Combinatorics, Probability and Computing (2009) 18, 775–801.
doi:10.1017/S0963548309990356 Printed in the United Kingdom

On the Random Satisfiable Process

MICHAEL KRIVELEVICH[1†], BENNY SUDAKOV[2‡]
and DAN VILENCHIK[3]

[1]School of Mathematical Sciences, Tel Aviv University, Tel Aviv, Israel
(e-mail: krivelev@post.tau.ac.il)

[2]Department of Mathematics, UCLA, Los Angeles, USA
(e-mail: bsudakov@math.ucla.edu)

[3]Department of Mathematics, UCLA, Los Angeles, USA
(e-mail: danny.vilenchik@gmail.com)

Received 28 July 2008; revised 21 June 2009

In this work we suggest a new model for generating random satisfiable k-CNF formulas. To generate such formulas. randomly permute all $2^k\binom{n}{k}$ possible clauses over the variables $x_1,\ldots,x_n$, and starting from the empty formula, go over the clauses one by one, including each new clause as you go along if, after its addition, the formula remains satisfiable. We study the evolution of this process, namely the distribution over formulas obtained after scanning through the first m clauses (in the random permutation's order).

Random processes with conditioning on a certain property being respected are widely studied in the context of graph properties. This study was pioneered by Ruciński and Wormald in 1992 for graphs with a fixed degree sequence, and also by Erdős, Suen and Winkler in 1995 for triangle-free and bipartite graphs. Since then many other graph properties have been studied, such as planarity and H-freeness. Thus our model is a natural extension of this approach to the satisfiability setting.

Our main contribution is as follows. For $m \geqslant cn$, $c = c(k)$ a sufficiently large constant, we are able to characterize the structure of the solution space of a typical formula in this distribution. Specifically, we show that typically all satisfying assignments are essentially clustered in one cluster, and all but $e^{-\Omega(m/n)}n$ of the variables take the same value in all satisfying assignments. We also describe a polynomial-time algorithm that finds w.h.p. a satisfying assignment for such formulas.

1. Introduction

Constraint satisfaction problems play an important role in many areas of computer science, *e.g.*, computational complexity theory [10], coding theory [16], and artificial intelligence

† Research supported in part by a USA–Israel BSF grant, by a grant from the Israel Science Foundation, and by the Pazy Memorial Award.
‡ Research supported in part by NSF CAREER award DMS-0812005 and a USA–Israel BSF grant.

[24], to mention just a few. The main challenge is to devise efficient algorithms for finding satisfying assignments (when such exist), or conversely to provide a certificate of unsatisfiability. One of the best-known examples of a constraint satisfaction problem is k-SAT, which is the first to be proved to be NP-complete. Although satisfactory approximation algorithms are known for several NP-hard problems, the problem of finding a satisfying assignment (if such exists) is not among them. In fact, Håstad [18] proved that it is NP-hard to approximate MAX-3SAT (the problem of finding an assignment that satisfies as many clauses as possible) within a ratio better than 7/8.

In trying to understand the inherent hardness of the problem, many researchers analysed structural properties of formulas drawn from different distributions. One such distribution is the *uniform* distribution, where instances are generated by picking m clauses uniformly at random out of all $2^k\binom{n}{k}$ possible clauses. Although many problems still remain unsolved, in general this distribution seems to be quite well understood (at least for some values of m and k). This is also true for the *planted k-SAT* model, where one first fixes some assignment ψ to the variables and then picks m clauses uniformly at random out of all $(2^k-1)\binom{n}{k}$ clauses satisfied by ψ. Comparatively, much less is known for variants of these distributions where extra conditions are imposed. These conditions distort the randomness in such a way that the 'standard' methods and tools employed to analyse the original distributions are *a priori* of little use in the new setting. Our work concerns the latter.

1.1. Our contribution

In this work we suggest a new model for generating random satisfiable k-CNF formulas. To generate such formulas, randomly permute all $2^k\binom{n}{k}$ possible clauses over the variables $x_1,\ldots,x_n$, and, starting from the empty formula, go over the clauses one by one, including each new clause as you go along if, after its addition to the formula, the formula remains satisfiable. We study the evolution of this process, namely the distribution over formulas obtained after scanning through the first m clauses (in the random permutation's order); we use $\mathcal{P}^{\text{sat}}_{n,m}$ to denote this distribution. Clearly, for every m, all formulas in $\mathcal{P}^{\text{sat}}_{n,m}$ are satisfiable (as every clause is included only if the formula obtained so far remains satisfiable).

Random processes with conditioning on a certain property being respected are widely studied in the context of graph properties. This study was pioneered by Ruciński and Wormald in 1992 [25] for graphs with a fixed degree sequence, and also by Erdős, Suen, and Winkler in 1995 for triangle-free and bipartite graphs [11]. Since then, many other graph properties have been studied, such as planarity [17], H-freeness [23], and also the property of being intersecting in the context of hypergraphs [6]. Thus our model is a natural extension of this approach to the satisfiability setting. The main difficulty when dealing with these restricted processes is that the edges of the random graph (and the clauses of the random k-CNF formula) are no longer independent, due to conditioning. Thus the rich methods that have been developed to understand the 'classical' random graph models, $G_{n,p}$ for example, do not carry over, at least not immediately, to the restricted setting.

Quite frequently in restricted random processes, the typical size of a final graph or formula (after all edges/clauses have been scanned) is a fascinating subject of study. This is, however, *not* the case here, as it is quite easy to see that, deterministically, the final random formula will have $(2^k - 1)\binom{n}{k}$ clauses and a unique satisfying assignment. Therefore, in the setting under consideration here, the process itself (*i.e.*, a typical development of a restricted random formula and of its set of satisfying assignments as the number m of scanned clauses grows) is much more interesting than the final result, and indeed in this paper we will study the development of a random satisfiable formula.

As it turns out, if m is chosen so that almost all k-CNF formulas with m clauses over n variables are satisfiable, then $\mathcal{P}^{\text{sat}}_{n,m}$ is statistically close to the uniform distribution over such formulas, since w.h.p. none of the m clauses will be rejected (by w.h.p. we mean 'with probability tending to 1 as n goes to infinity'). Therefore, if this is the case, then the clauses are practically independent of each other, and the 'usual' techniques apply. Remarkable phenomena occurring in the uniform distribution are *phase transitions*. With respect to the property of being satisfiable, such a phase transition also takes place: there exists a threshold $d = d(n, k)$ such that, for every fixed $\varepsilon > 0$, if $m/n \leqslant d - \varepsilon$, a random formula with m clauses over n variables is w.h.p. satisfiable, and if $m/n \geqslant d + \varepsilon$, a random formula is w.h.p. unsatisfiable [15]. Thus, while $\mathcal{P}^{\text{sat}}_{n,m}$ is statistically close to the uniform distribution for m/n below the threshold, it is not clear what a typical $\mathcal{P}^{\text{sat}}_{n,m}$ instance looks like when crossing this threshold (which is conjectured to be roughly 4.26 for 3SAT), and whether there exists a polynomial-time algorithm for finding a satisfying assignment for such instances.

In this work we analyse $\mathcal{P}^{\text{sat}}_{n,m}$ when m/n is some sufficiently large constant *above* the satisfiability threshold. The first part of our result is characterizing the structure of the solution space of a typical formula in $\mathcal{P}^{\text{sat}}_{n,m}$. By the 'solution space' of a formula we mean the set of all satisfying assignments (which is a subset of all 2^n possible assignments).

Definition 1. A variable is said to be *frozen* in a k-CNF F if, in every satisfying assignment of F, it takes the same value.

We state our results for $k = 3$ and remark that they can be extended to a general fixed $k \geqslant 3$, using arguments similar to those presented in this paper. See Section 7 for more details.

Theorem 1.1. *Let F be random 3CNF from $\mathcal{P}^{\text{sat}}_{n,m}$, $m/n \geqslant c$, for c a sufficiently large constant. Then w.h.p. F enjoys the following properties.*

(1) *All but $e^{-\Omega(m/n)}n$ variables are frozen.*
(2) *The formula induced by the non-frozen variables decomposes into connected components of at most logarithmic size.*
(3) *Letting $\beta(F)$ be the number of satisfying assignments of F, we have $\frac{1}{n}\log\beta(F) = e^{-\Omega(m/n)}$.*

Notice that item (3) in Theorem 1.1 follows directly from item (1). One immediate corollary of this theorem is as follows.

Corollary 1.2. *Let F be random 3CNF from $\mathcal{P}^{\text{sat}}_{n,m}$, $m/n \geqslant c \log n$, for c a sufficiently large constant. Then w.h.p. F has only one satisfying assignment.*

The corollary follows from item (3) in Theorem 1.1 since $e^{-\Omega(m/n)} = o(n^{-1})$ for $m/n \geqslant c \log n$, and therefore $\log \beta(F) = o(1)$, or in turn, $\beta(F) = 1 + o(1)$.

The characterization given by Theorem 1.1 is in sharp contrast to the structure of the solution space of $\mathcal{P}^{\text{sat}}_{n,m}$ formulas with m/n just below the threshold. Specifically, the *conjectured* picture, some supporting evidence of which was proved rigorously for $k \geqslant 8$ [2, 22, 1], is that typically random k-CNF formulas in the near-threshold regime have an exponential number of *clusters* of satisfying assignments. While any two assignments in distinct clusters disagree on at least εn variables, any two assignments within one cluster coincide on $(1 - \varepsilon)n$ variables (for some positive constant $\varepsilon < 1$). Furthermore, each cluster has a linear number of locally frozen variables (that is, frozen with respect to all satisfying assignments within that cluster). This structure seems to make life hard for most known SAT heuristics. One explanation seems to be that the algorithms do not 'steer' into one cluster but rather try to find a 'compromise' between the satisfying assignments in distinct clusters, which is in fact impossible.

Complementing this picture *rigorously*, we show that a typical formula in $\mathcal{P}^{\text{sat}}_{n,m}$ (in the above-threshold regime) can be solved efficiently.

Theorem 1.3. *There exists a deterministic polynomial-time algorithm that w.h.p. finds a satisfying assignment for 3CNF formulas from $\mathcal{P}^{\text{sat}}_{n,m}$, $m/n \geqslant c$, for c a sufficiently large constant.*

Our proof of Theorem 1.3 is constructive in the sense that we explicitly describe the algorithm.

Remark 1. Another natural problem to study is k-colourability. As for random k-CNF formulas, the random graph $G_{n,p}$ also goes through a phase transition with respect to the property of being k-colourable, as np grows. Analogously to the random k-CNF process that we defined, one can consider a restricted random graph process. Specifically, randomly order all $\binom{n}{2}$ edges of the graph, go over them in that order and include each new edge as long as the resulting graph remains k-colourable. Some of the results that we have for k-SAT extend to the k-colourability process. A more thorough discussion is given in Section 6.

