Recognizing more unsatisfiable random k-SAT instances efficiently

Joel Friedman[1], Andreas Goerdt[2⋆], and Michael Krivelevich[3] ⋆⋆

[1] Department of Mathematics, University of British Columbia,
Vancouver, BC V6T 1Z2, Canada
e-mail: jf@math.ubc.ca, homepage: www.math.ubc.ca/∼jf
[2] Fakultät für Informatik, TU Chemnitz, 09107 Chemnitz, Germany
e-mail: goerdt@informatik.tu-chemnitz.de
homepage: www.tu-chemnitz.de/informatik/HomePages/TI
[3] Department of Mathematics, Faculty of Exact Sciences, Tel Aviv University,
Tel Aviv 69978, Israel,
e-mail: krivelev@math.tau.ac.il

Abstract. It is known that random k-SAT instances with at least cn clauses where $c = c_k$ is a suitable constant are unsatisfiable (with high probability). We consider the problem to certify efficiently the unsatisfiability of such formulas. A backtracking based algorithm of Beame et al. shows that k-SAT instances with at least $n^{k-1}/(\log n)^{k-2}$ clauses can be certified unsatisfiable in polynomial time. We employ spectral methods to improve on this bound. We prove that for even $k \geq 4$ we present a polynomial time algorithm which certifies random k-SAT instances with at least $n^{(k/2)+o(1)}$ clauses as unsatisfiable (with high probability). For odd k we focus on 3-SAT instances and obtain an efficient algorithm for formulas with at least $n^{3/2+\varepsilon}$ clauses, where $\varepsilon > 0$ is an arbitrary constant.

Introduction

We study the complexity of certifying unsatisfiability of random k-SAT instances (or k-CNF formulas) over n propositional variables. All our discussion refers to k fixed and then letting n be sufficiently large. The probability space of random k-SAT instances has been widely studied in recent years for several good reasons. A somewhat arbitrary selection of the most recent literature is [Ac2000],[Fr99], [Be et al98], [AcSo2000], [AcMo2002], [AcPe2003].

⋆ Partially supported by the DFG.
⋆⋆ Partially supported by a USA-Israeli BSF grant and by a grant from the Israel Science Foundation.

One of the reasons for studying random k-SAT instances is that they have the following sharp threshold behaviour [Fr99]: there exists a function $c = c_k(n)$ such that for any $\varepsilon > 0$ formulas with at most $(1-\varepsilon) \cdot c \cdot n$ clauses are satisfiable whereas formulas with at least $(1+\varepsilon) \cdot c \cdot n$ are unsatisfiable with high probability (that means with probability tending to 1 when n goes to infinity). In fact, it is not known if $c_k(n)$ can be taken to be a constant (i.e., if the limit of $c_k(n)$ as $n \to \infty$ exists). It might be that $c_k = c_k(n)$ satisfying the aforementioned threshold property depends on n. However, it is known that c_k is at most $2^k \cdot \ln 2$ and the general conjecture is that c_k converges to a constant. For formulas with at least $2^k \cdot (\ln 2) \cdot n$ clauses the expected number of satisfying assignments of a random formula tends to 0 and the formulas are unsatisfiable with high probability. As the satisfiability problem for random 2-SAT instances is well known to be solvable in polynomial time, the most interesting case is that of 3-SAT. Accordingly much work is spent to approximate the value of c_3. The currently best results are that c_3 is at least 3.26 [AcSo2000] improved to the recent 3.52 [KaKiLa2002] and at most 4.601 [KiKrKr98] improved to 4.506 [DuBoMa2000]. For $k = 2$ the threshold is known, we have $c_2 = 1$ [ChRe92], [Go96].

The algorithmic interest in this threshold is due to the empirical observation that random k-SAT instances at the threshold, i.e. with around $c_k n$ random clauses are seemingly hard instances. The following behaviour has been reported consistently in experimental studies with suitably optimized backtracking algorithms searching for a satisfying assignment, see for example [SeMiLe96] [CrAu96]: the average running time is quite low for instances below the threshold. For 3-SAT instances we observe that almost all formulas with at most $4n$ clauses are satisfiable and it is quite easy to find a satisfying assignment. A precipitous increase in the average running time is observed at the threshold. For 3-SAT, about half of the formulas with $4.2n$ clauses are satisfiable and it is difficult to decide if a formula is satisfiable or not. Finally a speedy decline to lower complexity is observed beyond the threshold. For 3-SAT, almost all formulas with $4.5n$ clauses are unsatisfiable and the running time decreases again (in spite of the fact that now always the whole backtracking tree

must be searched.)

There are no general complexity theoretical results relating the threshold to hardness. The following observation is trivial: if we can efficiently certify almost all instances with dn clauses where d is above the threshold as unsatisfiable, then we can certify almost all instances with $d'n$ clauses , $d' > d$, as unsatisfiable by simply chopping off the superfluous clauses. The analogous fact holds below the threshold, where we extend a given formula with some random clauses. Of course similar remarks apply to the size of clauses.

The relationship of hardness and thresholds is not restricted to satisfiability. It has also been observed for k-colourability of random graphs with a linear number of edges. In [PeWe89] a peak in running time seemingly related to the threshold is reported. The existence of a threshold is proved in [AcFr99] but again the value and convergence to a constant are only known experimentally. For the subset sum problem which is of a quite different nature we also have the following relationship between threshold and hardness: the threshold is known, and some discussion related to hardness is found in [ImNa96]. For the similarly looking number partitioning problem threshold results can be found [BoChPi2001]

Abandoning the general complexity theoretic point of view and looking at concrete algorithms the following results are known for random k-SAT instances: all progress approximating the threshold from below is based on the analysis of rather simple polynomial time heuristics and is mostly restricted to clause size $k = 3$. In fact the most advanced heuristic being analyzed [AcSo2000] only finds a satisfying assignment with probability of at least ε, where $\varepsilon > 0$ is a small constant, for 3-SAT formulas with at most $3.26n$ clauses. The same applies to the recent improvement to $3.52n$ clauses in [KaKiLa2002]. The heuristic in [FrSu96] finds a satisfying assignment almost always for random 3-SAT instances with at most $3.003n$ clauses. On the other hand the progress made in approximating the threshold from above does not provide us at all with efficient algorithms certifying the unsatisfiability of the formula at hand. Only the expectation of the number of satisfying assignments is calculated

and is shown to tend to 0.

In fact beyond the threshold we have negative results: for arbitrary but fixed $d \geq 2^k \cdot \ln 2$ random k-SAT instances with dn clauses (are unsatisfiable and) have only resolution proofs with an exponential number, that is with at least $2^{\Omega(n)}$ clauses with high probability [ChSz88]. This has been improved upon by [Fu95], [BePi96], and [Be et al98] all proving (exponential) lower bounds for somewhat larger clause/variable ratios. Note that a lower bound on the size of resolution proofs provides a lower bound on the number of nodes in *any* classical backtracking tree as generated by any variant of the well known Davis-Putnam procedure.

Provably polynomial time results beyond the threshold are rather limited at present: in [Fu95] it is shown that random k-SAT formulas with at least n^{k-1} clauses allow for polynomial size resolution proofs with high probability. This is strengthened in [Be et al98] to the best result known at present: for at least $n^{k-1}/(\log\, n)^{k-2}$ random clauses a backtracking based algorithm proves unsatisfiability in polynomial time with high probability. (The result of Beame et al. is slightly stronger as it applies to formulas with $\Omega(n^{k-1}/(\log n)^{k-2})$ random clauses.)

We extend the region where a provably polynomial time algorithm exists. For even $k \geq 4$ we give an algorithm which works when the number of clauses is only n to a constant fraction of k (with high probability), that is for formulas with at least $\mathrm{POLY}(\log n) \cdot n^{k/2} = n^{(k/2)+o(1)}$ clauses, where POLY denotes a sufficiently fast growing polynomial (a degree of 7 is sufficient). To obtain our result we leave the area of strictly combinatorial algorithms considered up to this point. Instead we associate a graph with a given formula and show how to certify unsatisfiability of the formula with the help of the eigenvalue spectrum of a certain matrix associated to this graph. Note that our algorithm is not complete but only complete with high probability; in return for sacrificing completeness, we get small size proofs that are efficiently computable. Note that the eigenvalue spectrum can be approximated to arbitrary accuracy in polynomial time by standard linear algebra methods. With respect to odd k

we focus on the case $k = 3$ and show by extending the previous arguments that random 3-SAT instances with $n^{(3/2)+\varepsilon}$ clauses can be efficiently certified as unsatisfiable with high probability, again improving the $n^2/\log n$ bound of Beame et al. In very recent work the aformentioned results have been improved, for $k = 4$ we have an efficient certification algorithm for Cn^2 random clauses where C is a sufficiently large constant [CoGoLaSch2003]. And for $k = 3$ we have a bound of $\mathrm{POLY}(\log n) \cdot n^{3/2}$ clauses [GoLa2003]. It seems to be a very hard task to get a bound below $n^{k/2}$ random k-clauses; for one thing, this $n^{k/2}$ seems to be a natural barrier to the spectral techniques we use.

Eigenvalues are used in two ways in the algorithmic theory of random structures: they can be used to find a solution of an NP-hard problem in a random instance generated in such a way that it has a solution (not known to the algorithm). An example for 3-colourability is [AlKa94]. They can also be used to prove the absence of a solution of an NP-problem. However these applications are somewhat rare at the moment. The most prominent example here is the expansion property of random regular graphs [AlSp92]. Note that the expansion property is coNP-complete [Bl et al81] and the eigenvalues certify the absence of a non-expanding subset of vertices (which is the solution in this case). Our result is an example of the second kind and it might be worthwhile to investigate more systematically the existence of efficient algorithms for coNP-complete properties of random structures. A general overview on eigenvalues as applied to random graphs is [Al98].

We use the following notation throughout. The probabilistic model $\mathrm{Form}_{n,k,m}$ of k-CNF formulas with m clauses over n propositional variables is defined as follows: the probability space of clauses of size k, $\mathrm{Clause}_{n,k}$, is the set of ordered k-tuples of literals over n propositional variables $v_1, \ldots, v_n$. We write $l_1 \vee \ldots \vee l_k$ with $l_i = x$ or $l_i = \neg x$ where x is one of our variables. Our definition of $\mathrm{Clause}_{n,k}$ allows for clauses containing the same literal twice and clauses which contain a variable and its negation in order to simplify the subsequent presentation. We consider $\mathrm{Clause}_{n,k}$ as endowed with the uniform probability distribution: the probability of a clause

is given by $Pr(\, l_1 \vee \ldots \vee \, l_k \,) \, = \, (1/(2n))^k$. $\mathrm{Form}_{n,k,m}$ is the m-fold cartesian product space of $\mathrm{Clause}_{n,k}$. We write $F \, = \, C_1 \wedge \ldots \wedge \, C_m$ and $Pr(F) \, = \, (1/(2n))^{k \cdot m}$. The probability space $\mathrm{Form}_{n,k,p}$ is obtained by the following generation procedure: throw each clause with probability p into the formula to be generated, and the probability of a given formula with m clauses is $p^m \cdot (1-p)^{(2n)^k - m}$. These two spaces $\mathrm{Form}_{n,k,m}$ and $\mathrm{Form}_{n,k,p}$ are essentially equivalent when $m \, = \, p \cdot (2n)^k$. There are several other ways of defining k-SAT probability spaces (for example clauses might be sets of literals instead of sequences, tautological clauses could be forbidden). In spite of the fact that we have not really checked the details we feel confident and in line with general usage to assume that our results can be transferred to these other spaces, too. Note that the clause size k always is fixed.

