On the Satisfiability Threshold of Random Community-Structured SAT

Dina Barak-Pelleg[1] and **Daniel Berend**[2*]
[1] Department of Math, Ben-Gurion University of the Negev, Beer Sheva, Israel
[2] Departments of Math and of Computer Science, Ben-Gurion University of the Negev, Beer Sheva, Israel
dinabar@post.bgu.ac.il, berend@cs.bgu.ac.il

Abstract

For both historical and practical reasons, the Boolean satisfiability problem (SAT) has become one of central importance in computer science. One type of instances arises when the clauses are chosen uniformly randomly – random SAT. Here, a major problem, recently solved for sufficiently large clause length, is the satisfiability threshold conjecture. The value of this threshold is known exactly only for clause length 2, and there has been a lot of research concerning its value for arbitrary fixed clause length.

In this paper, we endeavor to study the satisfiability threshold for random industrial SAT. There is as yet no generally accepted model of industrial SAT, and we confine ourselves to one of the more common features of industrial SAT: the set of variables consists of a number of disjoint communities, and clauses tend to consist of variables from the same community.

Our main result is that the threshold of random community-structured SAT tends to be smaller than its counterpart for random SAT. Moreover, under some conditions, this threshold even vanishes.

1 Introduction

The Boolean satisfiability problem (SAT) is one of the most important problems in theoretical computer science. Historically, it was the first problem proved to be NP-complete [Cook, 1971]. Since the introduction of the problem, there has been growing interest and research of all its aspects.

In this problem, one is required to determine whether a certain Boolean formula is satisfiable. An instance of the problem consists of a Boolean formula in several variables $v_1, \dots, v_n$. The formula is usually given in conjunctive normal form (CNF). The basic building block of the formula is a *literal*, which is either a variable v_j or its negation $\overline{v}_j$. A *clause* is a disjunction of the form $l_1 \vee ... \vee l_k$ of several distinct literals. Thus, altogether, the formula looks like $C_1 \wedge\ C_2 \wedge\ ... \wedge\ C_m$, where each C_i is a clause, say $C_i = l_{i,1} \vee ... \vee l_{i,k_i}$. Given a formula, one may assign a TRUE/FALSE value to each of the variables $v_1, \dots, v_n$. The formula is *satisfiable*, or SAT, if there exists an assignment under which the formula is TRUE, and is *unsatisfiable*, or UNSAT, otherwise.

The k-satisfiability (k-SAT) problem is a special case of SAT, in which each clause is a disjunction of up to k literals. Some authors restrict k-SAT to instances with exactly k literals in each clause, which terminology we will follow here. Given n, m and k, let $\Omega(n,\ m,\ k)$ denote the set of all k-SAT instances with n variables and m clauses. A *random k-SAT instance* is a uniformly random element of $\Omega(n,\ m,\ k)$. Namely, each clause is selected uniformly randomly out of all $\binom{n}{k}2^k$ possible clauses of length k, and distinct clauses are independent. Note that two instances, differing in the order of the clauses only, are considered as distinct.

The ratio m/n is the *density* and denoted by α. This parameter turns out to be very important. If α is sufficiently small, then a large random instance with density α is SAT with high probability, whereas if α is sufficiently large then a large random instance is UNSAT with high probability. Despite its loose name, the notion of "with high probability" is well defined. Let $(E_j)_{j=1}^{\infty}$ be a sequence of events. The event E_j occurs *with high probability (w.h.p.)* if $P(E_j) \xrightarrow[j\to\infty]{} 1$. In our case, we take larger and larger random instances with some fixed density, and inquire whether they are SAT or UNSAT. For $k \geq 2$, denote [Achlioptas and Peres, 2004]:

$$r_k \equiv \sup\{\alpha : \text{A random density-}\alpha \text{ instance is SAT w.h.p.}\}\ ,$$
$$r_k^* \equiv \inf\{\alpha : \text{A random density-}\alpha \text{ instance is UNSAT w.h.p.}\}\ .$$

For $k = 2$, it was proved long ago [Chvátal and Reed, 1992], [Fernandez de la Vega, 1992] and [Goerdt, 1992] that $r_2 = r_2^* = 1$. The Satisfiability Threshold Conjecture claims that, in fact, $r_k = r_k^*$ for all k [Chvátal and Reed, 1992]. This conjectured common value is the *satisfiability threshold*. It has been a subject of interest among researchers, theoretically and empirically, to prove the conjecture for $k \geq 3$ and find the threshold. Recently, the conjecture has been proved for large enough k [Ding et al., 2015].