1.2. Related work and techniques

Almost all polynomial-time heuristics suggested so far for random instances (either SAT or graph optimization problems) were analysed when the input is sampled according to a planted-solution distribution, or various semi-random variants thereof. Alon and Kahale [3] suggest a polynomial-time algorithm based on spectral techniques that w.h.p. properly k-colours a random graph from the planted k-colouring distribution (the distribution of graphs generated by partitioning the n vertices into k equally sized colour classes, and including every edge connecting two different colour classes with probability $p = p(n)$), for graphs with average degree greater than some constant. In the SAT context, Flaxman's algorithm, drawing on ideas from [3], solves w.h.p. planted 3SAT instances where the clause-variable ratio is greater than some constant. Also, [13, 12, 19] address the planted 3SAT distribution.

On the other hand, very little work was done on non-planted distributions, such as $\mathcal{P}^{\text{sat}}_{n,m}$. In this context one can mention a work of Chen [7], who provides an *exponential*-time algorithm for the uniform distribution over satisfiable k-CNF formulas with exactly m clauses where m/n is greater than some constant. Ben-Sasson, Bilu and Gutfreund [5] also study this distribution but with $m/n = \Omega(\log n)$, a regime where the uniform distribution and the planted distribution essentially coincide (since typically there is only one satisfying assignment), and leave as an open question whether one can characterize the regime $m/n = o(\log n)$. This question was resolved in [9] (and in [8] for the uniform distribution over k-colourable graphs).

While some of the ideas suggested in these works have proved to be instrumental for our setting, most of their analytical methods break when considering $\mathcal{P}^{\text{sat}}_{n,m}$. In $\mathcal{P}^{\text{sat}}_{n,m}$, not only do clauses depend on each other (unlike the planted distribution where clauses are chosen independently), but the order in which they are introduced also plays a role (which is not the case in the uniform distribution studied in [9], although the clauses are not chosen independently). Therefore we had to come up with new analytical tools that might be of interest in other settings as well.

1.3. Structure of the paper

The rest of the paper is structured as follows. In Section 2 we discuss relevant structural properties that a typical formula in $\mathcal{P}^{\text{sat}}_{n,m}$ possesses; the proofs of some properties are postponed to Sections 4 and 5. One consequence of this discussion will be a proof of Theorem 1.1. We then prove Theorem 1.3 in Section 3 by presenting an algorithm and showing that it meets the requirements of Theorem 1.3. In Section 6 we discuss the k-colourability setting (mentioned in Remark 1) more elaborately, and concluding remarks are given in Section 7.

2. Properties of a random $\mathcal{P}^{\text{sat}}_{n,m}$ instance

This section contains the technical part of the paper. In it we analyse the structure of a typical formula in $\mathcal{P}^{\text{sat}}_{n,m}$. Here and throughout we think of m as cn, for c at least some sufficiently large constant.

2.1. Preliminaries and techniques

When analysing some structural properties of a random instance in $\mathcal{P}^{\text{sat}}_{n,m}$, it will be more convenient to analyse the same property under a somewhat different distribution, and then to go back to $\mathcal{P}^{\text{sat}}_{n,m}$ (maybe paying some factor in the estimate).

The variation we consider is $\mathcal{P}^{\text{sat}}_{n,p}$ and is defined as follows. Permute at random all possible $M = 8\binom{n}{3}$ clauses, go over the clauses in the permutation's order, and include each clause with probability $p = m/M$ if its addition also leaves the instance satisfiable. Let $\mathcal{P}_{n,p}$ be defined similarly, just without the conditioning (*i.e.*, all clauses chosen at random are included in the formula, thus making it not necessarily satisfiable).

Throughout we usually stick to the following notation. We shall use F to denote an instance sampled according to $\mathcal{P}_{n,p}$ and F^* to be the ordered subset of clauses of F distributed according to $\mathcal{P}_{n,p}^{\text{sat}}$.

Lemma 2.1. $\mathcal{P}_{n,m}^{\text{sat}} = \mathcal{P}_{n,p}^{\text{sat}} | \{$*exactly m clauses were considered for addition*$\}$.

Proof. To generate a formula according to $\mathcal{P}_{n,m}^{\text{sat}}$, we first pick a random permutation of the clauses and then scan one by one the first m clauses, skipping clauses whose addition will make the instance unsatisfiable. The key point is to notice that any ordered m-tuple of clauses is equally likely to be chosen as the first m clauses. This is exactly the case in $\mathcal{P}_{n,p}^{\text{sat}}$, when conditioning on the fact that exactly m clauses were considered for addition; any set of m clauses is equally likely, and also any permutation of them. □

Lemma 2.2. *Set* $M = 8\binom{n}{3}$. *For any property* A, *if* $p = m/M$ *then* $\Pr^{\mathcal{P}_{n,m}^{\text{sat}}}[A] \leqslant O(\sqrt{m}) \cdot \Pr^{\mathcal{P}_{n,p}^{\text{sat}}}[A]$.

Proof. Let X be a random variable counting the number of clauses that were considered for addition under $\mathcal{P}_{n,p}^{\text{sat}}$. X is distributed $\text{Binom}(8\binom{n}{3}, p)$, and therefore $E[X] = m$. Standard calculations show that $\Pr[X = m] = \Omega(\sqrt{m})$. Then

$$\Pr^{\mathcal{P}_{n,m}^{\text{sat}}}[A] = \Pr^{\mathcal{P}_{n,p}^{\text{sat}}}[A|X = m] = \frac{\Pr^{\mathcal{P}_{n,p}^{\text{sat}}}[A \wedge X = m]}{\Pr^{\mathcal{P}_{n,p}^{\text{sat}}}[X = m]} \leqslant O(\sqrt{m}) \cdot \Pr^{\mathcal{P}_{n,p}^{\text{sat}}}[A].$$ □

Remark 2. In the remainder of the section we analyse $\mathcal{P}_{n,p}^{\text{sat}}$ instead of $\mathcal{P}_{n,m}^{\text{sat}}$. When we use the expression 'with very high probability' (abbreviated w.v.h.p.) we will always mean with probability $1 - o(m^{-1/2})$. Lemma 2.2 will then imply that we can switch back to $\mathcal{P}_{n,m}^{\text{sat}}$ and still the property holds with probability $1 - o(1)$ (the usual interpretation of 'with high probability'). We will in fact prove that all the properties hold with probability $1 - o(n^{-3})$, which is always at least $1 - o(m^{-1/2})$ since $m = O(n^3)$.

2.2. The discrepancy property

A well-known result in the theory of random graphs is that a random graph w.v.h.p. will not contain a small yet unexpectedly dense subgraph. This is also the case for $\mathcal{P}_{n,p}$ (when considering the graph induced by the formula). In general, discrepancy properties play a fundamental role in the proof of many important structural properties such as expansion, the spectra of the adjacency matrix, *etc.*, and indeed in our case the discrepancy property plays a major role both in the algorithmic perspective and in the analysis of the clustering phenomenon. The following discussion rigorously establishes the above-stated fact.

Definition 2. We say that a 3CNF formula F on n variables is *ρ-proportional* if there exists *no* set U of variables such that:

- $|U| \leqslant n/10^6$,
- there are at least $\rho \cdot |U|$ clauses in F each containing at least two variables from U.

(We say that a clause C contains a variable x if x appears in C either as x or as $\bar{x}$. In this context we do not differentiate between the two cases.)

Proposition 2.3. *Let F be distributed according to $\mathcal{P}_{n,p}$ with $n^2p \geqslant d$, for d a sufficiently large constant, and set $\rho = n^2p/5500$. Then w.v.h.p. F is ρ-proportional.*

Remark 3. To see how Proposition 2.3 corresponds to the random graph context, consider the graph induced by the formula F (the vertices are the variables, and two variables are connected by an edge if there exists some clause containing them both) and observe that every clause that contains at least two variables from U contributes an edge to the subgraph induced by U. Thus, if we have many such clauses, this subgraph will be prohibitively dense. Since F is random so is its induced graph, and as we shall see the latter typically does not occur (we do not claim that the graph is random like the Erdős–Rényi random graph, for example, just that its edges are chosen according to some random rule).

Proof. The probability that a random formula F in $\mathcal{P}_{n,p}$ contains a set U of variables of size u that violates proportionality is at most (using the union bound)

$$\sum_{u=1}^{n/10^6} \binom{n}{u} \cdot \binom{8n\binom{u}{2}}{un^2p/5500} \cdot p^{un^2p/5500} = o(n^{-3}).$$

The first term accounts for the possible ways of choosing the variables of U, the second is to choose the $un^2p/5500$ clauses that contain at least two variables from U (out of at most $8n\binom{u}{2}$ possible ones), and the last term is just the probability of the chosen clauses actually appearing in F. To bound this sum we use the fact that $u \leqslant n/10^6$, the fact that n^2p can be arbitrarily large (constant), and the following standard estimate for the binomial coefficient:

$$\binom{n}{x} \leqslant \left(\frac{en}{x}\right)^x. \qquad \square$$

Corollary 2.4. *Let F^* be distributed according to $\mathcal{P}_{n,p}^{\text{sat}}$ with $n^2p \geqslant d$, for d a sufficiently large constant. Then w.v.h.p. F^* is $n^2p/5500$-proportional.*

The corollary follows easily by observing that the proportionality property is monotonically decreasing.

2.3. Crude characterization of the solution space's structure

In this section we make the first step towards proving Theorem 1.1 (clustering). We give a rather crude characterization of the structure of the solution space of a typical instance in $\mathcal{P}_{n,p}^{\text{sat}}$. This characterization will be refined below.

Definition 3. A 3CNF F is called *r-concentrated* if every two satisfying assignments ψ_1, ψ_2 of F are at Hamming distance at most r from each other.

Proposition 2.5. *Let F^* be distributed according to $\mathcal{P}_{n,p}^{\rm sat}$ with $n^2p \geqslant d$, for d a sufficiently large constant, and let $\rho = 30/(n^2p)$. Then w.v.h.p. F^* is ρn-concentrated.*

An immediate corollary of this proposition is that typically all satisfying assignments of a $\mathcal{P}_{n,p}^{\rm sat}$ instance can be enclosed in a ball of radius $30/(np)$ in $\{0,1\}^n$. This gives a 'first-order' characterization of the structure of the solution space.