1 Even Clause Size

1.1 From random formulas to random graphs

The following simple observation underlies our algorithm:

Lemma 1. *If a propositional formula F in k-CNF over n variables is satisfiable, there exists a subset S of at least $n/2$ variables such that F has no all-positive clause $x_1 \vee \ldots \vee x_k$ with $x_i \in S$ for all i or F has no all-negative clause $\neg x_1 \vee \ldots \vee \neg x_k$ with $x_i \in S$ for all i.*

Proof. Let $\mathcal{A} : \{v_1, \ldots, v_n\} \to \{0, 1\}$ be a satisfying assignment for F. $\mathcal{A}$ sets at least $n/2$ variables to 0 (=false) or to 1 (=true). In the first case the set of variables set to 0 satisfies the lemma otherwise the set of variables set to 1 does it. □

Our algorithm proves unsatisfiability by efficiently showing the *non*-existence of a set S as in the preceding lemma. To do this we translate the non-existence of S into a graph theoretical condition. To this end we assign two graphs to a formula. Let $F \in \mathrm{Form}_{n,k,m}$, where k is even, be given. The graph $G \, = \, G_F$ depends only on the sequence of all-positive clauses of F:

- The set of vertices of G is $V \, = \, V_F \, = \, \{x_1 \vee \ldots \vee x_{k/2} \,|\, x_i \text{ a variable}\}$. We have $|V| = n^{k/2}$ and V is independent of F.

– The set of edges of G, $E = E_F$ is given as follows. For $x_1 \vee \ldots \vee x_{k/2}$, $\neq$ $y_1 \vee \ldots \vee y_{k/2}$ we have $\{x_1 \vee \ldots \vee x_{k/2},\ y_1 \vee \ldots \vee y_{k/2}\} \in E$ iff $x_1 \vee \ldots \vee x_{k/2} \vee y_1 \vee \ldots \vee y_{k/2}$ (or $y_1 \vee \ldots \vee y_{k/2} \vee x_1 \vee \ldots \vee x_{k/2}$) is a clause of F. Note that it is possible that $|E| < m$ as clauses might induce no edge or two clauses induce the same edge. We do not allow for loops or multiple edges.

The graph H_F is defined in a totally analogous way for the all-negative clauses of F.

Recall that an *independent set* of a graph G is a subset of vertices W of G such that we have no edge $\{v, w\}$ in G where both $v, w \in W$. The next lemma follows directly from Lemma 1 as a set S of variables induces $|S|^{k/2}$ vertices in G_F consisting only of variables from S.

Lemma 2. *If $F \in Form_{n,k,m}$ is satisfiable then G_F or H_F has an independent set W of vertices with*

$$|W| \geq (n/2)^{k/2} = (1/2)^{k/2} \cdot |V|$$

Note, as k remains constant when n gets large this is a constant fraction of all vertices of G_F. We need to show that the distribution of G_F is just the distribution of a usual random graph. To this end let be $G_{n,m}$be the probability space of random graphs with n labelled vertices and m different edges. Each graph is equally likely, that is the probability of G is $Pr(G) = 1/\binom{\binom{n}{2}}{m}$.

Lemma 3. *(1) Conditional on the event in $Form_{n,k,m}$ that $|E_F| = r$ the graph G_F is a random member of the space $G_{\nu,r}$ where $\nu = n^{k/2}$ is the number of vertices of G_F.*
(2) Let $\varepsilon > 0$, with high probability the number of edges of G_F is between $m \cdot (1/2)^k \cdot (1 - \varepsilon)$ and $m \cdot (1/2)^k \cdot (1 + \varepsilon)$.

Proof. (1) The proof is a well known (but sometimes forgotten) trick: let G be a graph with vertices V_F as defined above and with edge set E with $|E| = r$. The set of formulas inducing this edge set can be constructed as follows: (1) make each edge into a clause (2^r possibilities); (2) put the clauses into r slots from altogether m slots such

that the leftmost slot is not empty; (3) fill each empty slot with a clause or its reversal[1] to its left. Each formula arises uniquely from some such procedure, and the number of formulas generated from each graph depends only on r and m but not on E itself, which shows the claim.

(2) The claim follows from the following statements which we prove further below:

- Let $\varepsilon > 0$ be fixed. The number of all-positive clauses of $F \in \mathrm{Form}_{n,k,m}$ is between $(1-\varepsilon) \cdot (1/2)^k \cdot m$ and $(1+\varepsilon) \cdot (1/2)^k \cdot m$ with high probability.

- The number of all-positive clauses like $x_1 \vee \ldots \vee x_{k/2} \vee x_1 \vee \ldots \vee x_{k/2}$, that is with the same first and second half, is $o(m)$.

- The number of unordered pairs of positions of a formula F in which we have all-positive clauses which induce only one edge, that is pairs of clauses $x_1 \vee \ldots \vee x_k$, $y_1 \vee \ldots \vee y_k$ where $\{x_1 \vee \ldots \vee x_{k/2},\, x_{k/2+1} \vee \ldots \vee x_k\} = \{y_1 \vee \ldots \vee y_{k/2},\, y_{k/2+1} \vee \ldots \vee y_k\}$ is also $o(m)$ with high probability.

This implies the claim of the lemma with the actual ε slightly lower than the ε from the first statement above because we have only $o(m)$ clauses inducing no additional edge.

The first statement: This statement follows with Chernoff bounds because the probability that a clause at a fixed position is all-positive is $(1/2)^k$ and clauses at different positions are independent.

The second statement: The probability that the clause at position i has the same first and second half is $(1/2n)^{k/2}$. The expected number of such clauses in a random F is therefore $m \cdot (1/2n)^{k/2} = o(m)$.

The third statement: We fix 2 slots of clauses $i \neq j$ of F. The probability that the clauses in these slots have the same set of first and second halves is bounded above by $(1/n)^k$ and the expected

[1] By the *reversal* of a clause we mean the clause obtained by interchanging the first and second half of its variables.

number of such unordered pairs is at most $m^2 \cdot (1/n)^k = O(m/n)$ provided $m = O(n^{k-1})$ which we can assume. Let X be the random variable counting the number of unordered pairs of positions with clauses with the same first and second half and let $\varepsilon > 0$. Markov's inequality gives us

$$Pr(X > n^{\varepsilon} \cdot EX) \leq EX/(n^{\varepsilon} \cdot EX) = 1/n^{\varepsilon}.$$

Therefore we get that with high probability $X \leq n^{\varepsilon} \cdot (m/n) = o(m)$. □

Spectral considerations Eigenvalues of matrices associated with general graphs are somewhat less common at least in Computer Science applications than those of regular graphs. The monograph [Ch97] is a standard reference for the general case. The easier regular case is dealt with in [AlSp92]. The necessary Linear Algebra details cannot all be given here. They are very well presented in the textbook [St88].

Let $G = (V, E)$ be an undirected graph (loopless and without multiple edges) with $V = \{1, \ldots, n\}$ being a standard set of n vertices. For $0 < p < 1$ we consider the matrix $A = A_{G,p}$ as in [KrVu2000] and [Ju82] which is defined as follows: The $n \times n$-matrix $A = A_{G,p} = (a_{i,j})_{1 \leq i,j \leq n}$ has $a_{i,j} = 1$ iff $\{i, j\} \notin E$ and $a_{i,j} = -(1-p)/p = 1 - 1/p$ iff $\{i, j\} \in E$. In particular $a_{i,i} = 1$. As A is real valued and symmetric A has n real eigenvalues when counting them with their multiplicities and allowing for 0 as an Eigenvalue (necessary when the matrix is not of full rank). We denote these eigenvalues by $\lambda_1(A) \geq \lambda_2(A) \geq \cdots \geq \lambda_n(A)$.

Recall that the *independence number* of G, denoted by $\alpha(G)$, is the size (= number of vertices) of a largest independent set of G. In general it is NP-hard to determine the independence number. But we have an efficiently computable bound:

Lemma 4. *(Lemma 4 of [KrVu2000]) For any p with $p > 0$ we have $\lambda_1(A_{G,p}) \geq \alpha(G)$.*

Proof. Let $l = \alpha(G)$. Let χ be the characteristic column vector of a largest independent set of G (i. e. taking the value 1 on elements from the set and 0 otherwise). The matrix $A_{G,p}$ has a $l \times l$-block which contains only 1's. This block of course is indexed with the vertices from our largest independent set. From the Courant-Fisher characterization of the eigenvalues of real valued symmetric matrices we get that

$$\lambda_1(A) \geq \frac{\chi^{tr} \cdot A \cdot \chi}{\chi^{tr} \cdot \chi} = l^2/l = l.$$

□

In order to bound the size of the eigenvalues of $A_{G,p}$ when G is a random graph we rely on a suitably modified version of the following theorem:

Theorem 5. *(Theorem 2 of [FuKo81]) Let for $1 \leq i, j \leq n$ and $i \leq j$ $a_{i,j}$ be independent , real valued random variables (not necessarily identically distributed) satisfying the following conditions:*

- *$|a_{i,j}| \leq K$ for all $i \leq j$,*
- *the expectation $Ea_{i,i} = \nu$ for all i,*
- *the expectation $Ea_{i,j} = 0$ for all $i < j$,*
- *the variance $Va_{i,j} = E[a_{i,j}^2] - (Ea_{i,j})^2 = \sigma^2$ for all $i < j$,*

where the values K, ν, σ are constants independent of n.

For $j \geq i$ let $a_{j,i} = a_{i,j}$ and let $A = (a_{i,j})_{1 \leq i,j \leq n}$ be the random $n \times n$-matrix defined by the $a_{i,j}$. Let the eigenvalues of A be $\lambda_1(A) \geq \lambda_2(A) \geq \cdots \geq \lambda_n(A)$. With probability at least $1 - (1/n)^{10}$ the matrix A is such that

$$\max\{|\lambda_i(A)| \mid 1 \leq i \leq n\} = 2 \cdot \sigma \cdot \sqrt{n} + O(n^{1/3} \cdot \log n) = 2 \cdot \sigma \cdot \sqrt{n} \cdot (1 + o(1)).$$

□

We intend to apply this theorem to the random matrix $A = A_{G,p}$ where G is a random graph from the probability space $G_{n,m}$. However, in this case the entries of A are not strictly independent and Theorem 5 cannot be directly applied. We first consider random graphs from the space $G_{n,p}$ and proceed to $G_{n,m}$ later on. Recall

that a random graph G from $G_{n,p}$ is obtained by inserting each possible edge with probability p independently of other edges.

For p constant and G a random member from $G_{n,p}$ the assumptions of Theorem 5 can easily be checked to apply to $A_{G,p}$. However, for sparser random graphs that is $p = p(n) = o(1)$ the situation changes. We have that $a_{i,j}$ can assume the value $-1/o(1) + 1$ and thus is not any more bounded above by a constant. The same applies to the variance: $\sigma^2 = (1-p)/p = 1/o(1) - 1$.

It can however be checked that the proof of Theorem 5 as given in [FuKo81] goes through as long as we consider matrices $A_{G,p}$ where $p = (\ln n)^7/n$. In this case we have that $K = n/(\ln n)^7 - 1$ and $\sigma^2 = n/(\ln n)^7 - 1$. With this modification and the other assumptions just as before the proof of [FuKo81] leads to:

Corollary 6. *With probability at least* $1 - (1/n)^{10}$ *the random matrix* A *satisfies*

$$\max\{|\lambda_i(A)| \mid 1 \le i \le n\} = 2\cdot\sigma\cdot\sqrt{n} + O(n/(\ln n)^{22/6}) = 2\cdot(1/(\ln n)^{7/2})\cdot n\cdot(1+o(1)).$$

Proof. We sketch the changes which need to be applied to the proof of Theorem 2 in [FuKo81]. These changes refer to the final estimates of the proof on page 237. We set

$$k := (\sigma/K)^{1/3}\cdot n^{1/6} = (\ln n)^{7/6}(1+o(1)),$$

in fact k should be the following even number. We set the error term

$$\nu := 50\cdot n/(\ln n)^{22/6}.$$

We have

$$2\cdot\sigma\cdot\sqrt{n} = 2\cdot n/(\ln n)^{7/2} = 2\cdot n/(\ln n)^{21/6}$$

which implies that $\nu = o(2\cdot\sigma\cdot\sqrt{n})$. Concerning the error estimate we get

$$\frac{\nu\cdot k}{2\cdot\sigma\cdot\sqrt{n}+v} = \frac{50\cdot(\ln n)^{7/6}}{(\ln n)^{1/6}}\cdot(1+o(1)) = 50\cdot\ln n\cdot(1+o(1)).$$

This implies the claim. □

Together with Lemma 4 we now get an efficiently computable certificate bounding the size of independent sets in random graphs from $G_{n,m}$.