As part of this research, lower and upper bounds were obtained on r_k and r_k^* for $k \geq 3$. In [Franco and Paull, 1983]

[*] Research supported in part by the Milken Families Foundation Chair in Mathematics

it was proven that $r_k^* \leq 2^k \ln 2$. This has been improved in [Kirousis et al., 1998] to $r_k^* \leq 2^k \ln 2 - \frac{1}{2}(1 + \ln 2) + o_k(1)$. From the other side, a sequence of successive improvements led finally to the bound $r_k \geq 2^k \ln 2 - \frac{1}{2}(1 + \ln 2) + o_k(1)$ [Coja-Oghlan and Panagiotou, 2016]. Thus, with the satisfiability conjecture settled in [Ding et al., 2015] for large k, it follows that $r_k = r_k^* = 2^k \ln 2 - \frac{1}{2}(1 + \ln 2) + o_k(1)$ for such k. For small values of k, more specific results were obtained. For $k = 3$, the best bounds seem to be $r_3 \geq 3.42$ [Kaporis et al., 2002] and $r_3^* \leq 4.506$ [Dubois et al., 2000]. Experiments and other results of heuristics, based on statistical physics considerations, indicate that $r_3 \approx 4.26$ [Mertens et al., 2006; Mézard and Zecchina, 2002], $r_4 \approx 9.93$, $r_5 \approx 21.12$, $r_6 \approx 43.37$, $r_7 \approx 87.79$ [Mertens et al., 2006].

Much more is known about 2-SAT. First, unlike k-SAT for $k \geq 3$, which is an NP-complete problem, 2-SAT instances may be solved by a linear time algorithm [Chvátal and Reed, 1992; Goerdt, 1992]. Also, there is quite precise information about 2-SAT for density very close to the threshold $r_2 = 1$ [Bollobás et al., 2001] and [Wilson, 1998].

It has been argued that instances of random k-SAT do not in fact represent real-world, or industrial, instances. One of the major differences between industrial and random SAT instances is that the set of variables in industrial instances often consists of a disjoint union of subsets, referred to as *communities*; clauses tend to comprise variables from the same community, with but a minority of clauses containing variables from distinct communities [Ansótegui et al., 2012; Newsham et al., 2014]. There are several additional differences. For example, the variables may be selected non-uniformly (say, according to a power-law distribution [Ansótegui et al., 2009; Giráldez-Cru and Levy, 2017]), and/or the clauses may be of non-constant length.

In this paper we work with a (generalization of a) model suggested by [Giráldez-Cru and Levy, 2015] . Our model is similar to the random model, except for the partition of the variables into communities. These communities are of the same size. There are several clause types, differing in the number of variables from the same or distinct communities in each clause. For example, a clause of type $(3, 2)$ is a clause of length 5, comprising 3 random variables from one random community and 2 from another.

Our focus is on the satisfiability threshold in this model. The question has been studied in [Giráldez-Cru and Levy, 2015], mostly experimentally, for the model suggested there. We show that the findings in that paper, whereby the threshold tends to be smaller when there are many single-community clauses, remain true in the general model. In fact, if the communities are small, the threshold may even be 0.

We present our model in Section 2. The main results are stated in Section 3, and the proofs follow in Section 4.

2 Random Industrial SAT

In industrial SAT, the strength of the community structure of an instance is usually measured by its modularity [Ansótegui et al., 2012; Giráldez-Cru and Levy, 2016; Park and Newman, 2003]. Roughly speaking, given a graph, its modularity gives an indication for the tendency of the vertices to be connected to other vertices which are similar to them in some way. In our case, an instance defines the following undirected graph. The set of nodes is the set of variables $\{v_1, \dots, v_n\}$. There is an edge (v_i, v_j) for $i \neq j$ if there exists a clause in the instance, containing both variables v_i (or its negation) and v_j (or its negation). Given an instance, high modularity indicates that there exists a partition of the set of variables into subsets, such that a large portion of the edges connect vertices of the same subset, compared to a random graph with the same number of vertices and same degrees [Newman, 2006; Park and Newman, 2003] .