Proof. Fix two assignments φ and ψ at distance αn, and let us bound Pr[φ and ψ satisfy F^*]. Assume without loss of generality that, say, φ is the all-TRUE assignment. We shall now upper-bound the probability of a set of clauses in $\mathcal{P}_{n,p}$ that may result in an instance F^* that is satisfied by both assignments. In particular a clause of the form $C_1 = (x \vee \bar{y} \vee \bar{z})$, where x is a variable on which φ and ψ disagree, and y, z are variables on which both agree, cannot be chosen to be $\mathcal{P}_{n,p}$. Let us call such a clause a type 1 clause. If a type 1 clause appears then either it is included in F^*, and then ψ cannot be a satisfying assignment, or it is rejected, and then φ is already at this point not a satisfying assignment. The same applies for clauses of the form $C_2 = (s \vee w \vee t)$, where on all three variables, s, w, t, both assignments disagree; call them type 2. It remains to upper-bound the probability of a $\mathcal{P}_{n,p}$ instance that does not contain type 1 and type 2 clauses. There are $\alpha n\binom{(1-\alpha)n}{2}$ type 1 clauses and $\binom{\alpha n}{3}$ type 2 clauses. The probability of none being chosen is

$$(1-p)^{\alpha n\binom{(1-\alpha)n}{2}+\binom{\alpha n}{3}} \leqslant \exp\left\{-p \cdot \left(\alpha n\binom{(1-\alpha)n}{2} + \binom{\alpha n}{3}\right)\right\}. \tag{2.1}$$

If $30/n^2p \leqslant \alpha \leqslant 1/2$, then

$$p \cdot \alpha n \binom{(1-\alpha)n}{2} \geqslant p\alpha n \cdot n^2/8 \geqslant 3n.$$

If $\alpha \geqslant 1/2$, then

$$p \cdot \binom{\alpha n}{3} \geqslant n \cdot n^2 p/48 \geqslant 3n.$$

In the last inequality we use the fact that n^2p can be arbitrarily large (specifically, greater than 144). In any case, the expression in (2.1) is at most 5^{-n}. Since we have no more than 4^n ways of choosing the pair φ, ψ, we deduce using the union bound that w.v.h.p. no such 'bad' pair exists. □

2.4. The core variables

We describe a subset of the variables, referred to as the *core variables*, which plays a crucial role in the understanding of $\mathcal{P}_{n,p}^{\rm sat}$. Recall that a variable is said to be frozen in F if, in every satisfying assignment, it takes the same value. The notion of a core captures this phenomenon. In addition, a core typically contains all but a small (though constant) fraction of the variables. This implies that a large fraction of the variables is frozen, a fact which must leave imprints on various structural properties of the formula. These imprints allow efficient heuristics to recover a satisfying assignment of the core. A second implication of this is an upper bound on the number of possible satisfying assignments, and on the distance between every such two. Thus the notion of a core plays a key role in obtaining a characterization of the cluster structure of the solution space.

Let us now proceed with a rigorous definition of a core. Before doing so, we take a long detour on expanding sets.

Definition 4 (Support). Given a 3CNF formula F and some assignment ψ to the variables, we say that a variable x *supports* a clause C (in which it appears) with respect to ψ if x is the only variable whose literal evaluates to true in C under ψ.

In some cases ψ might be a partial assignment (that is, some of the variables are not assigned). In such cases a variable x is said to support a clause C under ψ if all three variables of C are assigned a value under ψ and x is the only one of them whose literal evaluates to true.

In what follows $F[Z]$ stands for the subformula of F containing the clauses where all three variables belong to Z.

Definition 5 (Expanding set). Given a 3CNF formula F and an assignment ψ to the variables (not necessarily satisfying), a set of variables Z is called *t-expanding* in F with respect to ψ if every variable $x \in Z$ supports at least t clauses in $F[Z]$ with respect to ψ.

The following proposition illustrates the usefulness of Definition 5.

Proposition 2.6. *Let F be a 3CNF formula on n variables and let Z be a t-expanding set with respect to some assignment ψ. If in addition:*

- *ψ satisfies F,*
- *F is $n/10^6$-concentrated (Definition 3),*
- *F is t-proportional (Definition 2),*

then the variables in Z are frozen in F.

Proof. Let ψ be the satisfying assignment with respect to which Z is defined. By contradiction assume that there exists a satisfying assignment ψ' differing from ψ on Z. Then let $\emptyset \neq U \subseteq Z$ be the set of variables x of Z for which $\psi(x) \neq \psi'(x)$. Take $x \in U$ and consider all the clauses that x supports with respect to ψ in $F[Z]$. It must be that every such clause contains at least another variable y on which ψ and ψ' disagree (since every such clause is satisfied by ψ' but the literal corresponding to x is false under ψ'). Therefore y belongs to U by definition. We conclude that there exists a set U of variables and $t \cdot |U|$ clauses each containing at least two variables from U (no clause was counted twice since the supporter of a clause is unique by definition). Further, we assumed that F is $n/10^6$-concentrated and therefore $|U| \leqslant n/10^6$. Combining the latter two facts we derive a contradiction to the t-proportionality of F. □

Proposition 2.7. *Let F be distributed according to $\mathcal{P}_{n,p}$ with $n^2 p \geqslant d$, for d a sufficiently large constant. There exists a positive integer $t = t(n, p)$ such that w.v.h.p. there is a set Z*

of variables, and an assignment ψ such that:

- *Z is t-expanding with respect to ψ,*
- *$|Z| = (1 - e^{-\Omega(n^2p)})n$,*
- *F is $t/10$-proportional,*
- *ψ satisfies F^*,*
- *F^* is $n/10^6$-concentrated.*

The complete proof of this proposition is deferred to Section 4. Here and throughout the notation $f = \Omega(n^2p)$ (or $f = O(n^2p)$) stands for $f \geqslant cn^2p$ (or $f \leqslant cn^2p$) for some constant c that does not depend on n, p.

Corollary 2.8. *The set Z promised in Proposition 2.7 is frozen in F^*, and furthermore Z is t-expanding with respect to every satisfying assignment of F^*.*

Proof. To see why Z is frozen, let S be the set of clauses in F that are supported with respect to ψ. First observe that F^* is t-proportional as it is a subformula of F (and F is $t/10$-proportional and therefore also t-proportional). Furthermore S is contained in F^*. This is because ψ is a satisfying assignment of F^* throughout the entire generating process, and thus every clause in S that arrives is not rejected. Therefore Z is also t-expanding in F^* with respect to ψ. Finally apply Proposition 2.6 to F^*. The second part of the corollary is immediate from the fact that Z is frozen. □

Definition 6 (Self-contained sets). Given a 3CNF formula F we say that a set of variables Z is *r-self-contained* in F if every variable $x \in Z$ appears in at most r clauses in $F \setminus F[Z]$.

Finally, we are ready to define a core.

Definition 7 (Core). A set of variables $\mathcal{H}$ is called a *t-core* of F with respect to an assignment ψ if $\mathcal{H}$ is t-expanding in F with respect to ψ and also $(t/3)$-self-contained in F.

The property of being self-contained is necessary for the algorithmic part (the proof of Theorem 1.3, at least as our analysis proceeds).

Proposition 2.9. *Let F^* be distributed according to $\mathcal{P}^{\text{sat}}_{n,p}$ with $n^2p \geqslant d$, for d a sufficiently large constant. There exists a positive integer $t = t(n,p)$ such that w.v.h.p. there exists a satisfying assignment ψ of F^* and a t-core $\mathcal{H}$ with respect to ψ with the following properties:*

- *$|\mathcal{H}| = (1 - e^{-\Omega(n^2p)})n$,*
- *$\mathcal{H}$ is frozen in F^*,*
- *F^* is $t/10$-proportional.*

The proof of this proposition is best understood in the context of the proof of Proposition 2.7. Therefore the proof appears in Section 4.1.

Remark 4. Observe that if there exist two t-cores $\mathcal{H}_1$ and $\mathcal{H}_2$ that satisfy the conditions of Proposition 2.9, then their union $\mathcal{H}_1 \cup \mathcal{H}_2$ is also a t-core (since the core variables are frozen). Therefore we may speak of a unique maximal t-core. From now on, when we refer to a t-core, we mean the maximal one. Note that this maximal core is also frozen by Proposition 2.6. Therefore it can serve as a t-core for *any* satisfying assignment of F, and is thus effectively uniquely defined by the formula.

2.5. Satellite variables

In this section we isolate another set of variables which we call satellite variables. As it turns out, to prove Theorems 1.1 and 1.3, it is enough to distinguish between core and satellite variables and all other variables in V. Let us start with a formal definition of a satellite variable. In what follows, ℓ_z stands for a literal corresponding to a variable z (*i.e.*, $\ell = z$ or $\ell = \bar{z}$).

Definition 8. Given a formula F, an assignment φ and a core set $\mathcal{H}$ of F with respect to to φ, a variable x is called a 0-*satellite* of $\mathcal{H}$ if $x \in \mathcal{H}$. A variable x is called an i-*satellite* if F contains at least one clause of the form $(\ell_x \vee \ell_{z_1} \vee \ell_{z_2})$, where for $j = 1, 2$, z_j is a b-satellite for some $b < i$, $\varphi(\ell_{z_j}) =$ FALSE, and at least one of the z_js is an $(i-1)$-satellite. We say that x is a *satellite variable* if it is a b-satellite for some number $b \geqslant 1$.

Observe that if $\mathcal{H}$ is frozen in F then $\mathcal{H} \cup \mathcal{S}$ is frozen as well (this follows from a simple inductive argument).

Before we formally state the property involving the satellite variables we introduce additional notation. The connected components of a formula F are the subformulas $F[C_1], \ldots, F[C_k]$, where $C_1, C_2, \ldots, C_k$ are the connected components in the graph G_F induced by F (the vertices of G_F are the variables, and two variables are connected by an edge if there exists some clause containing them both). Given a set of variables A and an assignment φ we denote by $F_{\text{out}}(A, \varphi)$ the subformula of F which is the outcome of the following procedure: set the variables in A according to φ and simplify F (by simplify we mean remove every clause that contains a TRUE literal, and remove FALSE literals from the other clauses).

Proposition 2.10. *Let F^* be distributed according to $\mathcal{P}^{\text{sat}}_{n,p}$ with $n^2 p \geqslant d$, for d a sufficiently large constant. There exists a positive integer $t = t(n,p)$ such that w.v.h.p. there exists satisfying assignment ψ of F^* and a t-core $\mathcal{H}$ with respect to ψ with the following properties:*

- $|\mathcal{H}| \geqslant (1 - e^{-\Omega(n^2 p)})n$,
- *F^* is $t/10$-proportional,*
- *let $\mathcal{S}$ be its satellite variables; $\mathcal{H} \cup \mathcal{S}$ are frozen in F^*,*
- *the largest connected component in $F^*_{\text{out}}(\mathcal{H} \cup \mathcal{S}, \psi)$ is of size at most $\log n$.*

The new addition compared with Proposition 2.9 is the fact that we characterize the structure of the formula induced by the variables not in $\mathcal{H} \cup \mathcal{S}$.

Our proof strategy is as follows. Expose the first part of the random formula F and consider a t-core $\mathcal{H}$ promised w.v.h.p. by Proposition 2.9. We look at a ‘large’ connected component outside the core (if none exists then we are done) and consider the following 'shattering' procedure. Expose the second part of the random formula, and suppose for the time being that the core does not change (even if new clauses are included in F^*). Let x be a non-core variable after the first part, which lies in a spanning tree of a large connected component. The key observation is that, when resuming the random clause process, x becomes a satellite variable with high (constant) probability, in which case the spanning tree splits into parts. Since the tree is large, it contains many variables x, and therefore with very high probability at least one of them will become a satellite variable and shatter the tree. Finally, it remains to upper-bound the number of possible large trees *vs* the probability that such a tree does not survive. The complete proof is given in Section 5.