Corollary 7. *Let G be a random member from $G_{n,m}$ where $m = ((\ln n)^7/2) \cdot n$. and let $p = m/\binom{n}{2} = (\ln n)^7/(n-1)$. We have with high probability that*

$$\lambda_1(A_{G,p}) \leq 2 \cdot (1/(\ln n)^{7/2}) \cdot n \cdot (1 + o(1)).$$

Proof. The proof is a standard transfer from the random graph model $G_{n,p}$ to $G_{n,m}$. For G random from $G_{n,p}$ the induced random matrix $A_{G,p}$ satisfies the assumptions of the last corollary. We have that with probability at least $1 - (1/n)^{10}$ the eigenvalues of $A_{G,p}$ are bounded by $2 \cdot (1/(\ln n)^{7/2}) \cdot n \cdot (1 + o(1))$.

By the Local Limit Theorem for the binomial distribution named after de Moivre-Laplace the probability that a random graph from $G_{n,p}$ has *exactly* m edges is of $\Omega(1/(n \cdot p)^{1/2}) = \Omega(1/(\ln n)^{7/2})$. This implies the claim as the probability in $G_{n,p}$ that the eigenvalue is not bounded as claimed is $O((1/n)^{10}) = o(1/(\ln n)^{7/2})$. (We omit the formal conditioning argument.) □

1.2 The algorithm

We fix the clause size $k \geq 4$ and assume that k is even. We consider the probability space of formulas $\mathrm{Form} = \mathrm{Form}_{n,k,m}$ where the number of clauses is

$$m = 2^k \cdot (\ln n^{k/2})^7 \cdot n^{k/2} = 2^k \cdot (k/2)^7 \cdot (\ln n)^7 \cdot n^{k/2}.$$

Given a random formula F from Form the algorithm first considers the all-positive clauses from F and constructs the graph G_F. From Lemma 3 we know that $G = G_F$ is a random member of $G_{\nu,\mu}$ where $\nu = n^{k/2}$ and μ is at least $m \cdot (1/2)^k \cdot (1 - \varepsilon) = (\ln \nu)^7 \cdot \nu \cdot (1 - \varepsilon)$, where we fix $\varepsilon > 0$ sufficiently small, in fact $\varepsilon = 1/2$ will do. In case the number of edges is smaller than this bound the algorithm fails.

The algorithm determines the matrix $A = A_{G,p}$ where $p = \mu/\binom{\nu}{2} \geq (\ln \nu)^7/(\nu - 1)$. From Corollary 7 we get that with high probability

$$\lambda_1(A) \leq 2 \cdot (1/(\ln \nu)^{7/2}) \cdot \nu \cdot (1 + o(1)).$$

and thus $\lambda_1(A) < (1/2)^k \cdot \nu = (1/2)^k \cdot n^{k/2}$ with high probability. The algorithm approximates $\lambda_1(A)$ by a polynomial time algorithm. In case the preceding event does not occur the algorithm fails. By Lemma 4 G_F is now certain (not only certain with high probability) to have no independent set comprising $(1/2)^k \cdot n^{k/2}$ vertices.

The algorithm proceeds in the same way for the all negative clauses and the graph H_F. In case it succeeds (which happens with high probability) we have that F is unsatisfiable by Lemma 2. If the algorithm fails, which happens with probability $o(1)$ we do not know if the formula is satisfiable or not.

In case that the number of literals k is odd we can extend each clause by a random literal and apply the preceding algorithm. It succeeds whith high probability when the number of clauses is $2^{k+1} \cdot ((k+1)/2)^7 \cdot (\ln n)^7 \cdot n^{(k+1)/2}$. Concerning the probability space $\mathrm{Form}_{n,k,p}$ the algorithm succeeds with high probability when p is picked such that the expected number of clauses is at least $(1 + \varepsilon)\cdot$ "the bound above", because in this case the number of clauses is with high probability at least as large as this bound.

2 Clause size 3

From now on we restrict attention to the family of probability spaces $\mathrm{Form}_{n,p} = \mathrm{Form}_{n,3,p}$. We assume that $p = p(n) = 1/n^{1+\gamma}$ where $1/2 > \gamma > 0$ is a constant. Our formulas get sparser with increasing γ. Note that our space of formulas is analogous to the space of random graphs $G_{n,p}$. The number of clauses in a random instance from $\mathrm{Form}_{n,p}$ follows the binomial distribution with parameters $8n^3$ and p, $\mathrm{Bin}(8n^3, p)$ and the expected number of clauses is $8n^3 \cdot p = 8 \cdot n^{2-\gamma} = 8 \cdot n^{3/2+\varepsilon}$, where $\varepsilon = 1/2 - \gamma > 0$.

2.1 From random 3-SAT instances to random graphs

We state a graph theoretical condition which implies the unsatisfiability of a 3-SAT instance F over n propositional variables. To this end we again define the graphs G_F and H_F. $G_F = (V_F, E_F)$ is defined as follows:

- V_F is the set of ordered pairs over the n propositional variables. We have $|V_F| = n^2$.

- The edge (a_1, b_1)——(a_2, b_2) (where in order to avoid loops $(a_1, b_1) \neq (a_2, b_2)$ that is $a_1 \neq a_2$ or $b_1 \neq b_2$) is in E_F iff there exists a variable z such that F contains the two clauses $a_1 \vee a_2 \vee z$ and $b_1 \vee b_2 \vee \neg z$

The graph H_F is defined analogously but with different clauses: Its vertices are as before ordered pairs of variables, and (a_1, b_1)——(a_2, b_2) is an edge iff F has the clauses $\neg a_1 \vee \neg a_2 \vee z$ and $\neg b_1 \vee \neg b_2 \vee \neg z$ for a variable z. Note that in the case of G_F the clause pairs $a_1 \vee a_2 \vee z$, $b_1 \vee b_2 \vee \neg z$ and $a_2 \vee a_1 \vee z'$,$b_2 \vee b_1 \vee \neg z'$ induce the same edge (a_1, b_1)——(a_2, b_2). Of course analogous remarks apply to H_F.

Some remarks concerning the intuition of this definition follow from the previous case $k = 4$. The clause $a_1 \vee a_2 \vee b_1 \vee b_2$ is obtained by resolution [Sch89] with z from the two clauses $a_1 \vee a_2 \vee z$ and $b_1 \vee b_2 \vee \neg z$ which define an edge of G_F. Similarly we have that $\neg a_1 \vee \neg a_2 \vee \neg b_1 \vee \neg b_2$ is obtained from $\neg a_1 \vee \neg a_2 \vee z$ and $\neg b_1 \vee \neg b_2 \vee \neg z$. The correctness of resolution states that F is unsatisfiable if a set of resolvents (that is clauses obtained by resolution) of F is unsatisfiable.

For any z the number of clauses like $a_1 \vee a_2 \vee z$ and of clauses like $b_1 \vee b_2 \vee \neg z$ is concentrated at the expectation $\approx n^2 \cdot p = n^{1-\gamma} > n^{1/2}$ as $\gamma < 1/2$. Applying resolution with z to all these clauses gives $\approx n^{(1-\gamma)2} > n$ clauses $a_1 \vee a_2 \vee b_1 \vee b_2$. Doing this for all n variables z gives $> n^2$ all positive clauses of size 4. In the same way we get $> n^2$ all negative 4-clauses. Efficiently bounding the size of independent sets in these graphs we get an efficient algorithm which demonstrates unsatisfiability of these newly obtained 4-clauses. The

correctness of resolution implies that F itself is unsatisfiable.

Some detailed remarks concerning G_F: Only for technical reasons the variable z which is resolved is the *last* variable in our clauses. (Recall we consider clauses as ordered triples.) More important is the fact that the edge reflects the resolvent $a_1 \vee a_2 \vee b_1 \vee b_2$ *not* in the most natural way by the edge (a_1, a_2)——(b_1, b_2) but by (a_1, b_1)——(a_2, b_2). The variables of the vertices connected by the edge come from the *different* clauses taking part in the resolution step. The reason why this is important is to decrease the stochastic dependency of the edges of G_F when F is a random formula. Again more of a technical nature is the convention that the variables in the first position of each vertex come from the clause which contains the positive literal z, whereas the second variables b_1, b_2 come from the clause with $\neg z$.

Recall again that $\alpha(G)$ is the *independence number* of G that is the maximum number of vertices of an independent set of G.

Theorem 8. *If F is a 3-SAT instance over n variables which is satisfiable then we have:*

$$\alpha(G_F) \geq n^2/4 \quad or \quad \alpha(H_F) \geq n^2/4.$$

Proof. Let $\mathcal{A}$ be an assignment of the n underying propositional variables with 0, 1 (where 0 = false and 1 = true) which makes F true. We assume that $\mathcal{A}$ assigns 1 to at least $n/2$ variables. Let S be this set of variables. We show that the set of vertices $S \times S$ is an independent set of H_F. As this set has at least $(n/2)^2 = n^2/4$ vertices the claim holds.

Let (a_1, b_1)——(a_2, b_2) be an arbitrary edge from H_F. Then F has the clauses $\neg a_1 \vee \neg a_2 \vee z$ and $\neg b_1 \vee \neg b_2 \vee \neg z$ (or $\neg a_2 \vee \neg a_1 \vee z$ and $\neg b_2 \vee \neg b_1 \vee \neg z$). As the assignment $\mathcal{A}$ makes F true $\mathcal{A}$ sets at least one of the literals $\neg a_1$, $\neg a_2$, $\neg b_1$, or $\neg b_2$ to 1. (Here the correctness proof of resolution is hidden: The clause $\neg a_1 \vee \neg a_2 \vee \neg b_1 \vee \neg b_2$ is a resolvent of $\neg a_1 \vee \neg a_2 \vee z$ and $\neg b_1 \vee \neg b_2 \vee \neg z$ and we have: If $\mathcal{A}$ satisfies F then $\mathcal{A}$ satisfies all resolvents of F.) So $\mathcal{A}$ sets at least one of the variables a_1, a_2, b_1 or b_2 to 0. But this means that one

of these variables is not in our set S. This finishes our proof as now $(a_1, b_1) \notin S \times S$ or $(a_2, b_2) \notin S \times S$. As the initially chosen edge is arbitrary, we get that $S \times S$ is an independent set of H_F. Finally, if $\mathcal{A}$ sets at least half of the variables to 0 we apply the same argument to G_F. □

Now we will proceed in an analogous way as before: Given a random F from $\mathrm{Form}_{n,p}$ the graphs G_F and H_F are certain random graphs. With high probability our algorithm certifies that G_F has no independent set of size $\geq n^2/4$. The same applies to H_F. Therefore F is certified unsatisfiable.

The occurrence of different edges in G_F and H_F is not fully independent and techniques from the area of standard random graphs cannot be applied without further consideration. From now on we restrict attention to G_F, of course everything applies also to H_F. We collect some basics about G_F.