As in the regular model, we have n Boolean variables and m clauses in an instance. Each clause is chosen independently of the others. Each variable in each clause is negated with a probability of $\frac{1}{2}$, independently of the other variables. The model differs from the regular model in several aspects: There is a community structure on the set of variables, and we also do not necessarily assume all clauses to be of the same length. Specifically, the set of variables $\{v_1, \dots, v_n\}$ is divided into B disjoint (sets of variables referred to as) communities $\mathcal{C}_1, \mathcal{C}_2, ..., \mathcal{C}_B$. For simplicity, we assume all communities to be of the same size h, so that $n = B \cdot h$. As n grows, so do usually both B and h (although at times one of them may remain fixed), and we will write $B(n)$ and $h(n)$ when we want to relate to their dependence on n. For an ℓ-tuple $\mathbf{k} = (k_1, \dots, k_\ell)$ with positive, non-decreasing, integer entries, denote by $\Omega_B\left(n,\ \mathbf{k}\right)$ the set of all clauses formed of k_1 variables from some community $\mathcal{C}_{i_1}$, k_2 from another community $\mathcal{C}_{i_2}$,..., k_ℓ from some ℓ-th community $\mathcal{C}_{i_\ell}$, where the indices i_j are mutually distinct. We will always implicitly assume that $k_i \leq h$ for each i. Let $P_{\mathbf{k}}$ be the uniform measure on $\Omega_B\left(n,\ \mathbf{k}\right)$. Let $T \geq 1$ be some positive integer, and let P be a convex combination $P = \sum_{t=1}^{T} p_t \cdot P_{\mathbf{k}_t}$, where $p_t > 0$ for each t, $\sum_{t=1}^{T} p_t = 1$, and the vectors $\mathbf{k}_t = (k_{1t}, \dots, k_{\ell t})$ are mutually distinct. That is, P is a measure on $\bigcup_{t=1}^{T} \Omega_B\left(n,\ \mathbf{k}_t\right)$. Using similar notations to [Giráldez-Cru and Levy, 2015], denote by $F\left(n, m, B, P\right)$ the probability space of instances, with the sample space $\left(\bigcup_{t=1}^{T} \Omega_B\left(n,\ \mathbf{k}_t\right)\right)^m$ and the measure P^m. Instances in the model presented in [Giráldez-Cru and Levy, 2015] include clauses of two types: (i) all variables belong to the same community, and (ii) the variables belong to distinct communities. For some $0 < p < 1$, each clause is of type (i) with probability p and type (ii) with probability $1 - p$. With the above notation, their probability space is

$$F\left(n, m, B, p \cdot P_{(b)} + (1 - p) \cdot P_{(\underbrace{1, \dots, 1}_{b})}\right)$$

for some b. The instance $(\overline{v}_1 \vee v_2 \vee v_3) \wedge (v_2 \vee \overline{v}_6)$ is an instance in

$$F\left(9, 2, 3, 0.2 \cdot P_{(1,1)} + 0.8 \cdot P_{(3)}\right).$$

In particular, $F\left(n, m, 1, P_{(k)}\right)$ is the regular model of random k-SAT. Denote by $F\left(n, m, B, P\right)$-SAT the satisfiability problem, in which the probability space of the instances is $F\left(n, m, B, P\right)$.

As explained above, the clauses in an industrial instance tend to include variables from the same community. In this paper, we deal with the case where one (or more) of the programs $P_{\mathbf{k}_t}$, $1 \le t \le T$, takes clauses whose variables are from a single community, namely $\mathbf{k}_t = (k)$ for some t and k. In some results, we will further restrict ourselves to the case $T = 1$ of a single program.

3 The Main Results

In [Giráldez-Cru and Levy, 2015] it was observed empirically that, when the modularity of the variable incidence graph (VIG) of the instance increases, the threshold decreases. Now, the modularity in our case is larger when more clauses consist of variables from the same community and when the communities are small. Our first result is quite straightforward, but it already hints that instances in the model suggested in Section 2 tend to be no more satisfiable than random k-SAT instances. Note that the first part of the proposition is one of the initial results for random SAT [Franco and Paull, 1983] .

Proposition 3.1. *Let $\mathcal{I}$ be a random instance in $F(n, \alpha n, B, P)$.*

(a) Suppose that for each $P_{\mathbf{k}_t}$, $1 \le t \le T$, the clause length is at most k. If $\alpha > 2^k \ln 2$, then $\mathcal{I}$ is UNSAT w.h.p.

(b) Let $P = P_{(k)}$ for some $k \ge 2$.