One problem with the approach we just described is that we assumed that the core $\mathcal{H}$ established after the first round does not change when resuming the random clause process. This is not necessarily the case, as, for example, some core variables may violate the self-containment property and be removed, and this may cause a chain reaction of other variables leaving the core (maybe their support is too small, or they violate the self-containment requirement). However, w.v.h.p. all the variables that are removed from the core when resuming the random clause process remain satellite variables, and furthermore there are very few such variables.

Remark 5. In several papers that have studied planted-solution distributions, for example [3] and [14], a similar notion of a core appears (without the notion of satellite variables), and an analysis of the structure of the instance (k-colourable graph or k-CNF formula) induced on the non-core variables is also given. The main difference from our setting is the fact that the planted distribution is a product space, and therefore it was possible to prove that the core variables are distributed similarly to a uniformly random set of variables. In our case establishing such a property is a more challenging task. As it turns out, the approach that we take – defining the satellite variables – considerably simplifies the proof of this property.

2.6. The majority vote

Given a 3CNF formula F and a variable x, we let $N^+(x)$ be the set of clauses in F in which x appears positively (namely, as the literal x), and let $N^-(x)$ be the set of clauses in which x appears negatively (that is, as $\bar{x}$). The majority vote assignment over F, which we denote by MAJ, assigns every x according to the sign of $|N^+(x)| - |N^-(x)|$ (TRUE if the difference is positive and FALSE otherwise).

Proposition 2.11. *Let F^* be distributed according to $\mathcal{P}^{\text{sat}}_{n,p}$ with $n^2p \geqslant d$, for d a sufficiently large constant. Then w.v.h.p. every satisfying assignments of F^* differs from* MAJ *on at most $e^{-\Omega(n^2p)}n$ variables.*

Proof. Consider the following two-step procedure to generate F. In the first step go over the $M = 8\binom{n}{3}$ clauses and toss a coin with success probability p_1. We take the clauses that were chosen and put them first, ordered at random. Call F_1 this first part (and respectively define F_1^* in our standard way, *i.e.*, by scanning sequentially the clauses of F_1 and including those whose addition leaves the formula satisfiable.). Observe that F_1 is distributed according to $\mathcal{P}_{n,p_1}$. Then in the second round, every clause that was not chosen in the first round is included with probability p_2, and the chosen clauses are ordered at random and then concatenated after F_1. Call F_2 this last part. A straightforward calculation shows that $F = F_1 \cup F_2$ is distributed according to $\mathcal{P}_{n,p}$ when $p = p_1 + (1 - p_1)p_2$. Therefore we may think of F as generated in two steps (with the suitable choice of p_1, p_2). We will use this technique to prove several other properties as well.

Let d_0 be the constant promised in Proposition 2.9, and choose $d \geqslant 2000d_0$. Set $p_1 = p/2000$. By the choice of d_0 and Proposition 2.9, w.v.h.p. all but $e^{-\Omega(n^2p)}n$ variables are frozen in F_1^*, and without loss of generality assume that they all take the value TRUE.

We argue now that w.v.h.p. at this point all but $e^{-\Omega(n^2p)}n$ variables appear in no more than say $n^2p/30$ clauses in F_1 (and therefore also in F_1^*). The number of clauses of F_1 containing a given variable x is distributed binomially with parameters $8\binom{n-1}{2}$ and p_1 and thus has expectation $8\binom{n-1}{2}p_1 \leqslant n^2p/500$. Therefore, if $n^2p \geqslant C_1 \log n$ for some sufficiently large constant $C_1 > 0$, then w.v.h.p. every variable appears in no more than $n^2p/30$ clauses, by the Chernoff inequality.

Assume therefore that $n^2p \leqslant C_1 \log n$. Choose $c_2 = c_2(C_1) > 0$ sufficiently small so that $e^{-c_2n^2p}n \geqslant \sqrt{n}$. If F_1 has a set X of t variables, each contained in at least $n^2p/30$ clauses, then the total number of clauses of F_1 containing at least one variable from X is at least $t(n^2p/30)(1/3) = n^2pt/90$ (each clause is counted at most thrice). The probability of the latter event happening in $\mathcal{P}_{n,p_1}$ for some set of t variables is at most

$$\binom{n}{t}\binom{8t\binom{n-1}{2}}{\frac{n^2pt}{90}}p_1^{\frac{n^2pt}{90}} \leqslant \left(\frac{en}{t}\right)^t\left(\frac{360e}{2000}\right)^{\frac{n^2pt}{90}} < \left[\frac{en}{t}2^{-\frac{n^2p}{90}}\right]^t.$$

Choosing $t = e^{-c_2n^2p}n$ (and recalling that c_2 is chosen small enough so that, in particular, $t \geqslant \sqrt{n}$), we conclude that the above estimate is less than n^{-3}.

Let Z be then the set of frozen variables that appear in at most $n^2p/30$ clauses of F_1. Recall that we have assumed without loss of generality that they all froze to TRUE. By the above discussion together with Proposition 2.9, w.v.h.p.

$$|Z| \geqslant (1 - e^{-\Omega(n^2p)})n - e^{-\Omega(n^2p)}n \geqslant 0.999n.$$

Now let us consider the second iteration of coin flips. Fix $x \in Z$, observe that every clause containing x positively, if chosen in the second round, will be included in F^*. There are at least $4\binom{|Z|-1}{2} - n^2p/30$ such clauses with the other two variables from Z: call them 'good' clauses. As for clauses where x appears negatively, and the other two variables are in Z, there are only at most $3\binom{|Z|-1}{2}$ clauses such that if chosen will be included (since one way of negating the variables in Z results in a FALSE clause on frozen variables): call them 'bad' clauses. In addition there are at most $8(n - |Z|)n$ clauses, containing x and at least one variable outside Z, about which we say nothing, but let us adversely assume that x appears in all of them negatively, and if chosen are included in F^* (they are also part of the bad clauses).

In expectation, at least $p_2 \cdot \big(4\binom{|Z|-1}{2} - n^2p/30\big) \geqslant 1.8n^2p$ good clauses containing x will be chosen in the second round, and at most $p_2 \cdot \big(3\binom{|Z|-1}{2} + 8(n-|Z|)n\big) \leqslant 1.6n^2p$ bad clauses. (Recall that $199p/200 \leqslant p_2 \leqslant p$.)

Suppose that in the $n^2p/30$ clauses from the first round x also appears negatively. To conclude, for the majority vote of x to be wrong it must have been the case that the number of good clauses containing x or the number of bad clauses containing x deviates by at least $(1.8-1.6-1/30)n^2p/2$ from its expectation. But since both are binomially distributed with expectation $\Theta(n^2p)$, this happens with probability $e^{-\Omega(n^2p)}$. Using the linearity of expectation, all but $e^{-\Omega(n^2p)}n$ of the variables in Z are expected to have a 'proper' gap. To obtain concentration around this value we argue as for concentration of $|Z|$ above. Finally observe that $|Z| \geqslant (1-e^{-\Omega(n^2p)})n$, and therefore $|Z| - e^{-\Omega(n^2p)}n = (1-e^{-\Omega(n^2p)})n$ as required. □

2.7. Proof of Theorem 1.1

Theorem 1.1 follows from Proposition 2.10, which implies that all but $e^{-\Omega(n^2p)}n$ of the variables are frozen. Therefore, there are at most $2^{e^{-\Omega(n^2p)}n}$ possible ways to set the assignment of the remaining variables. Furthermore, Proposition 2.10 describes the formula induced by the non-frozen variables.

3. Proof of Theorem 1.3

In this section we prove that the algorithm SAT, which is described in Figure 1, meets the requirements of Theorem 1.3. The main principles underlying SAT were designed with the planted distribution in mind (see [14], for example). An additional ingredient that we add is a unit clause propagation step. Given a 1-2-3-CNF formula (namely a formula which contains clauses of size 1, 2 and 3), the unit clause propagation is the following simple heuristic:

> While there exists a clause of size 1, set the variable appearing in this clause in a satisfying manner, remove this clause and all other clauses satisfied by this assignment, and remove the FALSE literals of the variable from other clauses.

We say that F^* is *typical* in $\mathcal{P}^{\text{sat}}_{n,p}$ if it satisfies the properties stated in Propositions 2.10 and 2.11. The discussion in Section 2 guarantees that indeed this happens w.v.h.p. (that is, F^* is typical). Therefore, to prove Theorem 1.3 it suffices to consider a typical F^* and prove that SAT (always) finds a satisfying assignment for F^* in polynomial time. As the parameter t for SAT we use the t promised in Proposition 2.10.

We let $\mathcal{H}$ be the t-core promised in Proposition 2.10, let $\mathcal{S}$ be its satellite variables, and let φ be the satisfying assignment with respect to which $\mathcal{H}$ is defined. In all the following propositions we assume F^* is typical (for the sake of brevity we do not explicitly state it every time).

Proposition 3.1. *Let ψ_1 be the assignment defined in line 7 of* SAT*. Then ψ_1 agrees with φ on the assignment of all variables in $\mathcal{H}$.*

SAT(F, t)
Step 1: Majority vote
1. $\pi_1 \leftarrow$ majority vote over F.
Step 2: Reassignment
2. **for** $i = 1$ to $\log n$
3. **for all** variables x
4. **if** x supports fewer than $2t/3$ clauses with respect to π_i **then** $\pi_{i+1} \leftarrow \pi_i$ with x flipped.
5. **end for.**
6. **end for.**
Step 3: Unassignment
7. set $\psi_1 = \pi_{\log n}$, $i = 1$.
8. **while** $\exists x$ s.t. x supports fewer than t clauses with respect to ψ_i
9. set $\psi_{i+1} \leftarrow \psi_i$ with x unassigned.
10. $i \leftarrow i + 1$.
11. **end while.**
Step 4: Unit clause propagation
12. Let ξ be the final partial assignment obtained at Step 3.
13. Remove all clauses which are satisfied by ξ, and all FALSE literals from the remaining clauses.
14. Run the unit clause propagation algorithm on the resulting instance.
Step 5: Exhaustive search
15. Let F' be the formula remaining after the unit clause propagation of Step 4 terminates.
16. If some connected component of F' has more than $\log n$ variables, fail.
17. Exhaustively search and satisfy F', component by component.

Figure 1. The algorithm SAT.