An edge (a_1, b_1)——(a_2, b_2) in G_F is only possible if $a_1 \neq a_2$ or $b_1 \neq b_2$. We take a look at the structure of the clause sets which induce the fixed edge (a_1, b_1)——(a_2, b_2). The edge (a_1, b_1)——(a_2, b_2) is in G_F iff F contains at least one of the pairs of clauses $a_1 \vee a_2 \vee z$ and $b_1 \vee b_2 \vee \neg z$ or one of the pairs $a_2 \vee a_1 \vee z$ and $b_2 \vee b_1 \vee \neg z$ for a variable z.

Case 1: $a_1 \neq a_2$ and $b_1 \neq b_2$. In this case all z-clauses are different and all $\neg z$-clauses, too. As the z and $\neg z$ clauses are all different, too, we have $2n$ disjoint pairs of clauses which induce the edge (a_1, b_1)——(a_2, b_2).

Case 2: $a_1 = a_2$ and $b_1 \neq b_2$. In this case the clauses $a_1 \vee a_2 \vee z$ are all different. However $a_1 \vee a_2 \vee z = a_2 \vee a_1 \vee z$. The $\neg z$-clauses are all different and also the z- and $\neg z$-clauses. We have altogether $2n$ pairs of clauses where always two pairs have the common clause $a_1 \vee a_2 \vee z$. The last case $a_1 \neq a_2$ and $b_1 = b_2$ is analogous to the second case.

With these observations we can get a first impression of the probability of a fixed edge in G_F: If $a_1 \neq a_2$ and $b_1 \neq b_2$ the number of pairs of clauses which induce the edge (a_1, b_1)——(a_2, b_2) is distributed as $\mathrm{Bin}(2n,\, p^2)$. The probability that the edge is induced by two pairs of clauses is at most $\binom{2n}{2} \cdot p^4 = o(2np^2)$. This makes it intuitively clear that the probability of (a_1, b_1)——(a_2, b_2) being in G_F is about $2n \cdot p^2$.

If $a_1 = a_2$ and $b_1 \neq b_2$ we have that the number of clauses like $b_1 \vee b_2 \vee \neg z$ or $b_2 \vee b_1 \vee \neg z$ is distributed as $\mathrm{Bin}(2n,\, p)$. The probability of having at least two of these clauses is $O(n^2p^2) = o(2np)$. Restricting to the occurrence of at least one of these clauses it becomes intuitively clear that the probability of the edge (a_1, b_1)——(a_2, b_2) should also be about $2n \cdot p^2$.

Lemma 9. *We fix the edge* $e = (a_1, b_1)$——(a_2, b_2).
(a) For $a_1 \neq a_2$ *and* $b_1 \neq b_2$ *we have that*

$$Pr[F;\, e \text{ is an edge of } G_F] \;=\; 2n \cdot p^2 \cdot (1 \,+\, O(\frac{1}{n^{1+2\gamma}})).$$

(b) For $a_1 = a_2$ *and* $b_1 \neq b_2$ *this probability is*

$$2n \cdot p^2 \cdot (1 \,+\, O(\frac{1}{n^{1+\gamma}})).$$

The same applies of course to $a_1 \neq a_2$ *and* $b_1 = b_2$.

Proof. (a) Recalling the considerations just before this lemma we have for the probability that G_F has the edge e

$$Pr[F; e \text{ is an edge of } G_F] \;=\; 1 \,-\, (1 \,-\, p^2)^{2n}.$$

Using the binomial formula and simple further estimates like $\binom{2n}{i}(-p^2)^i \leq (2np^2)^i$ we get

$$\begin{aligned}
&1 \,-\, (1 \,-\, p^2)^{2n} \\
&= 1 \,-\, 1 \,+\, 2n \cdot p^2 \,-\, \sum_{i=2}^{2n} \binom{2n}{i} (-\,p^2)^i \\
&\geq 2n \cdot p^2 \,-\, \frac{4}{n^{2+4\gamma}} \cdot \sum_{i \geq 0} \left(\frac{2}{n^{1+2\gamma}}\right)^i.
\end{aligned}$$

In the same way

$$1-(1-p^2)^{2n} \leq 2n\cdot p^2 \;+\; \frac{4}{n^{2+4\gamma}}\cdot\sum_{i\geq 0}\Big(\frac{2}{n^{1+2\gamma}}\Big)^i$$

and the convergence of the geometric series and $2n\cdot p^2 \;=\; 2/n^{1+2\gamma}$ imply the claim.

(b) We have: The edge e is not in G_F iff for no z holds that F has the clause $a_1 \vee a_2 \vee z$ and one of the clauses $b_1 \vee b_2 \vee \neg z$ or $b_2 \vee b_1 \vee \neg z$. For a given z we get:

$Pr[F;\; F \;\text{ has }\; a_1 \vee a_2 \vee z \;\text{ and one of }\; b_1 \vee b_2 \vee \neg z,\;\; b_2 \vee b_1 \vee \neg z]$
$= \; p\cdot(2p-p^2) \;=\; 2p^2\cdot(1-(p/2))$.

As these triples of clauses are all disjoint for different z we get that the probability that the edge e is not in G_F is $(1-2p^2\cdot(1-(p/2)))^n$ and

$Pr[e \;\text{ is in }\; G_F]$
$= \; 1 \;-\; (1-2p^2\cdot(1-(p/2)))^n$
$\geq \; n\cdot 2p^2\cdot(1-(p/2)) \;-\; \frac{4}{n^{2+4\gamma}}\sum_{i\geq 0}\Big(\frac{2}{n^{1+2\gamma}}\Big)^i$
$= \; 2n\cdot p^2\cdot(1 \;+\; O(\frac{1}{n^{1+\gamma}}))$.

In the same way we get the upper bound

$$Pr[F;\; e \,\text{is in }\; G_F] \;\leq\; 2n\cdot p^2\cdot(1 \;+\; O(\frac{1}{n^{1+\gamma}}))$$

and the claim holds. □

The preceding lemma implies the following expectations:

Corollary 10. *(a) Let the random variable X denote the number of edges of G_F. For the expectation of X we get*

$$EX \;\;=\;\; n^{3-2\gamma}\cdot(1 \;+\; O(\frac{1}{n})).$$

(b) Let $X_{(a_1,b_1)}$ be the degree of the vertex (a_1,b_1) in G_F, then

$$E[X_{(a_1,b_1)}] \;\;=\;\; 2n^{1-2\gamma}\cdot(1+O(\frac{1}{n})).$$

Proof. (a) The expected number of edges like (a_1, b_1)——(a_2, b_2) with $a_1 \neq a_2$ and $b_1 \neq b_2$ in G_F is with Lemma 9 (a), as $2n \cdot p^2 = 2/n^{1+2\gamma}$,

$$\begin{aligned}
&\frac{n^2 \cdot (n-1)^2}{2} \cdot \frac{2}{n^{1+2\gamma}} \cdot (1 + O(\frac{1}{n^{1+2\gamma}})) \\
&= (n^{3-2\gamma} - 2n^{2-2\gamma} + n^{1-2\gamma}) \cdot (1 + O(\frac{1}{n^{1+2\gamma}})) \\
&= n^{3-2\gamma} \cdot (1 + O(\frac{1}{n}))
\end{aligned}$$

as $\gamma > 0$. For the expected number of edges (a_1, b_1)——(a_2, b_2) where $a_1 = a_2$ (and $b_1 \neq b_2$) we get with Lemma 9 (b)

$$\begin{aligned}
&\frac{n^2 \cdot (n-1)}{2} \cdot \frac{2}{n^{1+2\gamma}} \cdot (1 + O(\frac{1}{n^{1+\gamma}})) \\
&= (n^{2-2\gamma} - n^{1-2\gamma}) \cdot (1 + O(\frac{1}{n^{1+\gamma}})) \\
&= n^{2-2\gamma} \cdot (1 + O(\frac{1}{n})).
\end{aligned}$$

The same applies of course to these edges when $a_1 \neq a_2$ and $b_1 = b_2$. As $n^{2-2\gamma} = n^{3-2\gamma} \cdot 1/n$ the claim follows.

(b) Fixing the vertex (a_1, b_1) the number of edges (a_1, b_1)——(a_2, b_2) with $a_1 \neq a_2$ and $b_1 \neq b_2$ altogether is $(n-1)^2$ and for the expected number of such edges we get, as $2n \cdot p^2 = 2/n^{1+2\gamma}$,

$$\begin{aligned}
&(n-1)^2 \cdot \frac{2}{n^{1+2\gamma}} \cdot (1 + O(\frac{1}{n^{1+2\gamma}})) \\
&= (2n^{1-2\gamma} - \frac{4}{n^{2\gamma}} + \frac{2}{n^{1+2\gamma}}) \cdot (1 + O(\frac{1}{n^{1+2\gamma}})) \\
&= 2n^{1-2\gamma} \cdot (1 + O(\frac{1}{n})).
\end{aligned}$$

The number of edges (a_1, b_1)——(a_2, b_2) where $a_1 = a_2$ and $b_1 \neq b_2$ altogether is $n-1$ and for the expected number of these edges we get with Lemma 9 (b)

$$(n-1) \cdot \frac{2}{n^{1+2\gamma}} \cdot (1 + O(\frac{1}{n^{1+\gamma}})) = O(\frac{1}{n^{2\gamma}}) = n^{1-2\gamma} \cdot O(\frac{1}{n}).$$

Of course the same applies to these edges where $a_1 \neq a_2$ and $b_1 = b_2$. □

Observe that $n^2 \cdot 2n^{1-2\gamma} = 2 \cdot n^{3-2\gamma}$ reflecting the fact that the sum of the degrees of all vertices is two times the number of edges. The number of vertices is n^2 and the probability that a given edge occurs is $\approx 2/n^{1+2\gamma}$. Disregarding edge dependencies G_F is a random graph $G_{n^2,p'}$ where $p' \approx 2/n^{1+2\gamma} = 2n^{1-2\gamma}/n^2$. Note that the exponent is $1-2\gamma > 0$ as $0 < \gamma < 1/2$ The analogous situation in standard random graphs is the probability space $G_{n,p'}$ where $p' = n^{\varepsilon}/n$.

For random graphs according to $G_{n,p'}$ the degree of any vertex is with high probability concentrated at its expectation. This follows easily from tail bounds for the binomial distribution. In case of our graphs G_F and H_F there possibly is some overlap between clauses inducing edges. This entails a weak kind of stochastic dependency between the occurrence of different edges. Nevertheless the following concentration result holds.

Theorem 11. *With high probability we have for* all *vertices* (a_1, b_1) *of* G_F *that the degree of* (a_1, b_1) *is of* $2 \cdot n^{1-2\gamma}(1 + o(1))$.

Proof. Let the variables a_1, b_1 be given. For a variable z let X_z be the indicator random variable on the probability space $\mathrm{Form}_{n,p}$ of the event that there are (non-negated) variables a_2, b_2 such that $a_1 \vee a_2 \vee z \in F$ and $b_1 \vee b_2 \vee \neg z \in F$. Let Y_z indicate the event, $a_2 \vee a_1 \vee z \in F$ and $b_2 \vee b_1 \vee \neg z \in F$. Let $X = \sum X_z$ and Y the same for the Y_z. We show two points: The degree of the vertex (a_1, b_1) of G_F is with probability at least $1 - o(1/n^2)$ equal to $(X + Y) \cdot (1 + o(1))$. Moreover, with probability $1 - o(1/n^2)$ the random variable $X + Y$ is $2 \cdot n^{1-2\gamma}(1 + o(1))$.