(i) If $\alpha > r_k^*$, then $\mathcal{I}$ is UNSAT w.h.p.

(ii) If $h(n) = \Theta(n)$ and $\alpha < r_k$, then $\mathcal{I}$ is SAT w.h.p.

Our next result points out a significant difference between random instances and community-structured ones. One might expect the threshold to be different for community-structured instances, but it turns out that this difference may be not just quantitative. The following result shows that, surprisingly, under certain conditions the satisfiability threshold is 0. To this end, we will consider m as some function of n, not necessarily $m = \alpha n$, and write $m(n)$ instead of m.

For real functions f and g, we write $f = \Omega(g)$ if $g = O(f)$, and $f = \omega(g)$ if $g = o(f)$. We also write $f = \text{polylog}(g)$ if $f = O\left(\ln^\theta g\right)$ for some θ.

Theorem 3.2. *Let $\mathcal{I}$ be a random instance in $F(n, m(n), B, P)$, where $\mathbf{k}_t = (k)$ for some $1 \le t \le T$.*

(a) Let $h(n) = O(1)$.

(i) If $P = P_{(k)}$ for some k and $m(n) = o(n^{1-1/2^k})$, then $\mathcal{I}$ is SAT w.h.p.

(ii) If $m(n) = \Omega(n^{1-1/2^k})$, then $\mathcal{I}$ is UNSAT with probability bounded away from 0.

(iii) If $m(n) = \omega(n^{1-1/2^k})$, then $\mathcal{I}$ is UNSAT w.h.p.

(b) If $h(n) = o\left(\frac{\ln n}{\ln\ln n}\right)$ and $m(n) = \Omega\left(\frac{n}{\text{polylog}(n)}\right)$, then $\mathcal{I}$ is UNSAT w.h.p.

(c) If $h(n) = o(\ln n)$ and $m(n) = \Omega\left(n \cdot e^{-\beta\cdot\ln n/h(n)}\right)$ for some $\beta < 1/r_k^*$, then $\mathcal{I}$ is UNSAT w.h.p.

(d) Let $h(n) = O(\ln n)$ and $T = 1$. Then there exists some $\varepsilon_0 > 0$ such that, if $m(n) = \alpha n$ with $\alpha > r_k^* - \varepsilon_0$, then $\mathcal{I}$ is UNSAT w.h.p.

Parameters			Results
$h(n)$	$m(n)$	T	
$O(1)$	$o\left(n^{1-1/2^k}\right)$	1	SAT w.h.p.
$O(1)$	$\Omega\left(n^{1-1/2^k}\right)$	$*$	$\in (0, 1-\delta)$
$O(1)$	$\omega\left(n^{1-1/2^k}\right)$	$*$	UNSAT w.h.p.
$o\left(\frac{\ln n}{\ln\ln n}\right)$	$\Omega\left(n/\text{polylog}(n)\right)$	$*$	UNSAT w.h.p.
$o(\ln n)$	$\Omega\left(n^{1-1/(r_k^*+\varepsilon)h(n)}\right)$	$*$	UNSAT w.h.p.
$O(\ln n)$	$(r_k^* - \varepsilon)\, n$	1	UNSAT w.h.p.

Table 1: Asymptotic satisfiability of a random instance with small communities in $F(n, m(n), B, P)$.

Remark 3.3. *(a) The ε_0 in part (d) is effective. Namely, as will follow from the proof, one can present such an ε_0 explicitly (in terms of the implicit constant in the equality $h(n) = O(\ln n)$).*

(b) Still in case (d), one can deal with the general case of arbitrary T as long as the weight of the program (k) in P is sufficiently large.

In Theorem 3.2 there are three types of results for the asymptotic satisfiability of a random community-structured instance with n variables, $m(n)$ clauses, B communities of size $h(n) = n/B$, and probability measure P. Namely, either the probability of satisfiability tends to 0 as $n \to \infty$, or it tends to 1, or it is bounded away from 1. These results are summarized in Table 1. In general, we assume that $\mathbf{k}_t = (k)$ for some $1 \le t \le T$ and $k \ge 1$. In the third column we place a '1' or a '$*$', depending on whether T is required to be 1 or is arbitrary, respectively. The notation $\in (0, 1-\delta)$ indicates a probability bounded away from 1.

The proof of Theorem 3.2 will use the following lemma.