Proof. Let B_i be the set of core variables whose assignment in π_i disagrees with φ at the beginning of the ith iteration of the main for-loop: line 2 in SAT. It suffices to prove that $|B_{i+1}| \leqslant |B_i|/2$ (if this is true, then after $\log n$ iterations $B_{\log n} = \emptyset$). Observe that by Proposition 2.11, $|B_0| \leqslant n/10^7$ (as the majority vote error-rate $e^{-\Omega(n^2 p)}$ can be made arbitrarily small). By contradiction, assume that not in every iteration $|B_{i+1}| \leqslant |B_i|/2$, and let j be the first iteration violating this inequality. Consider a variable $x \in B_{j+1}$. If also $x \in B_j$, this means that x's assignment was not flipped in the jth iteration, and therefore x supports at least $2t/3$ clauses with respect to π_j. Since $\mathcal{H}$ is $t/3$-self-contained, at least $2t/3 - t/3 = t/3$ of these clauses contain only core variables. Since the literal of x is true in all these clauses, but in fact should be false under φ, each such clause must contain another variable on which φ and π_j disagree, that is, another variable from B_j. If $x \notin B_j$, this means that x's assignment was flipped in the jth iteration. This is because x supports fewer than $2t/3$ clauses with respect to π_j. Since x supports at least t clauses with respect to φ (t-expanding property of the core), it must be that in at least $t - 2t/3 = t/3$ of them, the literal of some other core variable evaluates to TRUE (not FALSE as it should be in φ). Let B'_{j+1} be the first (lexicographically) $n/10^7$ variables of B_{j+1} (or $B'_{j+1} = B_{j+1}$ in case B_{j+1} contains fewer than $n/10^7$ variables). Observe that in case we trim B_{j+1} it still holds that $|B'_{j+1}| \geqslant |B_j|/2$ since $|B_j| \leqslant n/10^7$ (B_0 is already so small, and by our assumption the sets $B_1, B_2, \ldots, B_j$ only decrease in size). Define $U = B_j \cup B'_{j+1}$. There are at least $t/3 \cdot |B'_{j+1}|$ clauses containing at least two variables from U (every clause is counted exactly once as the supporter of a clause is unique). Using our assumption, $|B'_{j+1}| \geqslant |B_j|/2$, we obtain $|U| = |B_j \cup B'_{j+1}| \leqslant |B_j| + |B'_{j+1}| \leqslant 3|B'_{j+1}|$, therefore $t/3 \cdot |B'_{j+1}| \geqslant (|U|/3) \cdot t/3 = (t/9)|U|$. Finally,

- $|B'_{j+1}| \leqslant n/10^7$,
- $|U| \leqslant 3|B'_{j+1}| \leqslant 3n/10^7 \leqslant n/10^6$,
- there are $t|U|/9$ clauses containing two variables from U.

The last two items contradict the $t/10$-proportionality of F^*. □

Proposition 3.2. *Let ξ be the partial assignment defined in line 12 of* SAT. *Then all assigned variables in ξ are assigned according to φ, and all the variables in $\mathcal{H}$ are assigned.*

Proof. By Proposition 3.1, ψ_1 coincides with φ (the satisfying assignment with respect to which $\mathcal{H}$ is defined) on $\mathcal{H}$. Furthermore, by the definition of t-core, every core variable supports at least t clauses with respect to φ, and also with respect to ψ_1 (the assignment at hand before the unassignment step begins). Hence all core variables survive the first round of unassignment. By induction it follows that the core variables survive all rounds. Now suppose by contradiction that not all assigned variables are assigned according to φ when the unassignment step ends. Let U be the set of variables that remain assigned when the unassignment step ends, and whose assignment disagrees with φ. Every $x \in U$ supports at least t clauses with respect to ξ (the partial assignment defined in line 12 of SAT), but each such clause must contain another variable on which ξ and φ disagree (since φ satisfies this clause). Thus, we have $t \cdot |U|$ clauses each containing at least two variables from U (again no clause is counted twice as the support of a clause is unique). Since $U \cap \mathcal{H} = \emptyset$ (by the first part of this argument) and $|\mathcal{H}| \geqslant (1 - e^{-\Omega(n^2 p)})n$, it follows that $|U| \leqslant e^{-\Omega(n^2 p)}n < n/10^6$, contradicting the $t/10$-proportionality of F^*. □

Proposition 3.3. *By the end of the unit clause propagation step, all the variables which get assigned are assigned according to φ; furthermore, the set of satellite variables $\mathcal{S}$ is assigned.*

Proof. The proof is by induction on the iterations of the unit clause propagation. The base case consists of clauses of the form $(x \vee \ell_z \vee \ell_y)$ where ℓ_z, ℓ_y are FALSE literals under ξ and x is unassigned. By the previous proposition, ξ can be extended to a satisfying assignment of F, but every such extension must set $x =$ TRUE. This is exactly what the unit clause propagation does. The step of the induction is proved similarly to the base case.

Now to the satellite variables. The previous proposition gives that $\mathcal{H}$ remains assigned according to φ. By the definition of satellite variables, $\mathcal{S}$ will be set in the unit clause propagation (the i-satellite variables will be set in iteration i of the unit clause propagation). □

Proposition 3.4. *The exhaustive search, Step 5 of* SAT, *completes in polynomial time with a satisfying assignment of F^*.*

Proof. By Proposition 3.3, the partial assignment at the beginning of the exhaustive search step is partial to the satisfying assignment φ of the entire formula. Therefore the exhaustive search will succeed. Further, observe that the unassigned variables are outside $\mathcal{H} \cup \mathcal{S}$. F^* is assumed to be typical (specifically, it satisfies the properties stated in Proposition 2.10) and this guarantees that the algorithm will not fail in line 16. This also guarantees that the running time of the exhaustive search will be at most polynomial. □

Theorem 1.3 follows.

4. Proof of Proposition 2.7

Let F be the random $\mathcal{P}_{n,p}$ instance, and let F^* be its satisfiable part. We divide the process of generating F into two steps, as in the proof of Proposition 2.11. In the first round go over the $M = 8\binom{n}{3}$ clauses and toss a coin with success probability $p_1 = p/2$. Take the clauses that were chosen and put them first ordered at random. In the second round, every clause that was not chosen is included with probability p_2, where p_2 satisfies $p_1 + (1 - p_1)p_2 = p$; then the included clauses are ordered at random and concatenated after the first part. Observe that this distribution is identical to $\mathcal{P}_{n,p}^{\text{sat}}$ as explained before.

Let t be such that F (and hence also F_1) is w.v.h.p. t-proportional (we can choose $t = n^2p/5500$ as asserted in Proposition 2.3). Also take n^2p sufficiently large so that F^* is w.v.h.p. $n/10^6$-concentrated (as required by Proposition 2.7, and as promised to be the case w.v.h.p. by Proposition 2.5).

Fix ψ to be some assignment (not necessarily a satisfying assignment of F^*), and let B_ψ be a random variable counting the number of variables whose support in F_1 with respect to ψ is smaller than $502t$. Fix some variable x, and without loss of generality assume x is TRUE in ψ. There are $\binom{n-1}{2}$ clauses that x supports with respect to ψ, each included with probability p_1. Therefore, in expectation x supports at least $n^2p_1/3 = n^2p/6$ clauses. Since the support of x is distributed binomially, the probability that x supports fewer than t clauses in F_1 with respect to ψ is at most $e^{-n^2p/50}$ (say, use the Chernoff bound). Finally observe that the set of clauses that x supports is disjoint from the set of clauses that $y \neq x$ supports. Therefore, the probability that there are at least $n/10^7$ such variables is at most $\binom{n}{n/10^7}e^{-(n^2p/50)\cdot(n/10^7)} < 3^{-n}$ for sufficiently large n^2p. Hence w.v.h.p. no such ψ exists (taking the union bound over all possible assignments).

In particular, w.v.h.p. every ψ that satisfies F_1^* has the desired property. Let now ψ be a satisfying assignment of F_1^* such that $B_\psi \leqslant n/10^7$, and consider the following procedure which, as we shall prove, produces a large $500t$-expanding set Z in F_1^* (and therefore also in F^* which contains F_1^*). When using the notation $F[A]$ for a formula F and a set of variables A, we mean all clauses in F in which all three variables belong to A.

Clearly, Z is $500t$-expanding in F_1 (by the construction). It remains to show that Z is large. By our assumption on B_ψ, step 1 removes at most $n/10^7$ variables; let A be those variables. It remains to prove that in the iterative step not too many variables were removed. Suppose by contradiction that in the iterative step more than $n/10^7$ variables were removed, and consider iteration $j = n/10^7$ and the set $W = \{a_1, \ldots, a_j\}$ ($a_i \in W$

1. set $Z_0 = V \setminus \{x \in V : x$ supports fewer than $502t$ clauses in F_1^* with respect to $\psi\}$; $i = 0$.
2. **while** there exists a variable $a_i \in Z_i$ that supports fewer than $500t$ clauses in $F_1[Z_i]$
   do $Z_{i+1} = Z_i \setminus \{a_i\}$; $i \leftarrow i + 1$.
3. let a_r be the last variable removed in step 2. Define $Z = Z_{r+1}$.

Figure 2. Building a t-expanding set.

is defined in line 2 of Figure 2). Every $a_i \in W$ appears in more than $502t - 500t = 2t$ clauses in which at least another variable belongs to $U = W \cup A$ (by the choice of Z_0 and the condition in line 2 that caused a_i to be removed). Therefore, by iteration $n/10^7$, the set U contains at most $n/10^7 + n/10^7 \leqslant n/10^6$ variables, and there are more than $2t \cdot |W| \geqslant 2t \cdot |U|/2 = t|U|$ clauses containing at least two variables from U (no clause is counted twice as the support of a clause is unique; $|W| \geqslant |U|/2$ by our assumption on the size of A and by the choice of $j = n/10^7$). This contradicts the t-proportionality of F_1. To conclude, $|Z| \geqslant \left(1 - 10^{-6}\right)n \geqslant 0.99n$.

It follows that w.v.h.p. for every satisfying assignment ψ of F_1^* there exists a $500t$-expanding set Z of variables of cardinality $|Z| \geqslant 0.99n$. Without loss of generality we can assume Z is of maximal size.

Observe that Z and F_1^* satisfy the conditions of Proposition 2.6 (that is, ψ is a satisfying assignment, F_1^* is t-proportional and $n/10^6$-concentrated) and therefore Z is frozen in F_1^*; without loss of generality assume that all variables in Z froze to TRUE. Since all variables of Z are frozen in F_1^*, we can take *the same* Z for every satisfying assignment ψ of F_1^*.

So let Z be as above, $|Z| \geqslant 0.99n$. Now we consider the second round of coin tosses; call the chosen clauses F_2. We prove that after adding them, with probability $1 - o(2^{-n})$ Z extends to a t-expanding set Z', $Z \subseteq Z'$, of the required size ($|Z'| \geqslant (1 - e^{-\Omega(n^2 p)})n$). Fix some variable $x \notin Z$ and observe that x supports $\binom{|Z|}{2}$ clauses, where x appears without negation and the other two variables are in Z and appear as negated. Since $x \notin Z$, we know by the maximality of Z that in the first iteration at most $500t$ such clauses were included. In expectation, F_2 contains at least $p_2\left(\binom{|Z|}{2} - 500t\right) \geqslant n^2 p/5 \geqslant 1000t$ such clauses (this is due to $p_2 \geqslant p/2$). If indeed at least $500t$ clauses are included then $Z \cup \{x\}$ is a $500t$-expanding set. The probability that fewer than $500t$ of them were included is $e^{-\Omega(n^2 p)}$ (again, the Chernoff bound). We can argue similarly about the number of clauses in F_2, containing x and two variables from Z, where all three variables appear as negated.