First we compute the value of $X + Y$. For a given variable z we have

$$
\begin{aligned}
&Pr[X_z = 1] \\
&= (1 - (1 - p)^n) \cdot (1 - (1 - p)^n) \\
&= 1 - 2 \cdot (1 - p)^n + (1 - p)^{2n}
\end{aligned}
$$

as $1 - (1-p)^n$ is the probability that at least for one a_2 the clause $a_1 \vee a_2 \vee z$ occurs in a random formula. As $0 \leq \gamma < 1/2$ we get from the binomial theorem and the formula for the geometric series that

$$(1-p)^n = 1 - 1/n^\gamma + 1/(2 \cdot n^{2\gamma}) + O(1/n^{3\gamma})$$

and

$$(1-p)^{2n} = 1 - 2/n^\gamma + 2/n^{2\gamma} + O(1/n^{3\gamma}).$$

Plugging these values into the formula for $\Pr[X_z = 1]$ we get

$$Pr[X_z = 1] = 1/n^{2\gamma} + O(1/n^{3\gamma}) = 1/n^{2\gamma} \cdot (1 + o(1)),$$

as we can assume that $\gamma > 0$.

As distinct X_z refer to disjoint sets of clauses, the X_z are stochastically independent and X follows the binomial distribution $\mathrm{Bin}(n\,,\, 1/n^{2\gamma} \cdot (1+o(1)))$. Therefore the expectation of X is $n^{1-2\gamma} \cdot (1 + o(1))$. As $\gamma < 1/2$ we have $1 - 2\gamma > 0$ and exponential tail bounds for the binomial distribution imply that the random variable X is with probability at least $1 - o(1/n^2)$ of $n^{1-2\gamma} \cdot (1 + o(1))$. Of course everything holds in the same way for Y and the required concentration result for $X + Y$ holds.

We come to the degree. The degree of $(a_1,\, b_1)$ in the graph G_F is the number of distinct vertices $(a_2,\, b_2) \neq (a_1,\, b_1)$ for which there is a variable z such that $a_1 \vee a_2 \vee z \in F$ and $b_1 \vee b_2 \vee \neg z \in F$ or $a_2 \vee a_1 \vee z \in F$ and $b_2 \vee b_1 \vee \neg z \in F$. We show that there exists a constant $C = C(\gamma)$ such that with probability at most $o(1/n^2)$ the degree of $(a_1,\, b_1)$ differs additively from the random variable $X + Y$ by more than this constant C. We need to analyze the cases in which a variable z with $X_z = 1$ or $Y_z = 1$ induces either no or strictly more than one additional edge incident with $(a_1,\, b_1)$.

Case 1. We have $X_z = 1$ but no additional edge is induced. This can only happen if $X_z = 1$ due to the clauses $a_1 \vee a_1 \vee z,\ b_1 \vee b_1 \vee \neg z \in F$ or if the edge induced by $a_1 \vee a_2 \vee z,\ b_1 \vee b_2 \vee \neg z \in F$ is also induced by another set of two clauses from F.

The first case occurs for at most C variables z with probability $1 - o(1/n^2)$. For constant C the expected number of sets of C variables z such that the two clauses above are in a random F is at most

$$n^C \cdot (1/n^{1+\gamma})^{2C} = 1/n^{(1+2\gamma)\cdot C}$$

and for $C = 2$, recalling $\gamma > 0$, this expectation is $o(1/n^2)$. By Markov's inequality the probability that there are at least 2 variables z accompanied by clauses as above in a random formula F is $o(1/n^2)$. The case $Y_z = 1$ goes in the same way.

The second case is slightly more complex to deal with, but follows with the same principle. First we consider the case that no additional edge is induced by the clauses $a_1 \vee a_2 \vee z,\ b_1 \vee b_2 \vee \neg z \in F$ due to a *disjoint* set of two clauses which yields the same edge. That is there exists an $z' \neq z$ such that $a_1 \vee a_2 \vee z',\ b_1 \vee b_2 \vee \neg z' \in F$ or there is a z' ($z' = z$ may be possible) such that $a_2 \vee a_1 \vee z',\ b_2 \vee b_1 \vee \neg z' \in F$. For a suitably chosen constant C this situation occurs for C variables z with probability $o(1/n^2)$. The expected number of sets of C variables z such that for each of these z one of the preceding possibilities occurs in a random F is bounded above by

$$n^C \cdot n^C \cdot n^{2C} \cdot 2^C \cdot (1/n^{1+\gamma})^{4C} = 2^C/n^{4C\gamma}.$$

For $C \geq 1$ and $\gamma > 0$ this expectation is $o(1)$. Picking $C > 1/(2\gamma)$ makes this expectation $o(1/n^2)$ and the result follows with Markov's inequality.

It may also be the case that no additional edge is induced due to a set of two clauses which however now is *not disjoint* to the two clauses $a_1 \vee a_2 \vee z,\ b_1 \vee b_2 \vee \neg z \in F$. In this case we must have $a_1 = a_2$ (or the same for the b_i's) and we have the clause $b_2 \vee b_1 \vee \neg z \in F$.

Therefore we have now $Y_z = 1$. For the expectation as before we get

$n^C \cdot n^{2C} \cdot 2^C \cdot (1/n^{1+\gamma})^{3C} = 1/n^{3C\gamma}$
which is $o(1/n^2)$ for $C \geq 1/(3\gamma)$ and $\gamma > 0$.

Case 2. We consider the situation $X_z = 1$ and strictly more than one edge is induced by the clauses with z $\neg z$ in the third position. We show that the number of edges incident with (a_1, b_1) induced by clauses with the fixed z, $\neg z$ in the end can be bounded above by a constant. For the appropriate expectation we get

$$n^C \cdot n^{C'} \cdot 2^C \cdot 2^{C'} \cdot (1/n^{1+\gamma})^{C+C'} = O(1/n^{\gamma \cdot (C+C')})$$

and picking $C + C'$ large enough we get that at least $C \cdot C'$ edges are induced due the clauses with z ,$\neg z$ in the end with probability $o(1/n^2)$.

We still need to bound the number of z's altogether such that strictly more than one edge is induced due to clauses with z, $\neg z$ in the end. The appropriate expectation is

$$n^C \cdot n^{3C} \cdot 2^C \cdot (1/n^{1+\gamma})^{3C} = 1/n^{(3\gamma-1)\cdot C}$$

which is $o(1)$ if $\gamma > 1/3$ which we can safely assume and $C \geq 1$. Picking $C > 2/(3\gamma - 1)$ the expectation is $o(1/n^2)$.

The claim of the theorem now follows from the preceding considerations. Picking C as the sum of all possible deviations from exactly one edge being induced by $X_z = 1$ or $Y_z = 1$ should do. □

2.2 Spectral considerations

In this section we prove a general relationship between the size of an independent set in a graph and the eigenvalues of its adjacency matrix. Then we prove that the random graphs G_F and H_F satisfy certain eigenvalue bounds with high probability. These eigenvalue bounds certify that the graphs G_F and H_F do not have independent

sets as required by Theorem 8 in order to be satisfiable. Again background from spectral graph theory can be found for regular graphs in [AlSp92] and for the general case in [Ch97]. The linear algebra required is well presented in [St88].

Let $G = (V, E)$ be an undirected graph and $A = A_G$ the adjacency matrix of G. Let A's eigenvalues be ordered as $\lambda_1(A) \geq \cdots \geq \lambda_n(A)$, with $n = |V|$. We abbreviate $\lambda_i = \lambda_i(A)$. We say that G is *ν-separated* if $|\lambda_i| \leq \nu\lambda_1$ for $i > 1$. With $\lambda = \max_{i>1} |\lambda_i|$ this reads $\lambda \leq \nu\lambda_1$. We say that G is *ϵ-balanced* for some $\epsilon > 0$ if there is a real d such that the degree of each vertex is between $d(1-\epsilon)$ and $d(1+\epsilon)$. As opposed to Lemma 4 the subsequent theorem only makes use of the eigenvalues of the adjacency matrix of the graph considered.

Theorem 12. *If G is ν-separated and ϵ-balanced, then G contains no independent set of size $> (n/5) + n \cdot f(\nu, \epsilon)$ where $f(\nu, \epsilon)$ tends to 0 as ν, ϵ tend to 0.*

We remark that this theorem can probably be greatly improved upon (see the remark in the proof). But this weak theorem does preclude independent sets of size $n/4$ for small ν, ϵ, and that is all we need here.

Proof. Let S be an independent subset of vertices of G. We will bound $|S|$. Let $T = V \setminus S$. Let χ_S, χ_T be the characteristic functions (represented as column vectors) of S, T respectively (i.e. taking the value 1 on the set and 0 outside of the set). As S is an independent set and G is ε-balanced, we have

$$d(1-\epsilon)|S| \leq \Big| \text{ edges leaving } S \Big| = < A_G\chi_S, \chi_T > . \qquad (1)$$

Note that $A_G\chi_S$ is the column vector whose i'th entry is the number of edges going from vertex i into the set S. Recall that $T = V \setminus S$ and $< \cdots, \cdots >$ is the standard inner product of two vectors. We show further below that

$$< A_G\chi_S, \chi_T > \leq \; d(1+\varepsilon) \cdot (1/2 \; + \; \nu) \cdot \sqrt{|S||T|}. \qquad (2)$$

Abbreviating $\theta = |S|/n$ we get from (1) and (2):

$|S| \leq \frac{1+\varepsilon}{1-\varepsilon}(1/2 + \nu) \cdot \sqrt{|S||T|}$.
This implies that

$$\sqrt{\frac{|S|}{|T|}} \leq \frac{1+\varepsilon-\varepsilon+\varepsilon}{1-\varepsilon}\left(\frac{1}{2}+\nu\right)$$
$$= \left(1 + \frac{2\varepsilon}{1-\varepsilon}\right)\left(\frac{1}{2} + \nu\right)$$
$$= 1/2 + g(\varepsilon, \nu)$$

where $g(\varepsilon, \nu)$ goes to 0 when ε and ν do. Next we get:

$$\frac{\theta}{1-\theta} \leq (1/2 + g(\nu, \varepsilon))^2 = 1/4 + g(\nu, \varepsilon) + g(\nu, \varepsilon)^2$$

We set $f(\nu, \varepsilon) = (4/5)(g + g^2)$ and can easily conclude:

$$\theta \leq (1-\theta)\cdot(1/4) + (1-\theta)\cdot g + (1-\theta)\cdot g^2 \leq (1/4) - \theta\cdot(1/4) + g + g^2$$

$$\Rightarrow \theta \leq (1/4)(4/5) + (4/5)(g+g^2) = 1/5 + f,$$

which is the theorem.