Lemma 3.4. *Consider the spaces $F(n, m(n), B, P)$ and $F(n, m'(n), B, P)$, where $m'(n) = \omega(m(n))$. If a random instance in $F(n, m(n), B, P)$ is UNSAT with probability bounded away from 0, then a random instance $\mathcal{I}'$ in $F(n, m'(n), B, P)$ is UNSAT w.h.p.*

In the proof of Theorem 3.2 (and that of Theorem 3.6) we use some results regarding the classical "balls and bins" problem. In this problem, there are M balls and B bins. Each ball is placed uniformly randomly into one of the bins, independently of the other balls. One quantity of interest is the *maximum load*, which is the maximum number of balls in any bin. There are several papers studying the size of the maximum load, as well as generalizations of this problem. It seems that [Raab and Steger, 1998] contains all previous results. Our next result seems not to be covered by previous results regarding the balls and bins problem. It will be employed in the proof of Theorem 3.2, and may be of independent interest.

Proposition 3.5. *Consider the balls and bins problem with B bins and $M = M(B)$ balls. Let $s \geq 1$ be an arbitrarily fixed integer.*

(a) If $M(B) = o(B^{1-1/s})$, then the maximum load is at most $s-1$ w.h.p.

(b) If $M(B) = \Omega(B^{1-1/s})$, then the maximum load is at least s, with probability bounded away from 0.

(c) If $M(B) = \omega(B^{1-1/s})$, then the maximum load is at least s w.h.p.

As noted earlier, random 2-SAT is much better understood than random k-SAT for general k. This enables us to obtain a stronger result than Theorem 3.2 in the case $P = P_{(2)}$.

Theorem 3.6. *Let $\mathcal{I}$ be a random instance in $F\left(n, \alpha n, B, P_{(2)}\right)$.*

(a) There exists an $0 < \varepsilon_0 < 1$ such that, if $h(n) = o\left(\sqrt{n}\right)$ and $\alpha > 1 - \varepsilon_0$, then $\mathcal{I}$ is UNSAT w.h.p.

(b) For $h(n) = \Theta(\sqrt{n})$:

(i) If $1 - \varepsilon_0 < \alpha < 1$, where ε_0 is as in (a), then $\mathcal{I}$ is SAT with probability bounded away from both 0 and 1.

(ii) If $\alpha = 1$ then $\mathcal{I}$ is UNSAT w.h.p.

(c) For $h(n) = \omega(\sqrt{n})$ with $h(n) = o(n)$:

(i) If $\alpha < 1$ then $\mathcal{I}$ is SAT w.h.p.

(ii) If $\alpha = 1$ then $\mathcal{I}$ is UNSAT w.h.p.

(d) For $h(n) = \Theta(n)$:

(i) If $\alpha < 1$ then $\mathcal{I}$ is SAT w.h.p.

(ii) If $\alpha = 1$ then $\mathcal{I}$ is SAT with probability bounded away from both 0 and 1.

$\boldsymbol{\alpha}$ $\boldsymbol{h(n)}$	$\in (1-\varepsilon_0, 1)$	$= 1$
$= o\left(\sqrt{n}\right)$	UNSAT w.h.p	UNSAT w.h.p.
$= \Theta\left(\sqrt{n}\right)$	$\in (\delta, 1-\delta)$	UNSAT w.h.p.
$\in \omega\left(\sqrt{n}\right) \cap o\left(n\right)$	SAT w.h.p.	UNSAT w.h.p
$= \Theta\left(n\right)$	SAT w.h.p	$\in (\delta, 1-\delta)$

Table 2: Asymptotic satisfiability of a random instance in $F\left(n, \alpha n, B, P_{(2)}\right)$.

Remark 3.7. *Similarly to Remark 3.3.b, one can deal with the more general case of arbitrary T, as long as one of the programs $P_{\mathbf{k}_t}$ is of the form (k) and is of sufficiently large weight.*

Similarly to Table 1, we summarize the results of the theorem 3.6 in Table 2. Here, we always assume $k = 2$, $m(n) = \alpha n$ and $T = 1$. Similarly to Theorem 3.2, we have three types of results for the asymptotic satisfiability of a random instance in $F\left(n, \alpha n, B, P_{(2)}\right)$, two of which are the same as in Table 1. Namely, either the probability of satisfiability tends to 0 as $n \to \infty$, or it tends to 1, or it is bounded away both from 0 and 1. In the latter case we shall use the notation $\in (\delta, 1-\delta)$.