Call a variable x *good* if it participates in at least $500t$ clauses in F_2 where the other two variables are from Z and are negated and x is not negated, and is also in at least $500t$ clauses in F_2 where the other two variables are from Z and all three variables are negated; otherwise x is called *bad*. Observe that for every good x, for every satisfying assignment ψ of $F_1^* \cup F_2^*$, we can add x to Z, regardless of whether ψ sets x to TRUE or FALSE. The above argument shows that the expected number of bad variables is $e^{-\Omega(n^2 p)}n$. Applying standard concentration techniques, we can derive that w.v.h.p. the number of bad variables is w.v.h.p. $e^{-\Omega(n^2 p)}n$ as well (the events 'x_i is bad' and 'x_j is bad' are independent since they involve disjoint sets of clauses).

1. set $H_0 = Z$ and $i = 0$.
2. **while** there exists a variable $a_i \in H_i$ such that:
   - a_i appears in more than $t/11$ clauses in $F \setminus F[H_i]$, **or**,
   - a_i supports fewer than $10t/11$ clauses in $F[H_i]$,

   do $H_{i+1} = H_i \setminus \{a_i\}$.
3. let a_r be the last variable removed in step 2. Define $\mathcal{H} = H_{r+1}$.

Figure 3. Building a t-core.

So we have proved that w.v.h.p. there exists a $500t$-expanding set Z' (which contains Z) and $|Z'| = |Z| + (1 - e^{-\Omega(n^2p)})|V \setminus Z| \geqslant (1 - e^{-\Omega(n^2p)})|V|$.

In conclusion, we have shown that there exists a $500t$-expanding set Z' in F^* of cardinality $|Z'| = (1 - e^{-\Omega(n^2p)})n$ with respect to ψ, where ψ is some satisfying assignment of F^* (in fact this is true with respect to all satisfying assignments of F^* by the frozenness property). Scaling everything down (setting $t' = 500t$), Z' is t'-expanding and (at least) $t'/500$-proportional. This completes the proof of the proposition.

Remark 6. Note that here we proved $t'/500$-proportionality, which is stronger than we are required to prove ($t'/10$-proportionality). In general, we prefer a clear and brief presentation over optimizing the constants in the proofs. Later we will use this slackness in other proofs that rely on this one.

4.1. Proof of Proposition 2.9

Let Z be the t-expanding set promised by Proposition 2.7. Consider the procedure in Figure 3, which will produce a t'-core (for $t' = 10t/11$). Recall that by using the notation $F[A]$ for formula F and set of variables A, we mean all clauses in F in which all three variables belong to A.

First let us explain why indeed $\mathcal{H}$ is a t'-core. By its construction $\mathcal{H}$ is $10t/11$-expanding (or t'-expanding). Further, $\mathcal{H}$ is $t/11$-self-contained, or $t'/10$-self-contained (which also implies $t'/3$-self-contained as $1/3 > 1/10$).

Remark 7. By the definition of a core we are required to prove only that it is $t'/3$-self-contained, but we shall need this slackness in the proof of Proposition 2.10.

It remains to prove that $|\mathcal{H}| \geqslant (1 - e^{-\Omega(n^2p)})n$. By Proposition 2.7, $|H_0| \geqslant (1 - e^{-n^2p/c_1})n$ for some constant $c_1 > 0$ independent of n, p. Let $A = V \setminus H_0$, and note that $|A| \leqslant e^{-n^2p/c_1}n$. Suppose that the iterative procedure (line 2) removed more than $e^{-n^2p/c_1}n$ variables. Consider iteration $j = e^{-n^2p/c_1}n$ and the set $W = \{a_1, \ldots, a_j\}$ ($a_i \in W$ is defined in line 2 of Figure 3). Define $U = W \cup A$. One possibility for the removal of a_i is that it appears in at least $t/11$ clauses in which at least another variable belongs to U. Another is that a_i supports fewer than $10t/11$ clauses with respect to $F[H_i]$. In the latter case a_i must support at least $t - 10t/11 = t/11$ clauses in $F \setminus F[H_i]$ (by the choice of $a_i \in Z$). In any case a_i appears in at least $t/11$ clauses with at least another variable from U. Therefore, by iteration j, there exists a set U containing at most $2e^{-n^2p/c_1}n \ll n/10^6$ variables, and there are at least $(t/33) \cdot |W| \geqslant (t/33) \cdot (|U|/2) = t|U|/66$ clauses containing at least two variables from it (we divide $t/11$ by 3 as a clause could have been counted three times). This, however, contradicts the $t/500$-proportionality of F (recall that in Proposition 2.7, when proving the existence of a t-expanding set Z, we actually proved that F is $t/500$-proportional: see Remark 6).

Finally, the core variables are frozen as they are a subset of Z, and Z – the t-expanding set – is frozen (by Proposition 2.6).

5. Proof of Proposition 2.10

Let F be the random $\mathcal{P}_{n,p}$ instance, F^* its satisfiable part. We divide the process of generating F into two steps, as in the proof of Proposition 2.11. In the first round go over the $M = 8\binom{n}{3}$ clauses and toss a coin with success probability $p_1 = p/2$. Take the clauses that were chosen and put them first. In the second round, every clause that was not chosen is included with probability p_2, where p_2 satisfies $p_1 + (1 - p_1)p_2 = p$. Let F_1^* be the part of F^* that corresponds to the first iteration (F_1^* is distributed according to $\mathcal{P}^{\text{sat}}_{n,p_1}$). Let F_2 be the clauses that were chosen in the second round. Also, we say that a formula F is *bounded* if no variable appears in more than n clauses.

By Proposition 2.9, if we take n^2p to be sufficiently large, then F_1^* has w.v.h.p. a t-core $\mathcal{H}$ with respect to to a satisfying assignment ψ with the following properties (the last property did not appear in Proposition 2.9; we justify it immediately after):

- F_1 is $t/500$-proportional (Remark 6),
- $\mathcal{H}$ is $t/10$-self-contained and not only $t/3$-self-contained (Remark 7),
- $|\mathcal{H}| \geqslant (1 - e^{-\Omega(n^2p)})n$,
- $\mathcal{H}$ is frozen,
- F_1^* is bounded.

In F_1 every variable is expected to appear in $O(n^2p)$ clauses, and we may assume that $n^2p = O(n^{1/2})$ (if not, then, in particular, w.v.h.p. $\mathcal{H} = V$ and the entire discussion in this section is unnecessary). Standard calculations then show that w.v.h.p. no variable appears in more than n clauses of F_1.

We now discuss what happens to $\mathcal{H}$ in the second round, that is, when adding F_2. We will be interested in large connected components of F_1 whose vertices are not in $\mathcal{H}$ (Proposition 5.3), and also in vertices that may leave $\mathcal{H}$ due to F_2 (Propositions 5.1 and 5.2). The key to understanding the transformation that $\mathcal{H}$ and the connected components undergo lies in the notion of satellite variables.

Proposition 5.1. *There exists w.v.h.p. a satisfying assignment ψ of F^* and a set $\mathcal{H}' \subseteq \mathcal{H}$ of variables which is a $t/2$-core of F^* with respect to ψ, with $|\mathcal{H}'| \geqslant (1 - e^{-\Omega(n^2p)})n$.*

Proof. We call a variable $x \in \mathcal{H}$ *dirty* if in F_2 there exists a clause C containing x and some variable not in $\mathcal{H}$. Let D be the set of dirty variables. For a specific x, there are $e^{-\Omega(n^2p)}n^2$ clauses such that, if chosen to F_2, will make x dirty. The probability that any of them appears is at most $p_2 \cdot e^{-\Omega(n^2p)}n^2 = e^{-\Omega(n^2p)}$ (since $e^{-\Omega(n^2p)}$ is much smaller than n^2p_2

for sufficiently large p). Linearity of expectation gives $E[|D|] = e^{-\Omega(n^2p)}n$. Also observe that D satisfies the Lipschitz condition with difference 3 (as every new clause can affect 3 new variables). Therefore concentration is also obtained. Let us assume from now on that indeed $|D| = e^{-\Omega(n^2p)}n$.

Consider $\mathcal{H}$ after scanning F_2 (to complete F^*) and set $\mathcal{H}_0 = \mathcal{H} \setminus D, i = 0$. In a very similar way to the procedure in Figure 3, consider the following iterative procedure:

> While there exists $x \in \mathcal{H}_i$ such that x supports fewer than $t/2$ clauses in $F[\mathcal{H}_i]$ with respect to ψ, or appears in more than $t/6$ clauses where some variable belongs to $V \setminus \mathcal{H}_i$, define $\mathcal{H}_{i+1} = \mathcal{H}_i \setminus \{x\}, i = i + 1$.

Set $\ell = e^{-cn^2p}n$, where c is some constant satisfying $|D| \leqslant e^{-cn^2p}n$. Suppose that the iterative process reached iteration ℓ, and let W_ℓ be the set of variables that were removed in iterations $1 \dots \ell$, let $U = W_\ell \cup D$, and observe that $|W_\ell| \geqslant |U|/2$ by our choice of c. Take $x \in W_\ell$, if x was removed in iteration i, because it appeared in more than $t/6$ clauses where some variable belongs to $V \setminus \mathcal{H}_i$. Then, since x was part of $\mathcal{H}$ to begin with, and $\mathcal{H}$ was $t/10$-self-contained, x must appear in at least $t/6 - t/10 = t/15$ clauses in which at least another variable belongs to U. If x was removed because it supports fewer than $t/2$ clauses in $F[\mathcal{H}_i]$, then x was again part of $\mathcal{H}$, and therefore it supports at least t clauses in $F[\mathcal{H}]$, and hence it must support (and, in particular, appear in) at least $t - t/2 = t/2$ clauses in which some variable belongs to U. At any rate, every $x \in W_\ell$ appears in at least $t/15$ clauses in which at least another variable belongs to U. Finally,

- there are $t|W_\ell|/15 \cdot 1/3 \geqslant t|U|/90$ clauses containing at least two variables from U (we divide by 3 as every clause might have been over-counted up to 3 times, and we use the fact that $|W_\ell| \geqslant |U|/2$),
- $|U| = |D| + |W_\ell| = e^{-\Omega(n^2p)}n < n/10^6$ (we used our estimate on $|D|$, and the fact that we look at the iterative process until iteration ℓ, therefore $|W| \leqslant \ell = e^{-\Omega(n^2p)}n$).

Combining these two facts contradicts the $t/250$-proportionality of F^*, which holds w.v.h.p.. (F_1 is w.v.h.p. $t/500$-proportional, and so is F_2, as they are almost identically distributed, and there is enough slackness in the choice of constants to accommodate this difference. Hence $F = F_1 \cup F_2$ is w.v.h.p. $t/250$-proportional, and F^* is too.) Therefore, if we let W denote the set of variables that were removed in the iterative step, in all iterations, then w.v.h.p. $|W| \leqslant \ell$. Now set $t' = t/2$, and let $\mathcal{H}' = \mathcal{H} \setminus \{D \cup W\}$. We have shown that the set $\mathcal{H}'$ is a t'-core of the required size. Further, F^* is (at least) $t'/10$-proportional, as required by Proposition 2.10.