We need to show inequality (2). Let $u_1, \ldots, u_n$ be an orthonormal basis of the n-dimensional vectorspace over the reals where u_i is an eigenvector with eigenvalue λ_i of A_G. We can decompose the adjacency matrix as

$$A_G = \lambda_1 \cdot u_1 \cdot u_1^T + \lambda_2 \cdot u_2 \cdot u_2^T + \cdots + \lambda_n \cdot u_n \cdot u_n^T,$$

where $u_i^T = (u_{i,1}, \ldots, u_{i,n})$ is the transpose of the column vector u_i. Note that $\lambda_i \cdot (u_i \cdot u_i^T) \cdot v = \lambda_i \cdot v$ if $v = \alpha \cdot u_i$ and $\lambda_i \cdot (u_i \cdot u_i^T) \cdot v = 0$ for v orthogonal to u_i. Let

$$\mathcal{E} = A_G - \lambda_1 \cdot u_1 \cdot u_1^T = \sum_{i \geq 2} \lambda_i \cdot u_i \cdot u_i^T.$$

and represent χ_S, χ_T over the basis of the u_i:

$$\chi_S = \sum_{i=1}^{n} \alpha_i \cdot u_i \quad \text{and} \quad \chi_T = \sum_{i=1}^{n} \beta_i \cdot u_i.$$

Recall the fact known as Parseval's equation:

$$|S| = ||\chi_S||^2 = \sum \alpha_i^2 \text{ and } |T| = ||\chi_T||^2 = \sum \alpha_i^2.$$

We get:

$$\begin{aligned}
&< A_G\chi_S, \chi_T > = < \lambda_1 \cdot u_1 \cdot u_1^T \cdot \chi_S + \cdots + \lambda_n \cdot u_n \cdot u_n^T \cdot \chi_S \,,\, \chi_T > \\
&= < (\lambda_1 \cdot (u_1^T \cdot \chi_S)) \cdot u_1 \,,\, \chi_T > + < \mathcal{E} \cdot \chi_S \,,\, \chi_T > \\
&= (\lambda_1 (u_1^T \cdot \chi_S)) \cdot (u_1^T \cdot \chi_T) + < \mathcal{E} \cdot \chi_S \,,\, \chi_T > .
\end{aligned}$$

We bound the two summands separately. Because of the orthornormality of the u_i we get:

$$\begin{aligned}
&< \mathcal{E}\chi_S \,,\, \chi_T > \\
&= < \mathcal{E} \cdot (\alpha_1 u_1 + \cdots + \alpha_n u_n) \,,\, \chi_T > \\
&= < \lambda_2\alpha_2 u_2 + \cdots + \lambda_n\alpha_n u_n \,,\, \beta_1 u_1 + \cdots + \beta_n u_n > \\
&= < \lambda_2\alpha_2 u_2 \,,\, \beta_2 u_2 > + < \lambda_3\alpha_3 u_3 \,,\, \beta_3, u_3 > + \cdots + < \lambda_n\alpha_n u_n \,,\, \beta_n u_n > \\
&= \lambda_2\alpha_2\beta_2 + \lambda_3\alpha_3\beta_3 + \cdots + \lambda_n\alpha_n\beta_n \\
&\leq \lambda \cdot \sum_{i>1} |\alpha_i\beta_i| \\
&\leq \lambda \cdot \sqrt{\sum_{i>1} \alpha_i^2} \cdot \sqrt{\sum_{i>1} \beta_i^2} \\
&\leq \lambda \cdot \sqrt{\sum_{i\geq 1} \alpha_i^2} \cdot \sqrt{\sum_{i\geq 1} \beta_i^2} \\
&= \lambda \cdot \sqrt{|S|} \cdot \sqrt{|T|}, \\
&\leq \nu \cdot d(1+\varepsilon)\sqrt{|S||T|}
\end{aligned}$$

where the last step holds because λ_1 is bounded above by the maximal degree of a vertex, the step before last is Parseval's equation and the third step before last is Cauchy-Schwarz inequality, $\sum |\alpha_i\beta_i| \leq \sqrt{\sum \alpha_i^2} \cdot \sqrt{\sum \beta_i^2}$.

Now we come to the other summand, $(\lambda_1(u_1^T \cdot \chi_S)) \cdot (u_1^T \cdot \chi_T)$. Let α, β be the average values of u_1 on S, T respectively, that is $\alpha =$

$(\sum u_{1,j})/|S|$ where the sum goes over $j \in S$. With the inequality of Cauchy-Schwarz we get:

$$\alpha^2 \;=\; \frac{\left(\sum (u_{1,j} \cdot 1)\right)^2}{|S|^2} \leq \frac{(\sum u_{1,j}^2) \cdot (\sum 1)}{|S|^2} \;=\; \frac{\sum u_{1,j}^2}{|S|}$$

which implies $\alpha^2|S| \leq \sum u_{1,j}^2$. As $T = V \setminus S$ we get

$$\alpha^2|S| \,+\, \beta^2|T| \leq \sum_{j=1}^{n} u_{1,j}^2 \;=\; ||u_1||^2 \;=\; 1.$$

Using the fact that the geometric mean is bounded by the arithmetic mean this implies

$$\alpha\sqrt{|S|} \cdot \beta\sqrt{|T|} \;=\; \sqrt{\alpha^2|S| \cdot \beta^2|T|} \leq \frac{\alpha^2|S| \,+\, \beta^2|T|}{2} \leq \frac{1}{2}.$$

(The weakness of this theorem undoubtedly comes from the pessimistic first estimate, which is only close to the truth when $\alpha^2|S|$ is close to $\beta^2|T|$). This implies as λ_1 is bounded above by the maximum degree that

$$\lambda_1 \cdot (u_1^T \chi_S) \cdot (u_1^T \chi_T) \;\leq\; d(1+\varepsilon)\alpha|S| \cdot \beta|T| \;\leq\; (1/2) \cdot d(1+\varepsilon)\sqrt{|S||T|}$$

and we get (2) finishing the proof. □

We next show that the graphs G_F, and H_F are ν-separated for a small ν. We do this by applying the trace method, see for example [Fr91]. For our purposes it is sufficient to use an elementary version of this method. We first give a general outline of this method and then we apply it to the $G_{n,p}$ model of random graphs. Finally we proceed to the technically more complex graphs G_F, H_F. For $A = A_G$ an adjacency matrix we have from linear algebra that for each $k \geq 0$ $\text{Trace}(A^k) = \sum_{i=1}^{n} \lambda_i^k$. (The trace of a matrix is the sum of the elements on the diagonal.) $\text{Trace}(A^k)$ can be calculated from the underlying graph as:
$\text{Trace}(A^k) = \Big|$ closed walks of length k in the underlying graph $\Big|$. A closed walk of length k in G is a walk like

$$a_0 \overset{e_1}{\text{———}} a_1 \overset{e_2}{\text{———}} a_2 \overset{e_3}{\text{———}} \cdots \overset{e_{k-1}}{\text{———}} a_{k-1} \overset{e_k}{\text{———}} a_k \;=\; a_0.$$

Note that the e_i and a_i need by no means be different, only $a_{i-1} \neq a_i$ as we assume the graph loopless. If k is even we have that all $\lambda_i^k \geq 0$ and we get $\text{Trace}(A^k) = \sum_{i=1}^{n} \lambda_i^k \geq \lambda_1^k + \max_{i>1} \lambda_i^k$. Abbreviating $\lambda = \max_{i>1} |\lambda_i|$ we get further

$$\lambda^k \leq \text{Trace}(A^k) - \lambda_1^k = \sum_{i=2}^{n} \lambda_i^k.$$

If the underlying adjacency matrix A is random this applies in particular to the expected values:

$$E[\lambda^k] \leq E[\text{Trace}(A^k)] - E[\lambda_1^k] = E[\sum_{i=2}^{n} \lambda_i^k]. \tag{3}$$

Now assume that k is an even constant, that n is a variable and sufficiently large (having a model that allows a variable n such as $G_{n,p}$) and that $E[\lambda^k] = o(\lambda_1^k)$ with high probability. Then we have that the graph underlying A is ν-separated for any fixed $\nu > 0$ with high probability which is easy to see:

$$Pr[\lambda > \nu\lambda_1] = Pr[\lambda^k > (\nu\lambda_1)^k] = Pr\left[\lambda^k > \frac{(\nu\lambda_1)^k}{E[\lambda^k]} \cdot E[\lambda^k]\right] \leq o(1)$$

where we apply Markov's inequality , use the fact that $(\nu\lambda_1)^k / E[\lambda^k] = 1/o(1)$ with high probability as k and ν are constant. The last estimate of course implies that for each $\nu > 0$ almost all graphs considered are ν-separated. (The idea considering the k-th power of the eigenvalues seems to be to increase the gap between the largest eigenvalue and the remaining eigenvalues.)

Now we apply this to the $G_{n,p}$ model of random graphs where $p = n^\delta/n$ and $\delta > 0$ is a constant. We calculate the two expected values in equation (3) separately. The largest eigenvalue is always between the minimum and maximum degree of any vertex, as follows from the Courant-Fisher characterization of the eigenvalues of real symmetric matrices. From Chernoff bounds we know that for any $\varepsilon > 0$ with high probability all vertices x from a random G satisfy $(1-\varepsilon)n^\delta \leq$ the degree of $x \leq (1+\varepsilon)n^\delta$. Thus with high probability

$(1-\varepsilon)n^\delta \leq \lambda_1 \leq (1+\varepsilon)n^\delta$. As the probability bounds of the degree estimate follow directly from Chernoff bounds, the probability that the estimate does not hold is exponentially low, that is at most $n \cdot \exp(-\Omega(n^\delta))$. The same probability estimate applies to the estimate for λ_1. (The exponentially small term is for a single fixed vertex, the factor n is necessary as we have n vertices.) As k is even $\lambda_1^k \geq 0$ and

$$E[\lambda_1^k] \geq \left((1-\varepsilon)n^\delta\right)^k (1 - n\exp(-\Omega(n^\delta)) = (1-\varepsilon)^k n^{\delta k} - o(1).$$

Next we come to the expectation of the trace in (3). For $\boldsymbol{a} = (a_0, \ldots, a_{k-1}, a_k = a_0)$ let walk($\boldsymbol{a}$) be the indicator random variable of the event that the walk given by $\boldsymbol{a}$ is possible in a random graph, that is all edges $e_i = (a_{i-1}, a_i) = \{a_{i-1}, a_i\}$ for $1 \leq i \leq k$ occur. Then

$$E[\text{Trace}(A^k)] = E[\sum_{\boldsymbol{a}} \text{walk}(\boldsymbol{a})] = \sum_{\boldsymbol{a}} P[\text{walk}(\boldsymbol{a}) = 1]$$

by linearity of expectation and as the walk($\boldsymbol{a}$) are indicator random variables. To calculate the preceding sum we distinguish three types of possible walks $\boldsymbol{a}$. A walk is *distinct* iff all edges e_i are distinct. A walk is *duplicated* iff each edge among the e_i occurs at least twice. A walk is *quasi-distinct* iff some edges among the e_i occur at least twice and some only once. (Notice that for quasi-distinct walks and duplicated walks, one type does not subsume the other— indeed, quasi-distinct walks must have an edge that occurs only once; furthermore the arguments given below differ essentially for the quasi-distinct and duplicated cases.) For $\boldsymbol{a}$ distinct we have that $Pr[\text{walk}(\boldsymbol{a}) = 1] = p^k$ and the number of $\boldsymbol{a}$'s altogether which can induce a distinct walk is at most n^k because we can choose $a_0, \ldots, a_{k-1}$. Hence, the expected number of distinct walks is bounded above by $n^{\delta k}$. (Compare our estimate for $E[\lambda_1]$.)

For $\boldsymbol{a}$ duplicated we parametrize further with respect to the number j, with $1 \leq j \leq k/2$, of different edges among the e_i. Any walk which is possible at all is generated at least once by the following process: 1. Pick the j positions among k possible positions where each of the j different edges occurs for the first time in $\boldsymbol{e} = (e_1, \ldots, e_k)$:

$\binom{k}{j} \leq k^k$ possibilities. 2. For each of the remaining $k-j$ positions specify which of preceding first occurences of an edge is to be used in this position: $\leq j^{k-j} \leq k^k$ possibilities. 3. Specify the vertices incident with the edges picked in 1: As the j edges from 1. must induce a connected graph the number of possibilities here is at most n^{j+1}. Now we get for the expected number of duplicated walks

$$\sum_{\boldsymbol{a} \text{ duplicated}} Pr[\text{walk}(\boldsymbol{a}) = 1] \leq \sum_{j=1}^{k/2} k^{2k} \cdot n^{j+1} \cdot (n^\delta/n)^j$$

$$\leq \sum_{j=1}^{k/2} k^{2k} \cdot n \cdot n^{\delta j} \leq (k/2) \cdot k^{2k} \cdot n \cdot n^{\delta k/2}.$$

Note that $1 + \delta k/2 < \delta k$ iff $2/\delta < k$. So in order for this estimate to be true k must increase if δ gets smaller.