4 Proofs

Proof of Proposition 3.1:

(a) We follow the proof in the random model [Franco and Paull, 1983]. Fix a truth assignment and consider $\mathcal{I}$. Each clause has at most k literals. The variables are negated with probability $1/2$ independently of each other, and hence each clause is satisfied with probability of $1 - 2^{-k}$, independently of the other clauses. The expected number of satisfying truth assignments is therefore $2^n \cdot \left(1 - 2^{-k}\right)^{\alpha n}$. As $\alpha > 2^k \ln 2$, we have

$$2^n \cdot \left(1 - 2^{-k}\right)^{\alpha n} \xrightarrow[n \to \infty]{} 0.$$

Thus, $\mathcal{I}$ is UNSAT w.h.p.

(b) A random instance $\mathcal{I}$ in $F\left(n, \alpha n, B, P_{(k)}\right)$ decomposes into B sub-instances $\mathcal{I}_i$, $1 \leq i \leq B$, where each $\mathcal{I}_i$ is formed of those clauses consisting of variables solely from $\mathcal{C}_i$. Obviously, $\mathcal{I}$ is SAT if and only if all $\mathcal{I}_i$-s are such. For $1 \leq i \leq B$, let $U_i = 1$ if $\mathcal{I}_i$ is satisfiable, and $U_i = 0$ otherwise. The variable $U = \prod_{i=1}^{B} U_i$ indicates whether $\mathcal{I}$ is satisfiable. Let W_i denote the number of clauses in $\mathcal{I}_i$. Clearly,

$$W_i \sim B\left(\alpha n, 1/B\right), \qquad 1 \leq i \leq B.$$

(i) Suppose first that $h(n) = \omega(1)$. Let α_i denote the density of the sub-instance $\mathcal{I}_i$, $1 \leq i \leq B$. There exists an i with $\alpha_i \geq \alpha$, and therefore $\alpha_i > r_k^*$. It follows that $\mathcal{I}_i$ is UNSAT w.h.p., and hence so is $\mathcal{I}$. The case $h(n) = O(1)$ follows in particular from part (a.iii) of this theorem (to be proved below).

(ii) In this case $B(n) = \Theta(1)$. Without loss of generality assume $B(n) = B$ is fixed. For $\gamma > 0$, let

$$W_{<\gamma} = \bigcap_{i=1}^{B} \left\{W_i < \gamma \cdot h(n)\right\}. \tag{1}$$

Let α' be an arbitrary fixed number, strictly between α and r_k. Then:

$$\begin{aligned} P\left(U = 1\right) &= P\left(W_{<\alpha'}\right) P\left(U = 1 \left| W_{<\alpha'}\right.\right) \\ &\quad + P\left(\overline{W}_{<\alpha'}\right) P\left(U = 1 \left| \overline{W}_{<\alpha'}\right.\right) \\ &\geq P\left(W_{<\alpha'}\right) P\left(U = 1 \left| W_{<\alpha'}\right.\right). \end{aligned} \tag{2}$$

By the weak law of large numbers,

$$\frac{W_i}{n/B} \xrightarrow[n\to\infty]{P} \alpha, \qquad 1 \le i \le B,$$

and therefore

$$\begin{aligned} P\left(\overline{W}_{<\alpha'}\right) &= P\left(\bigcup_{i=1}^{B}\{W_i \ge \alpha' \cdot n/B\}\right) \\ &\le \sum_{i=1}^{B} P\left(W_i \ge \alpha' \cdot n/B\right) \\ &= B \cdot P\left(W_1 \ge \alpha' \cdot n/B\right) \xrightarrow[n\to\infty]{} 0. \end{aligned}$$

Hence

$$P\left(W_{<\alpha'}\right) = 1 - P\left(\overline{W}_{<\alpha'}\right) \xrightarrow[n\to\infty]{} 1.$$

Now consider the second factor on the right-hand side of (2). Clearly,

$$\begin{aligned} P\left(U = 1 \,|W_{<\alpha'}\,\right) &\ge \prod_{i=1}^{B} P\left(U_i = 1 \,|W_i = \alpha' \cdot n/B\,\right) \\ &= P\left(U_1 = 1 \,|W_1 = \alpha' \cdot n/B\,\right)^B . \end{aligned} \tag{3}$$

As $\alpha' < r_k$ and B is fixed, the right-hand side of (3) converges to 1 as $n \to \infty$. Hence $\mathcal{I}$ is SAT w.h.p.