Finally observe that $\mathcal{H}$ is frozen and hence $\mathcal{H}' \subseteq \mathcal{H}$ is frozen too. Therefore, although $\mathcal{H}'$ is defined with respect to ψ, it will be a core of F^* regardless of which satisfying assignments survive at the end (as it will be a core with respect to all F^*'s satisfying assignments, and at least one is guaranteed to survive). □

Proposition 5.2. *Let $\mathcal{S}'$ be the satellite variables of $\mathcal{H}'$. If F^* is $t/10$-proportional then $\mathcal{H} \setminus \mathcal{H}' \subseteq \mathcal{S}'$.*

Proof. Let $A = \mathcal{H} \setminus \mathcal{H}'$. Let $\mathcal{S}'$ be all the satellite variables of $\mathcal{H}'$, and by contradiction assume that the set $B = A \setminus \mathcal{S}'$ is non-empty. Every x in B belongs to $\mathcal{H}$ and therefore supports at least t clauses where the other two variables appear in $\mathcal{H}$. Observe that in none of these t clauses are the other two variables in $\mathcal{H}' \cup \mathcal{S}'$ (as otherwise x is in $\mathcal{S}'$). Therefore we have found a set B, $|B| = e^{-\Omega(n^2 p)}n \leqslant n/10^6$, for which there are at least $t|B|$ clauses containing two variables from B. This contradicts the $t/10$-proportionality of F^*. □

In the proof of Proposition 2.10 we consider two 'types' of satellite variables. The first type, which we just met, are the variables in $\mathcal{H} \setminus \mathcal{H}'$. The second type, which we will make use of in the proof of Proposition 5.3 ahead, are satellite variables of $\mathcal{H}$ whose 'job' is to shatter the large connected components in the formula induced by variables not in $\mathcal{H}$ (when exposing the second part of F). In some sense these two types represent competing processes. The one is variables leaving $\mathcal{H}$, but still remaining satellite variables; the other is new variables attaching to $\mathcal{H}$ as satellite variables.

Recall our notation $F_{\text{out}}(A, \varphi)$ (A a set of variables, φ an assignment), which stands for the subformula of F which is the outcome of the following procedure. Set the variables in A according to φ and simplify F (by simplify we mean remove every clause that contains a TRUE literal, and remove FALSE literals from the other clauses). The connected components of a formula F are the subformulas $F[C_1], \ldots, F[C_k]$, where $C_1, C_2, \ldots, C_k$ are the connected components in the graph G_F induced by F (the vertices of G_F are the variables, and two variables are connected by an edge if there exists some clause containing them both).

Proposition 5.3. *Let $\mathcal{H}$ be a t-core of F_1^*, let $\mathcal{S}$ be the set of all satellite variables of $\mathcal{H}$ in F^*, and let ψ be a satisfying assignment of F^*. Then the largest connected component in $F^*_{\text{out}}[\mathcal{H} \cup \mathcal{S}, \psi]$ is w.v.h.p. of size at most* $\log n$.

First let us show why Proportion 5.3 completes the proof of Proposition 2.10. Take $\mathcal{H}'$ for the core to be given by Proposition 2.10, and denote by $\mathcal{S}'$ its satellite variables. Observe that (under the assumption of proportionality and by Proposition 5.2) $\mathcal{H} \subseteq \mathcal{H}' \cup \mathcal{S}'$. Hence, by the definition of satellite variables $\mathcal{H} \cup \mathcal{S} \subseteq \mathcal{H}' \cup \mathcal{S}'$. Also recall that if $\mathcal{H}'$ is frozen, then by definition $\mathcal{H}' \cup \mathcal{S}'$ is too.

5.1. Proof of Proposition 5.3

Let us refine the process of generating F. First we generate F_1 (and F_1^*), and fix $\mathcal{H}$ according to F_1^*. Then in the second round (F_2) first toss the coins of clauses C such that at most one literal in C belongs to $\mathcal{H}$, and call $J \subseteq F_2$ the set of clauses that were chosen. Finally, toss the coins of the other clauses (the ones that were not picked in the first step and contain at least two variables from $\mathcal{H}$), and call $K \subseteq F_2$ the set of clauses that were chosen. In this new terminology $F = F_1 \cup J \cup K$, and set $F' = F_1 \cup J$. To prove Proposition 5.3 it suffices to consider only trees of size $\log n$ in F'. This is because: (a) every connected component of size at least $\log n$ contains a tree of size $\log n$, and (b) only the clauses of F' may contribute edges to the connected components of $F^*_{\text{out}}[\mathcal{H} \cup \mathcal{S}, \psi]$.

We will prove Proposition 5.3 as follows. Fix an arbitrary tree T on r vertices, and let $V(T)$ denote its set of vertices. The following two conditions are necessary for T to belong to $F^*_{\mathrm{out}}[\mathcal{H} \cup \mathcal{S}, \psi]$:

- $A = \{$there exists a subformula of F' that induces $T\}$,
- $B = \{$the clauses in K do not prevent the following from holding: $V(T) \cap \mathcal{S} = \emptyset\}$.

The probability that $F^*_{\mathrm{out}}[\mathcal{H} \cup \mathcal{S}, \psi]$ contains a tree of size at least r is at most

$$\sum_{T:|V(T)|=r} \Pr[A \wedge B] = \sum_{T:|V(T)|=r} \Pr[A] \cdot \Pr[B|A]$$
$$\leqslant \left(\max_{T:|V(T)|=r} \Pr[B|A] \right) \cdot \left(\sum_{T:|V(T)|=r} \Pr[A] \right) \equiv q \cdot h.$$

Our next goal is to bound q and h, and then to show that $q \cdot h = o(n^{-3})$ for $r = \log n$. In fact we shall prove that $q \cdot h = o(n^{-\Omega(n^2 p)})$ for $r = \log n$. The next two lemmas establish the desired bounds (we use $d = n^2 p$).

Lemma 5.4. $h = \sum_{T:|V(T)|=r} \Pr[A] \leqslant n(100d)^r$.

Lemma 5.5. $q = \max_{T:|V(T)|=r} \Pr[B|A] \leqslant e^{-dr/8}$.

To conclude, for $r = \log n$,

$$q \cdot h \leqslant n(100d)^{\log n} \cdot n^{-d/8} \leqslant n^{1+\log(100d)-d/8} = o(n^{-\Omega(d)}).$$

The last equality is true since $d/8 \gg 1 + \log(100d)$ for sufficiently large $d = n^2 p$. We shall now prove the two lemmas.

Proof of Lemma 5.4. The quantity h to be estimated is obviously the expected number of trees of size $r = \log n$ induced by a formula $F' \subseteq F$ and is therefore at most the expected number of such trees induced by F itself. We thus estimate from above the latter quantity.

Let T be a fixed tree on r variables (a tree in the regular graph sense), and let F_T be a fixed collection of clauses such that each edge of T is induced by some clause of F_T; we call such an F_T an *inducing* set of clauses. We say that a clause set F_T is *minimal* with respect to T if by deleting a clause from F_T, T is no longer induced by the new formula. By the definition of minimality, $|F_T| \leqslant |E(T)| = |V(T)| - 1$ (as T is a tree). In our argument we shall be interested only in (T, F_T) such that F_T is a minimal set of clauses that induces T.

Given a tree T of size r, we estimate the number of ways to extend T to a minimal inducing set F_T. Every clause in F_T can cover either one or two edges of T (it cannot cover three edges or we have a cycle in T). Following the argument in [14], let $N_{T,s}$ be the number of ways to pair $2s$ edges of T to form s clauses in F_T that cover two edges. There are 8 ways to set the polarity of variables in every clause of F_T (and there are $r - 1 - s$ such clauses), and at most n^{r-1-2s} ways to choose the third variable in the $r - 1 - 2s$ clauses that cover exactly one edge. Using this terminology, the expected number of r-trees induced by a random formula F, generated according to $\mathcal{P}_{n,p}$ with $n^2 p = d$, is at most

$$\sum_{r\text{-trees}} \sum_{s=0}^{r/2} N_{T,s} 8^{r-1-s} n^{r-1-2s} \left(\frac{d}{n^2}\right)^{r-1-s} \leqslant \sum_{r\text{-trees}} \left(\sum_{s=0}^{r/2} N_{T,s}\right) (8d)^r n^{1-r}. \tag{5.1}$$

Our next task is to obtain useful upper bounds on the sum $\sum_{s=0}^{r/2} N_{T,s}$. To this end, let us fix a degree sequence $(d_1, \dots, d_r)$ for T, and consider the following procedure for *properly* pairing edges. By proper we mean that every pair of edges can be covered by a 3CNF clause; for example, we cannot pair the edges (x_1, x_2) and (x_3, x_4) as they result in a 4CNF clause. For each vertex, we specify a permutation of the edges incident to that vertex. Then we iterate through the vertices, and for each vertex, we iterate through the edges and pair up each unpaired edge with the edge given by the permutation associated with the current vertex (and leave the edge unpaired if the permutation sends the edge to itself). Any pairing of edges which can be covered by clauses can be generated this way by choosing the permutations to transpose each pair of edges to be covered by a single clause and to leave fixed all the other edges. Since there are $d_i!$ different permutations for vertex i, we have

$$\sum_{s=0}^{r/2} N_{T,s} \leqslant \prod_{i=1}^{r} d_i!.$$

A classical result by Prüfer is that the number of r-trees with degree sequence $(d_1, \dots, d_r)$ equals $\binom{r-2}{d_1-1,\dots,d_r-1}$ (see, for example, [21, Section 4.1, p. 33]). There are $\binom{n}{r}$ ways to choose the r vertices of the tree. So (5.1) is at most

$$\sum_{d_1+\cdots+d_r=2(r-1)} \binom{n}{r} \binom{r-2}{d_1-1,\dots,d_r-1} \left(\prod_{i=1}^{r} d_i!\right) (8d)^r n^{1-r}$$
$$\leqslant \sum_{d_1+\cdots+d_r=2(r-1)} \left(\prod_{i=1}^{r} d_i\right) (8d)^r n.$$

By convexity, for $(d_1, \dots, d_r)$ with $d_1 + \cdots + d_r = 2(r-1)$, the product $\prod_{i=1}^{r} d_i$ is maximized when $d_1 = \cdots = d_r$, and so $\prod_{i=1}^{r} d_i \leqslant 2^r$. The number of ways to choose positive integers $(d_1, \dots, d_r)$ so that $d_1 + \cdots + d_r = 2(r-1)$ is $\binom{2r-3}{r-1}$, which is less than 2^{2r}. Hence, the expected number of r-trees induced by a random formula F is at most $n \cdot 2^{2r} \cdot 2^r \cdot (8d)^r \leqslant n(100d)^r$. □

Proof of Lemma 5.5. Fix a tree T in $F' = F_1 \cup J$ on r vertices (recall that J is the set of clauses that contain at most one variable from $\mathcal{H}$), and consider the set of clauses that have at least two variables in $\mathcal{H}$, for which we now toss their coins (we use K to denote the set of clauses that were chosen among the latter).