For the number of quasi-distinct walks we first assume that the last edge, e_k, is a unique edge of the walk. As our walks are quasi-distinct there are at most $1 \leq j \leq k-2$ first occurrences of different edges in $(e_1, \ldots, e_{k-1})$. This implies that the expected number of quasi-distinct walks with the last edge unique is bounded by

$$\sum_{j=1}^{k-2} k^{2k} \cdot n^{j+1} \cdot \left(\frac{n^\delta}{n}\right)^j \cdot \frac{n^\delta}{n} \leq k \cdot k^{2k} \cdot n^{\delta(k-1)}.$$

We account for those quasi-distinct walks where the last edge is not unique by shifting a unique edge to the end and counting as before. As there are k positions where a unique edge might occur we need an additional factor of k. Note that always $\delta(k-1) < \delta k$.

Summing the preceding estimates we get

$$E[\text{Trace}(A^k)] \leq n^{\delta k} + k^2 \cdot k^{2k} \cdot (nn^{\delta k/2} + n^{\delta(k-1)}) = n^{\delta k} + o(n^{\delta k})$$

if $k > 2/\delta$ is an even constant which we assume. Now equation (3) implies

$$E[\lambda^k] \leq (1 - (1-\varepsilon)^k) \cdot n^{\delta k} + o(n^{\delta k})$$

as k even. As $\varepsilon > 0$ can be chosen arbitrarily small, k is constant and the preceding estimate holds whenever n is sufficiently large this means that $E[\lambda^k] = o(n^{\delta k})$. Now fix $\nu > 0$. By the general principle stated above we get that $Pr[\lambda > \nu\lambda_1] = o(1)$ meaning that almost all graphs are ν-separated. Applying Theorem 12 we can for almost all graphs from $G_{n,p}$ efficiently certify that they do not have independent sets with much more than $n/5$ vertices.

The treatment of our graphs G_F, H_F based on the method above is more technical but follows the same principles.

Theorem 13. *For $F \in Form_{n,p,3}$ let $A = A_F$ be the adjacency matrix of G_F and let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{n^2}$ be the eigenvalues of A. Then*

$$E\left[\sum_{i=1}^{n^2} \lambda_i^k\right] \;=\; E[\mathit{Trace}(A_F^k)]$$

$$\leq \left(2n^{1-2\gamma}\right)^k \;+\; c \cdot k^4 \cdot k^{4k} \cdot 2^k \cdot (n^{(1-2\gamma)(k-1)} \;+\; n^2 \cdot n^{(1-2\gamma)k/2}),$$

where c is a constant (possibly $c = 100$ should be enough). If $k > 4/(1-2\gamma)$ the preceding estimate is $(2n^{1-2\gamma})^k \;+\; o\left((2n^{1-2\gamma})^k\right)$.

Proof. For any F we have that $\mathrm{Trace}(A_F) \;=\; \Big|$closed walks of length k in $G_F\Big|$. A typical closed walk of length k is

$$(a_0, b_0)\text{—}(a_1, b_1)\text{—}\cdots\text{—}(a_{k-1}, b_{k-1})\text{—}(a_k, b_k) \;=\; (a_0, b_0)$$

with the only constraint that adjacent vertices (= pairs of propositional variables) are different. Now consider a step (a_{i-1}, b_{i-1})—(a_i, b_i) of this walk. For this step to be possible in G_F the formula F must have one of the following $2n$ pairs of clauses: $a_{i-1} \vee a_i \vee z$, $b_{i-1} \vee b_i \vee \neg z$ for a propositional variable z or the other way round, that is a_i, b_i first. We say that pairs of the first type induce the step (a_{i-1}, b_{i-1})—(a_i, b_i) with sign $+1$ whereas the second type induces this step with sign -1. For two sequences of clauses $\boldsymbol{C} = (C_1, C_2, \ldots, C_k)$ where the last literal of each C_i is a positive literal and $\boldsymbol{D} = (D_1, D_2, \ldots, D_k)$, where the last literal of each D_i is negative, and a a sequence of signs $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_k)$ we say that $\boldsymbol{C}, \boldsymbol{D}, \boldsymbol{\varepsilon}$

induce the walk above iff for each i the pair of clauses C_i, D_i induces the i'th step of the walk with sign given by ε_i. Note that the occurrence of the clauses D_i and C_j in a random F is independent as these clauses are always distinct. We say that F induces the walk above iff we can find sequences of clauses $\boldsymbol{C}, \boldsymbol{D} \subseteq F$ (the C_i, D_i need not necessarily be all distinct) and a sequence of signs $\boldsymbol{\varepsilon}$ such that $\boldsymbol{C}, \boldsymbol{D}, \boldsymbol{\varepsilon}$ induce the given walk. We observe:

- G_F allows for a given walk iff F induces this walk as defined above.

- Three sequences $\boldsymbol{C}, \boldsymbol{D}, \boldsymbol{\varepsilon}$ induce at most one walk, but one walk can be induced by many $\boldsymbol{C}, \boldsymbol{D}, \boldsymbol{\varepsilon}$'s. (Without the $\boldsymbol{\varepsilon}$ it is possible that $\boldsymbol{C}, \boldsymbol{D}$ induce several walks.)

Thus we get that $\mathrm{Trace}(A_F^k)$ can be bounded above by the number of different sequences $\boldsymbol{C}, \boldsymbol{D}, \boldsymbol{\varepsilon}$ with $\boldsymbol{C}, \boldsymbol{D} \subseteq F$ inducing a closed walk of length k and this estimate transfers to the expectation over a random formula F. The notions distinct, quasi-distinct, and duplicated are defined as for graphs: The sequence $\boldsymbol{C}$ is *distinct* iff all component clauses C_i are different. $\boldsymbol{C}$ is *quasi-distinct* iff some C_i's occur at least two times and some only once. $\boldsymbol{C}$ is *duplicated* iff all C_i's which occur in $\boldsymbol{C}$ occur at least twice in $\boldsymbol{C}$. The same notions apply to $\boldsymbol{D}$. We decompose the expected number of $\boldsymbol{C}, \boldsymbol{D}, \boldsymbol{\varepsilon}$'s which induce a closed walk of length k according to all combinations of $\boldsymbol{C}, \boldsymbol{D}$ being distinct, quasi-distinct or duplicated.

The expected number of $\boldsymbol{C}, \boldsymbol{D}, \boldsymbol{\varepsilon}$'s with $\boldsymbol{C}, \boldsymbol{D}$ both distinct can be bounded as follows: The number of possible sequences $\boldsymbol{C}, \boldsymbol{D}$ altogether can be bounded above by n^{3k}. Note that we have unique correspondence between 3 sequences of variables $\boldsymbol{a} = (a_0, \ldots, a_{k-1}, a_k = a_0), \boldsymbol{b} = (b_0, \ldots, b_{k-1}, b_k = b_0)$, and $\boldsymbol{z} = (z_1, \ldots, z_k)$ and two sequences of clauses which might induce a closed walk of length k, $\boldsymbol{C}, \boldsymbol{D}$: $C_i = a_{i-1} \vee a_i \vee z_i$ and $D_i = b_{i-1} \vee b_i \vee \neg z_i$. To account for the signs we get an additional factor of 2^k. The probability that $\boldsymbol{C}, \boldsymbol{D} \subseteq F$ for $\boldsymbol{C}, \boldsymbol{D}$ distinct is $(1/n^{1+\gamma})^{2k}$ and we can bound the expectation by

$$2^k \cdot n^{3k} \left(\frac{1}{n^{1+\gamma}}\right)^{2k} = (2n^{1-2\gamma})^k.$$

Recall that $\gamma < 1/2$ and the degree of each vertex is concentrated at $2n^{1-2\gamma}$. Note also the analogous situation for $G_{n,p}$ above.

For $\boldsymbol{C}$ distinct and $\boldsymbol{D}$ duplicated we parameterize further with respect to the number l with $1 \leq l \leq k/2$ of different clauses D_i. Each $\boldsymbol{D}$ possible at all is generated at least once as follows: 1. Pick l positions among the k possible positions where the different clauses D_i occur for the first time: $\binom{k}{l} \leq k^k$ possibilities. 2. For each of the $k-l$ remaining positions pick one of the preceding positions which were picked in 1: $\leq l^{k-l} \leq k^k$ possibilities. 3. Pick the clauses and the signs for the l positions from 1:$\leq n^{l+1} \cdot n^l \cdot 2^l$ possibilities.(Pick the b_i's, pick the z_i's, pick a sign.) 4. The remaining D_i's are specified through 2. and 3. but perhaps we can pick a sign: 2^{k-j} possibilities. 5. Pick the sequence $\boldsymbol{C}$: n^k possibilities. (We can only choose the a_i's.) For the expectation we get observing that $l \leq k/2 < k$

$$\sum_{l=1}^{k/2} k^{2k} \cdot 2^k \cdot n^{l+1} \cdot n^l \cdot n^k \cdot \left(\frac{1}{n^{1+\gamma}}\right)^{l+k}$$

$$\leq \sum k^{2k} \cdot 2^k \cdot n^{l+1} \cdot \left(\frac{1}{n^{\gamma}}\right)^{2l}$$

$$= \sum_{l=1}^{k/2} k^{2k} \cdot 2^k \cdot n \cdot n^{(1-2\gamma)l} \leq k \cdot k^{2k} \cdot 2^k \cdot n \cdot n^{(1-2\gamma)k/2}$$

where we use $\gamma < 1/2$ and have omitted some of the factors $1/n^{\gamma}$ not necessary. Observe that $1+(1-2\gamma)k/2 < (1-2\gamma)k$ iff $k > 2/(1-2\gamma)$. For this to hold k must become large when γ approaches $1/2$.

Now let $\boldsymbol{C}$ be distinct and $\boldsymbol{D}$ be quasi-distinct. We first consider the case that the last clause from $\boldsymbol{D}$, D_k is unique. We parameterize further with respect to the number l with $1 \leq l \leq k-2$ of different clauses in $(D_1, \ldots, D_{k-1})$. The number of possibilities for $(D_1, \ldots D_{k-1})$ altogether is bounded by: $k^{2k} \cdot n^{l+1} \cdot n^l$. For $D_1, \ldots, D_{k-1}, D_k$ altogether we have only an additional factor of n. For the sign we get 2^k and for $\boldsymbol{C}$ we only have at most n^k possibilities. Thus for the expectation we get noting that always $l+2 \leq k$

$$\sum_{l=1}^{k-2} k^{2k} \cdot 2^k \cdot n^{l+1} \cdot n^l \cdot n \cdot n^k \cdot \left(\frac{1}{n^{1+\gamma}}\right)^{l+k} \cdot \frac{1}{n^{1+\gamma}}$$

$$\leq \sum k^{2k} \cdot 2^k \cdot n^{l+1} \cdot \left(\frac{1}{n^\gamma}\right)^{l+l+2}$$
$$= \sum k^{2k} \cdot 2^k \cdot n^{(1-2\gamma)(l+1)} \leq k \cdot k^{2k} \cdot 2^k \cdot n^{(1-2\gamma)(k-1)}$$

Again we have omitted some unnecessary $1/n^\gamma$'s. To account for the fact that the unique D_i is in between we get an additional factor of k.