□

As mentioned in Section 3, the proofs of Theorem 3.2 and Theorem 3.6 make use of some results concerning the balls and bins problem. Let L be the maximum load for M balls and B bins. By [Raab and Steger, 1998], for any $\delta > 0$,

$$L \ge \begin{cases} \dfrac{\ln B}{\ln \frac{B \ln B}{M}}, & \frac{B}{\text{polylog}(B)} < M = o(B \ln B), \\ (d_c - \delta)\ln B, & M = c \cdot B \ln B, \end{cases} \tag{4}$$

w.h.p.for an appropriate constant $d_c > c$.

Remark 4.1. *The constant d_c, in the second part of (4), is the unique solution of the equation*

$$1 + x(\ln c - \ln x + 1) - c = 0$$

in (c, ∞) (see [Raab and Steger, 1998, Lemma 3]). A routine calculation shows that $d_c = c + c \cdot u(1/c)$, where the function u is the unique non-negative function defined implicitly by the equation

$$-u(w) + (1 + u(w))\ln(1 + u(w)) = w, \qquad (w \ge 0).$$

The function $u(w)$ has been studied in [Kolchin et al., 1978, pp. 101–102], and in particular expressed as a power series in $\sqrt{w}$ near 0.

The fact that $d_c > c$ is the reason that the threshold in Theorem 3.2.d is strictly less than r_k^*. One can easily bound $d_c - c$ from below. In fact, write $d_c = c + \varepsilon$. Then

$$\begin{aligned} 1 &= -(c+\varepsilon)(\ln c - \ln(c+\varepsilon) + 1) + c \\ &= -(c+\varepsilon)\ln c + (c+\varepsilon)\ln(c+\varepsilon) - \varepsilon \\ &< (c+\varepsilon)\cdot\varepsilon/c - \varepsilon = \varepsilon^2, \end{aligned}$$

and hence $d_c > c + \sqrt{c}$.

Proof of Theorem 3.2: We follow the notations used in the proof of Proposition 3.1. Recall that $\mathcal{I}_i$ is the sub-instance formed of those clauses in $\mathcal{I}$ consisting of variables solely from $\mathcal{C}_i$, and W_i is the number of clauses in $\mathcal{I}_i$, $1 \le i \le B$. Denote $W_{\max} = \max\{W_1, \ldots, W_B\}$.

We may assume that $T = 1$ in all parts of the theorem. In fact, in the general case, if the weight of the program $P_{(k)}$ in P is p, then w.h.p. there will be at least $\frac{p}{2} \cdot m(n)$ clauses consisting of clauses of type (k). To see it, denote by $\mathcal{I}^{(k)}$ the sub-instance of $\mathcal{I}$ obtained by taking the clauses from the program of type (k) and by $m'(n)$ the number of clauses in $\mathcal{I}^{(k)}$. Note that $m'(n)$ is $B(m(n), p)$-distributed. Employing Chernoff's bound we deduce a lower bound $p \cdot m(n)/2$ on $m'(n)$ w.h.p. Since $p \cdot m(n)$ tends to infinity with n, $\mathcal{I}^{(k)}$ is UNSAT w.h.p., which implies that so is $\mathcal{I}$.

Since $T = 1$, each clause corresponds to some community. Consider clauses as balls, and communities as bins. The process of selecting the clauses, as far as the community to which the variables in each clause belong, is analogous to that of placing $m(n)$ balls in B bins uniformly at random. The idea of the proof in parts (b)-(d) will be to prove that w.h.p. we have $W_{\max}/h(n) > r_k^*$. This will imply that there is at least one sub-instance $\mathcal{I}_i$ with density larger than r_k^*. Thus, already $\mathcal{I}_i$ is UNSAT w.h.p., and consequently so is $\mathcal{I}$.

(a) Without loss of generality, assume that $h(n) = h > 0$ is fixed.

(i) By Proposition 3.5.a, there is no sub-instance with more than $2^k - 1$ clauses w.h.p. Since instances with less than 2^k clauses are certainly satisfiable w.h.p., all $\mathcal{I}_i$-s are SAT, and hence so is $\mathcal{I}$.