Assume without loss of generality that the assignment ψ, with respect to which $\mathcal{H}$ is defined, is the all-TRUE assignment. Look at a variable $x \in V(T)$. We call a clause $(x \vee \bar{z}_1 \vee \bar{z}_2)$, $(\bar{x} \vee \bar{z}_3 \vee \bar{z}_4)$, where the z_is are some variables in $\mathcal{H}$, a type 1 (resp. type 2) clause. If clauses of both types appear in K then x surely belongs to $\mathcal{S}$ (and therefore

$V(T) \cap \mathcal{S} \neq \emptyset$). We call $x \notin \mathcal{H}$ *elusive* if at least one of the two types of clauses did not appear in K.

Set $\rho = 1 - e^{-\Omega(n^2 p)}$, since $|\mathcal{H}| \geqslant \rho n$ there are at least $\binom{\rho n}{2} \geqslant (\rho n)^2/3$ clauses of type 1. We assume that F_1 is bounded and hence every variable appears in at most n clauses, therefore at most n clauses of type 1 have been included in F_1. An identical argument applies to clauses of type 2. Note also that the clauses of J cannot belong to any of the types. Therefore the probability that no clause of type 1 belongs to K is at most $(1-p_2)^{(\rho n)^2/3-n} \leqslant e^{-d/7}$ (here we use that $d = n^2 p$ is large, $p_2 = (p-p_1)/(1-p_1)$, $p_1 = p/2$). The same is true by symmetry for clauses of type 2. Let E_x be the event that x is elusive, and let P_i be the event that no clause of type i for x appeared, for $i = 1, 2$ (namely, $E_x = P_1 \vee P_2$). Then

$$\Pr[E_x] = \Pr[P_1 \vee P_2] \leqslant \Pr[P_1] + \Pr[P_2] \leqslant 2e^{-d/7} \leqslant e^{-d/8}.$$

Further, observe that for $x \neq y$ the events E_x and E_y are independent as they involve disjoint sets of clauses (each variable supports its own set of clauses). Recall the events A, B, which were defined above. In this terminology we just upper-bounded the probability of B given A, and therefore the following is true:

$$\Pr[B|A] \leqslant \left(e^{-d/8}\right)^r = e^{-dr/8}.$$

Since our upper bound on $\Pr[B|A]$ only depends on the fact that $|V(T)| = r$, then also

$$q = \max_{T:|V(T)|=r} \Pr[B|A] \leqslant e^{-dr/8}. \qquad \square$$

6. *k*-colourability

In this section we will discuss, in a high-level fashion, how one can obtain similar results those we have for k-SAT for the random graph process (of k-colourability). Before we start our discussion let us recall the algorithm due to Alon and Kahale for colouring k-colourable graphs [3]. The first step of the algorithm is a spectral step; specifically a k-colouring of the graph (not necessarily proper) is obtained by looking at some eigenvectors of the graph (that hopefully reflect in some sense a proper k-colouring). Then, this initial k-colouring is refined using a series of combinatorial steps (very similar to our Steps 2–4 in algorithm SAT), until possibly a proper k-colouring is reached (or the algorithm fails). The algorithm was analysed on graphs drawn from the planted distribution first defined at [20] (the distribution is defined by the following procedure. Partition the vertex set into k colour classes of size n/k each: $V_1, V_2, \ldots, V_k$. Next, include every $V_i - V_j$ edge with probability p). The algorithm was shown to find w.h.p. a proper k-colouring of the graph when $np \geqslant ck^2$, for c some sufficiently large constant.

Let us consider the following distribution. Randomly permute all $\binom{n}{2}$ possible edges. Go over the edges in the permutation's order and add each edge if, after its addition, the graph remains k-colourable. It is possible to prove that the algorithm of Alon and Kahale also works for graphs drawn from the distribution we just described for the same edge density (maybe the constant c is different). The main challenge is to re-prove the spectral properties of the graph. The basic idea is to notice that w.h.p. every k-colouring has all of its colour classes of linear size, and also to prove discrepancy properties (similar, yet more elaborate, to Proposition 2.3). Another crucial ingredient in the proof is establishing a similar notion of a core (Definition 7).

Unfortunately, at this point we are still unable to answer a seemingly much simpler question: How many edges will such a graph typically contain by the end of the process? We expect the answer to be about $\binom{k}{2}\left(\frac{n}{k}\right)^2$, which would correspond to the case where a unique final k-colouring is nearly balanced.

7. Discussion

As we have already mentioned, only a vanishing proportion of 3CNFs with m clauses over n variables are satisfiable when m/n is above the threshold. In recent years, several papers have studied different distributions over satisfiable 3CNF formulas in the above-threshold regime, more precisely some sufficiently large constant factor above the threshold. In particular, [14] considered the planted 3SAT distribution, and [9] addressed the planted and uniform distributions, both papers developing new analytical and algorithmic techniques. Our work joins this line of research by studying a new distribution over satisfiable 3CNF formulas, and once again introducing new analytical ideas to face the intricacies of $\mathcal{P}^{\text{sat}}_{n,m}$. Furthermore, one interesting conclusion emerges from combining [14, 9] and our result. In all three distributions the instances show basically the same uni-cluster structure of the solution space, and the same algorithm solves them all. This gives rise to the following question: Does forcing (in some 'natural' way) the unlikely event of being satisfiable in the above-threshold regime generally result in the structure suggested by Theorem 1.1 (for clause-variable ratio greater than some sufficiently large constant)? This question has been answered positively for the planted and uniform distributions, and in this paper for the random satisfiable 3CNF process.

It is possible to generalize our result for k-CNF formulas for any fixed $k \geqslant 3$. In that case we shall require $m/n \geqslant c(k)$. Here is a formal statement.

Theorem 7.1. *Let F be random k-CNF from $\mathcal{P}^{\text{sat}}_{n,m}$, $m/n \geqslant c$, for $c = c(k)$ a sufficiently large constant. Then w.h.p. F enjoys the following properties.*

(1) *All but $e^{-\Omega(m/n)}n$ variables are frozen.*

(2) *The formula induced by the non-frozen variables decomposes into connected components of at most logarithmic size.*

(3) *Letting $\beta(F)$ be the number of satisfying assignments of F, we have $\frac{1}{n}\log\beta(F) = e^{-\Omega(m/n)}$.*

(4) *There exists a polynomial-time algorithm that finds a satisfying assignment for F.*

The constant $c(k)$ scales exponentially with k. The exponential dependency on k is inevitable as the satisfiability threshold itself scales exponentially with k (asymptotically $2^k \ln 2$).

References

[1] Achlioptas, D. and Coja-Oghlan, A. (2008) Algorithmic barriers from phase transitions. In *Proc. 49th Annual IEEE Symposium on Foundations of Computer Science*, pp. 793–802.

[2] Achlioptas, D. and Ricci-Tersenghi, F. (2006) On the solution-space geometry of random constraint satisfaction problems. In *Proc. 38th ACM Symposium on Theory of Computing*, pp. 130–139.
[3] Alon, N. and Kahale, N. (1997) A spectral technique for coloring random 3-colorable graphs. *SIAM J. Comput.* **26** 1733–1748.
[4] Alon, N. and Spencer, J. (2000) *The Probabilistic Method*, 2nd edn, Wiley-Interscience Series in Discrete Mathematics and Optimization, Wiley-Interscience, New York.
[5] Ben-Sasson, E., Bilu, Y. and Gutfreund, D. (2002) Finding a randomly planted assignment in a random 3CNF. Manuscript.
[6] Bohman, T., Frieze, A., Martin, R., Ruszinko, M. and Smyth, C. (2007) Randomly generated intersecting hypergraphs II. *Random Struct. Alg.* **30** 17–34.
[7] Chen, H. (2004) An algorithm for SAT above the threshold. In *Theory and Applications of Satisfiability Testing: 6th International Conference* (SAT 2003), Vol. 2919 of *Lecture Notes in Computer Science*, Springer, pp. 14–24.
[8] Coja-Oghlan, A., Krivelevich, M. and Vilenchik, D. (2007) Why almost all k-colorable graphs are easy. In *Proc. 24th Symposium on Theoretical Aspects of Computer Science*, Vol. 4393 of *Lecture Notes in Computer Science*, Springer, pp. 121–132.
[9] Coja-Oghlan, A., Krivelevich, M. and Vilenchik, D. (2007) Why almost all satisfiable k-CNF formulas are easy. In *13th Conference on Analysis of Algorithms: DMTCS Proceedings*, pp. 89–102.
[10] Cook, S. (1971) The complexity of theorem-proving procedures. In *Proc. 3rd ACM Symposium on Theory of Computing*, pp. 151–158.
[11] Erdős, P., Suen, S. and Winkler, P. (1995) On the size of a random maximal graph. *Random Struct. Alg.* **6** 309–318.
[12] Feige, U., Mossel, E. and Vilenchik, D. (2006) Complete convergence of message passing algorithms for some satisfiability problems. In *Approximation, Randomization, and Combinatorial Optimization: Algorithms and Techniques* (Proc. RANDOM 2006), Vol. 4110 of *Lecture Notes in Computer Science*, Springer, pp. 339–350.
[13] Feige, U. and Vilenchik, D. (2004) A local search algorithm for 3SAT. Technical report, The Weizmann Institute of Science.
[14] Flaxman, A. (2003) A spectral technique for random satisfiable 3CNF formulas. In *Proc. 14th ACM–SIAM Symposium on Discrete Algorithms*, pp. 357–363.
[15] Friedgut, E. (1999) Sharp thresholds of graph properties, and the k-sat problem. *J. Amer. Math. Soc.* **12** 1017–1054.
[16] Gallager, R. (1963) *Low-Density Parity-Check Codes*, MIT Press, Cambridge.
[17] Gerke, S., Schlatter, D., Steger, A. and Taraz, A. (2008) The random planar graph process. *Random Struct. Alg.* **32** 236–261.
[18] Håstad, J. (2001) Some optimal inapproximability results. *J. Assoc. Comput. Mach.* **48** 798–859.
[19] Krivelevich, M. and Vilenchik, D. (2006) Solving random satisfiable 3CNF formulas in expected polynomial time. In *Proc. 17th ACM–SIAM Symposium on Discrete Algorithms*, pp. 454–463.
[20] Kučera, L. (1977) Expected behavior of graph coloring algorithms. In *Proc. Fundamentals of Computation Theory*, Vol. 56 of *Lecture Notes in Computer Science*, Springer, Berlin, pp. 447–451.
[21] Lovász, L. (1993) *Combinatorial Problems and Exercises*, 2nd edn, Elsevier, Amsterdam.
[22] Mezard, M., Mora, T. and Zecchina, R. (2005) Clustering of solutions in the random satisfiability problem. *Phys. Review Letters* **94** 197–205.
[23] Osthus, D. and Taraz, A. (2001) Random maximal H-free graphs. *Random Struct. Alg.* **18** 61–82.
[24] Pearl, J. (1988) *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann, San Francisco, CA, USA.
[25] Ruciński, A. and Wormald, N. (1992) Random graph processes with degree restrictions. *Combin. Probab. Comput.* **1** 169–180.