Now we come to the case that both $\boldsymbol{C}$ and $\boldsymbol{D}$ are duplicated. We parameterize further with respect to the number j ($1 \leq j \leq k/2$) of different C_i's and $l (1 \leq l \leq k/2)$ of different D_i's. Assume first $j \leq l$, the other case follows symmetrically. The number of different sequences $\boldsymbol{C}, \boldsymbol{D}, \boldsymbol{\varepsilon}$ altogether is bounded by $k^{4k} \cdot 2^k \cdot n^{j+1} \cdot n^j \cdot n^{l+1}$ as we choose the z_i's with $\boldsymbol{C}$. Summing over j, l this gives for the expectation a bound of

$$k^2 \cdot k^{4k} \cdot 2^k \cdot n^2 \cdot n^{(1-2\gamma)k/2}.$$

Note that $\gamma < 1/2$ and $2+(1-2\gamma)k/2 < (1-2\gamma)k$ iff $k > 4/(1-2\gamma)$.

Now we look at the case $\boldsymbol{C}$ duplicated and $\boldsymbol{D}$ quasi-distinct. We first assume that the last clause of $\boldsymbol{D}$ is unique. We parameterize further with respect to j, the number of distinct clauses in $\boldsymbol{C}$, and l, the number of distinct clauses in $(D_1, \ldots, D_{k-1})$. First assume that $j \leq l+1$. The number of possible sequences $\boldsymbol{C}, \boldsymbol{D}, \boldsymbol{\varepsilon}$ altogether is bounded above by $k^{4k} \cdot 2^k \cdot n^{j+1} \cdot n^j \cdot n^{l+1}$ as we choose the z_i's with the j different C_i. For the expectation we get with the sum going over $1 \leq j \leq k/2, 1 \leq l \leq k-2, j \leq l+1$

$$\sum k^{4k} \cdot 2^k \cdot n^{j+1} \cdot n^j \cdot n^{l+1} \cdot \left(\frac{1}{n^{1+\gamma}}\right)^{j+l} \cdot \frac{1}{n^{1+\gamma}}$$
$$\leq \sum k^{4k} \cdot 2^k \cdot n \cdot n^j \cdot \left(\frac{1}{n^\gamma}\right)^{2j}$$
$$\leq k^2 \cdot k^{4k} \cdot 2^k \cdot n \cdot n^{(1-2\gamma)k/2}.$$

Again note that $1+(1-2\gamma)k/2 < (1-2\gamma)k$ for k a sufficiently large constant.

The case $l+1 \leq j$ yields in the same way, now choosing the z_i with the D_i's:

$$\sum k^{4k} \cdot 2^k \cdot n^{l+1} \cdot n^l \cdot n \cdot n^{j+1} \cdot \left(\frac{1}{n^{1+\gamma}}\right)^{j+l} \cdot \frac{1}{n^{1+\gamma}}$$
$$\leq\ k^2 \cdot k^{4k} \cdot 2^k \cdot n^2 \cdot n^{(1-2\gamma)k/2}.$$

For the unique clause of $\boldsymbol{D}$ being somewhere in between we need an additional factor of k.

Now finally we look at the case that $\boldsymbol{C}, \boldsymbol{D}$ are both quasi-distinct. First assume that the two last clauses C_k, D_k are unique and let j be the number of different C_i among the first $k-1$ C_i's and l the same for the D_i's. First assume the $j \leq l$. We get for the expectation

$$\sum_{1 \leq j \leq l \leq k-2} k^{4k} \cdot 2^k \cdot n^{j+1} \cdot n^j \cdot n \cdot n^{l+1} \cdot \left(\frac{1}{n^{1+\gamma}}\right)^{l+j} \cdot \left(\frac{1}{n^{1+\gamma}}\right)^2$$
$$\leq\ k^2 \cdot k^{4k} \cdot 2^k \cdot n^{(1-2\gamma)(k-1)}.$$

For the unique clauses standing at different positions we get the same estimate which altogether accounts for another factor of k^2. For $l \leq j$ we get another factor of 2. □

Now our algorithm is obvious: We pick ε, ν sufficiently small such that the $f(\nu, \varepsilon)$ from Theorem 12 is $< 1/20$. Given $F \in \text{Form}_{n,p}, p = 1/n^{1+\gamma}$, we construct G_F. We check if maximum degree/minimum degree $\leq (1+\varepsilon)/(1-\varepsilon)$. This holds with high probability, in case it does not the algorithm fails. Now we determine λ_1 and λ. With high probability we have that $\lambda \leq \nu\lambda_1$. If the last estimate does not hold, we fail. By Theorem 12 the algorithm now has certified that G_F has no independent set of size $\geq n^2/4$. We do the same for H_F. With high probability we succeed and by Theorem 8 F is certified unsatisfiable.

Conclusion

Our algorithm works with high probability with respect to the binomial space $\text{Form}_{n,p}$ where p is such that the *expected number* of clauses is the announced $n^{3/2+\varepsilon}$. In case we always draw *exactly* $n^{3/2+\varepsilon}$ clauses the algorithm should also work, as can be shown by applying very similar arguments to those given in the paper. Note

that we generate formulas in $\mathrm{Form}_{n,p}$ with exactly the expectation many clauses with probability bounded below only $\Omega(1/\sqrt{n})$. However after generating exactly $n^{3/2+\varepsilon}$ clauses one can formally delete some clauses with the appropriate low probability. This yields a subset having high probability on $\mathrm{Form}_{n,p}$ with an appropriate p and the preceding consideration applies to these new formulas. Thus showing the unsatisfiability of the original formula.

References

[Ac2000] Dimitris Achlioptas. Setting 2 variables at a time yields a new lower bound for random 3-SAT. In Proceedings SToC 2000, ACM, 28–37.

[AcFr99] Dimitris Achlioptas, Ehud Friedgut. A sharp threshold for k-colourability. Random Structures and Algorithms 14, 1999, 63–70.

[AcMo97] Dimitris Achlioptas, Mike Molloy. Analysis of a list colouring algorithm on a random graph. In Proceedings FoCS 1997, IEEE, 204–212.

[AcMo99] Dimitris Achlioptas, Mike Molloy. Almost all graphs with 2.522n edges are not 3-colourable. Electronic J. Combin. 6, 1999, Research paper 29.

[AcMo2002] Dimitris Achlioptas, Cristopher Moore. The Asymptotic Order of the Random k-SAT Threshold. In Proceedings FoCS 2002, IEEE, 779 – 788.

[AcPe2003] Dimitris Achlioptas, Yuval Peres. The Threshold for Random k-SAT is $2^k(\ln 2 + o(1))$. In Proceedings SToC 2003, ACM, 223–231.

[AcSo2000] Dimitris Achlioptas, Gregory B.Sorkin. Optimal myopic algorithms for random 3-SAT. In Proceedings 41st FoCS 2000, IEEE, 590-600.

[Al98] Noga Alon, Spectral techniques in graph algorithms, In Proceedings LATIN 1998, LNCS 1380, 206–215.

[AlKa94] Noga Alon, Nabil Kahale. A spectral technique for colouring random 3-colorable graphs. In Proceedings SToC 1994, ACM, 346-355.

[AlSp92] Noga Alon, Joel H. Spencer. The probabilistic method. Wiley & Sons Inc, 1992.

[Be et al98] Paul Beame, Richard Karp, Toniann Pitassi, Michael Saks. On the complexity of unsatisfiability proofs for random k-CNF formulas. In Proceedings 30th SToC 1998, ACM, 561-571.

[BePi96] Paul Beame, Toniann Pitassi. Simplified and improved resolution lower bounds. In Proceedings FoCS 1996, IEEE, 274-282.

[Bl et al81] Manuel Blum, Richard Karp, Oliver Vornberger, Christos H. Papadimitriou, Mihalis Yannakakis. The complexity of testing whether a graph is a superconcentrator. Information Processing Letters 13, 1981, 164-167.

[Bo85] Bela Bollobás. Random Graphs. Academic Press, 1985.

[BoChPi2001] Christoph Borgs, Jennifer Chayes, Boris Pittel. Sharp threshold and scaling window for the integer partitioning problem, In Proceedings 33rd SToC 2001, ACM, 330-336.

[Ch97] Fan R. K. Chung. Spectral Graph Theory. American Mathematical Society, 1997.

[ChRe92] Vašek Chvatal, Bruce Reed. Mick gets some (the odds are on his side). In Proceedings 33nd FoCS 1992, IEEE, 620-627.

[ChSz88] Vašek Chvatal, Endre Szemerédi. Many hard examples for resolution. Journal of the ACM 35(4), 1988, 759-768.

[CoGoLaSch2003] Amin Coja-Oghlan, Andreas Goerdt, André Lanka, Frank Schädlich. Certifying Unsatisfiability of Random $2k$-SAT Formulas using Approximation Techniques. In Proceedings FCT 2003, LNCS.

[CrAu96] J. M. Crawford, L. D. Auton. Experimental results on the crossover point in random 3SAT. Artificial Intelligence 81, 1996, 31–57.

[DuBoMa2000] Olivier Dubois, Y. Boufkhad, Jacques Mandler. Typical Random 3-SAT Formulae and the Satisfiability Threshold. In Proceedings 11th SoDA, 2000, SIAM, 126 – 127.

[DuZi98] Paul E. Dunne, Michele Zito. An improved upper bound for the non-3-colourability threshold. Information Processing Letters 65, 1998, 17–23.

[Fr91] Joel Friedman. On the second eigenvalue and random walks in random d-regular graphs. Combinatorica 11(4),1991, 331 - 362.

[Fr99] Ehud Friedgut. 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.

[FrSu96] Alan M. Frieze, Stephen Suen. Analysis of two simple heuristics on a random instance of k-SAT. Journal of Algorithms 20(2), 1996, 312-355.

[Fu95] Xudong Fu. On the complexity of proof systems. PhD Thesis, University of Toronto, 1995.

[FuKo81] Z. Füredi, J. Komlós. The eigenvalues of random symmetric matrices. Combinatorica 1(3), 1981, 233-241.

[Go96] Andreas Goerdt. A threshold for unsatisfiability. Journal of Computer and System Sciences 53, 1996, 469-486.

[GoLa2003] Andreas Goerdt, André Lanka. Recognizing more random unsatisfiable 3-SAT instances efficiently. In Proceedings Workshop on Typical Case Complexity and Phase Transitions. Satellite Workshop to LiCS 2003, Ottawa.

[GoKr2000] Andreas Goerdt, Michael Krivelevich. Efficient recognition of random unsatisfiable k-SAT instances by spectral methods. In Proceedings STACS 2001, LNCS 2010, 294–304.

[ImNa96] Russell Impagliazzo, Moni Naor. Efficient cryptographic schemes provably as secure as subset sum. Journal of cryptology 9, 1996, 199-216.

[Ju82] Ferenc Juhász. The asymptotic behaviour of Lovász theta function for random graphs. Combinatorica 2(2), 1982, 153-155.

[KaKiLa2002] Alexis C. Kaporis, Lefteris M. Kirousis, Efthimios G. Lalas. The Probabilistic Analysis of a Greedy Satisfiability Algorithm. In Proceedings 10th ESA, 2002, LNCS, 574 – 585.

[KiKrKr98] Lefteris M. Kirousis, Evangelos Kranakis, Danny Krizanc, Yiannis Stamatiou. Approximating the unsatisfiability threshold of random formulas. Random Structures and Algorithms 12(3), 1998, 253-269.

[KrVu2000] Michael Krivelevich, Van H. Vu. Approximating the independence number and the chromatic number in expected polynomial time. In Proceedings ICALP 2000, LNCS 1853, 13-24.

[PeWe89] A. D. Petford, Dominic Welsh. A Randomised 3-colouring algorithm. Discrete Mathematics 74, 1989, 253-261.

[Sch89] Uwe Schöning. Logic for Computer Scientists. Birkhäuser, Boston, 1989

[SeMiLe96] Bart Selman, David G. Mitchell, Hector J. Levesque. Generating hard satisfiability problems. Artificial Intelligence 81(1-2), 1996, 17-29.

[St88] Gilbert Strang. Linear Algebra and its Applications. Harcourt Brace Jovanovich, Publishers, San Diego, 1988.