\ldots,b_k}\sum_{c_1,\ldots,c_k}\sum_{z_1,\ldots,z_k}(B_{b_1b_2z_1}\cdot B_{c_1c_2z_1}+B_{b_2b_1z_1}\cdot B_{c_2c_1z_1})\cdot\ldots\\ &\qquad\cdot(B_{b_kb_1z_k}\cdot B_{c_kc_1z_k}+B_{b_1b_kz_k}\cdot B_{c_1c_kz_k}).\end{aligned}$$

Performing the multiplications between the brackets we get 2^k terms X_j and

$$\mathrm{Trace}[\mathbf{A}^k]=\sum_{b_1,\ldots,b_k}\sum_{c_1,\ldots,c_k}\sum_{z_1,\ldots,z_k}\sum_{j=1}^{2^k}X_j$$

where each X_j has the appearance

$$X_j = B_{\beta_1}\cdot B_{\gamma_1}\cdot B_{\beta_2}\cdot B_{\gamma_2}\cdot\ldots\cdot B_{\beta_k}\cdot B_{\gamma_k}$$

with $\beta_i = b_ib_{i+1}z_i$ and $\gamma_i = c_ic_{i+1}z_i$ or $\beta_i = b_{i+1}b_iz_i$ and $\gamma_i = c_{i+1}c_iz_i$ for $1\le i<k$ and analogously with 1 instead of $i+1$ for $i=k$. Note that we can always assume that $\beta_i\neq\gamma_i$.

Let $B=(b_1,\ldots,b_k,c_1,\ldots,c_k)$ and $Z=(z_1,\ldots,z_k)$. We let $|B|=|\{b_1,\ldots,b_k,c_1,\ldots,c_k\}|$ and $|Z|=|\{z_1,\ldots,z_k\}|$ be the number of different elements of B and Z. We need to show

$$\mathbf{E}[\mathrm{Trace}[\mathbf{A}^k]]=\sum_{b=1}^{2k}\sum_{z=1}^{k}\sum_{\substack{B\\|B|=b}}\sum_{\substack{Z\\|Z|=z}}\sum_{j=1}^{2^k}\mathbf{E}[X_j]\le(\ln^4 n\cdot f)^k.$$

This sum can be shortened to

$$\mathbf{E}[\mathrm{Trace}[\mathbf{A}^k]]=\sum_{b=1}^{k+2}\sum_{z=1}^{k/2}\sum_{\substack{B\\|B|=b}}\sum_{\substack{Z\\|Z|=z}}\sum_{j=1}^{2^k}\mathbf{E}[X_j]\le(\ln^4 n\cdot f)^k.$$

Fix B with $|B|=b$, Z with $|Z|=z$, and let $X_j=B_{\beta_1}\cdot B_{\gamma_1}\cdot\ldots\cdot B_{\beta_k}\cdot B_{\gamma_k}$ be a term corresponding to B and Z. We show if $z>k/2$ or $b>k+2$ then there exists a factor B_δ inside X_j which occurs only once. In this case we have that $\mathbf{E}[X_j]=0$ by (9) as this B_δ is independent from the remaining factors of X_j.

Going along X_j from left to right there are exactly z slots where an element from Z occurs for the first time. At each such slot we get 2 B's which do not occur to the left in X_j. Thus we get at least $2z$ different B's in X_j. In order that each of these $2z$ B's occurs at least twice in X_j we must have that $2k\ge 4z$ or $z\le k/2$.

Again we go from left to right over X_j. The first two B-factors can use maximally 4 elements from B for the first time. All the remaining B-factors use at most two elements from B for the first time. This is because two elements are already determined by the predecessor. Thus except for the first two B-factors we need at least $b-4$ different B-factors. Altogether we need at least $2+(b-4)=b-2$ different B-factors. Again we

must have that $2(b-2) \leq 2k$ or $b \leq k+2$ in order that each of these different B-factors occurs at least twice in X_j.

Let B_α be a factor which occurs exactly r-times with $r \geq 2$ in X_j. Then we have that

$$E[B_\alpha^r] \;=\; p \cdot (-1)^r \;+\; (1-p) \cdot \left(\frac{p}{1-p}\right)^r \;\leq\; p \;+\; \frac{p^r}{(1-p)^{r-1}} \leq 2p$$

assuming $p \leq 1/2$. As we have at least $\max\{2z, b-2\}$ different B-factors in X_j we bound

$$\mathbf{E}[X_j] \leq (2p)^{\max\{2z,b-2\}}$$

which is independent from X_j, B, and Z. Therefore we only need to show that

$$\sum_{b=1}^{k+2} \sum_{z=1}^{k/2} \sum_{\substack{B \\ |B|=b}} \sum_{\substack{Z \\ |Z|=z}} 2^k \cdot (2p)^{\max\{2z,b-2\}} \leq (\ln^4 n \cdot f)^k.$$

Given b, each B with $|B| = b$ is obtained at least once by first picking a subset of b elements from V, $\leq n^b$ possibilities to choose, and second by placing the elements picked into $2k$ slots, $\leq b^{2k}$ possibilities. As we can assume $b \leq 2k$ we have at most $n^b(2k)^{2k}$ possibilities. Similarly we can bound the number of sequences Z with $|Z| = z$ by $n^z z^k$ and for $z \leq k$ we get a bound of $n^z k^k$. Therefore we have

$$\sum_{b=1}^{k+2} \sum_{z=1}^{k/2} \sum_{\substack{B \\ |B|=b}} \sum_{\substack{Z \\ |Z|=z}} 2^k \cdot (2p)^{\max\{2z,b-2\}} \leq \sum_{b=1}^{k+2} \sum_{z=1}^{k/2} 2^{3k} \cdot n^{b+z} \cdot k^{3k} \cdot (2p)^{\max\{2z,b-2\}}$$

We calculate next that

$$n^{b+z} \cdot (2p)^{\max\{2z,b-2\}} \leq (2f)^{\max\{2z,b-2\}} \cdot n^2.$$

First, let $2z > b-2$ then we have that $b \leq 2z+1$ and

$$n^{b+z} \cdot (2p)^{\max\{2z,b-2\}} \leq n^{3z+1}(2p)^{2z} = n^{3z+1}(2fn^{1/2}/n^2)^{2z} = n(2f)^{2z}.$$

Second let $b-2 \geq 2z$ then $z \leq b/2-1$ and we get

$$n^{b+z} \cdot (2p)^{\max\{2z,b-2\}} \leq n^{b+b/2-1}(2p)^{b-2} = n^{b+b/2-1}(2fn^{1/2}/n^2)^{b-2} = n^2(2f)^{b-2}.$$

As $b \leq k+2$ and $z \leq k/2$, we have $\max\{2z, b-2\} \leq k$ and we need to show

$$\sum_{b=1}^{k+2} \sum_{z=1}^{k/2} 2^{3k} \cdot k^{3k} \cdot n^2 \cdot (2f)^k \leq (\ln^4 n \cdot f)^k. \tag{11}$$

There is no restriction on k by now and we pick k as the smallest even integer $\geq \ln n$. For n sufficiently large we now get

$$\sum_{b=1}^{k+2} \sum_{z=1}^{k/2} 2^{3k} \cdot k^{3k} \cdot n^2 \cdot (2f)^k \leq (k+2) \cdot (k/2) \cdot 2^{4k} \cdot k^{3k} \cdot n^2 \cdot f^k \leq (\ln^4 n)^k \cdot f^k,$$

which yields (11) and the lemma. □