(ii) By Proposition 3.5.b, the probability that there is an $\mathcal{I}_i$, $1 \le i \le B$, with at least 2^k clauses, is bounded away from 0. Assume, say, that $W_1 \ge 2^k$. Then, with probability at least

$$\left(1/\binom{h}{k}\right)^{2^k} \cdot (2^k)!/2^{k2^k},$$

all 2^k distinct clauses consisting of the variables $v_1, \ldots, v_k$ have been drawn. As the instance is UNSAT if it contains all these 2^k clauses, the probability for our instance to be UNSAT is bounded away from 0.

(iii) Follows from the previous part and Lemma 3.4.

(b) In view of part (a.iii), we may assume $h(n) \to \infty$. We may also assume that $m(n) = n/\ln^\theta n$ for some $\theta \ge 1$. Clearly, $m(n) \le B$. On the other hand,

$$m(n) = \frac{n}{\ln^\theta n} \ge \frac{B}{(2\ln B)^\theta} = \frac{B}{\text{polylog}(B)}.$$

Thus, by (4), w.h.p., the maximum load is at least

$$\frac{\ln B}{\ln \dfrac{B \ln B}{m(n)}} \ge \frac{\frac{1}{2}\cdot \ln n}{\ln\left(\dfrac{n \cdot \ln n}{n/\ln^\theta n}\right)} \ge \frac{\ln n}{2(\theta+1)\ln\ln n}. \tag{5}$$

Now, there are $h(n) = o(\ln n / \ln\ln n)$ variables in each community. By (5), w.h.p., the density of the sub-instance $\mathcal{I}_i$ with the maximal number of clauses is at least

$$\frac{W_{\max}}{h(n)} \geq \frac{\frac{1}{2(\theta+1)} \cdot \ln n / \ln\ln n}{o\left(\ln n / \ln\ln n\right)} \xrightarrow[n\to\infty]{} \infty.$$

Hence, this $\mathcal{I}_i$ is UNSAT w.h.p., and therefore so is $\mathcal{I}$.

(c) By (4), w.h.p., the number of clauses in the sub-instance $\mathcal{I}_i$ with the maximal number of clauses is at least

$$W_{\max} \geq \frac{\ln B}{\ln \dfrac{B \ln B}{m(n)}} = \frac{\ln n(1 - o(1))}{\ln \dfrac{B \ln B}{m(n)}}.$$

For a large enough n

$$\ln \frac{B \ln B}{m(n)} \leq \ln \left(\frac{\dfrac{n}{h(n)} \cdot \ln n}{n \cdot e^{-\beta \ln n / h(n)}} \right)$$
$$= \ln \frac{\ln n}{h(n)} + \beta \cdot \frac{\ln n}{h(n)}.$$

As $\beta < 1/r_k^*$, for large enough x we have $\ln x + \beta x < x / r_k^*$. Hence, for large enough n we have

$$\ln \frac{B \ln B}{m(n)} \leq \frac{1}{r_k^*} \cdot \frac{\ln n}{h(n)}.$$

This implies that the density of the sub-instance $\mathcal{I}_i$ with the maximal number of clauses is at least

$$\frac{W_{\max}}{h(n)} \geq \frac{1}{\ln \dfrac{B \ln B}{m(n)}} \cdot \frac{\ln n(1 - o(n))}{h(n)} > r_k^*,$$

and thus UNSAT w.h.p. Consequently, so is $\mathcal{I}$.

(d) In view of the previous part, we may assume that $h(n) = \theta \ln n$ for some $\theta > 0$. Choose c_0 such that

$$d_{c_0} = \theta r_k^*,$$

where d_c is as in (4). Let $\alpha > c_0/\theta$, and put $c = \alpha\theta$. Thus, $c > c_0$ and $d_c > d_{c_0}$. Let $\delta < d_c - \theta r_k^*$. We have

$$m(n) = n \cdot \alpha = \frac{nc}{\theta} = (1 + o(1)) \cdot \frac{nc \ln B}{\theta \ln n}$$
$$= (c + o(1)) \cdot B \ln B.$$

By (4), the size of the largest sub-instance is $W_{\max} \geq (d_c - \delta) \ln B$ w.h.p. Hence, w.h.p. the density of this sub-instance is

$$\frac{W_{\max}}{h(n)} \geq \frac{(d_c - \delta) \ln B}{h(n)} = \frac{(d_c - \delta) \cdot (1 - o(1)) \ln n}{\theta \ln n}$$
$$= \frac{d_c - \delta}{\theta} - o(1) = r_k^* + \frac{d_c - \delta - \theta r_k^*}{\theta} - o(1).$$

Letting $\varepsilon_0 = r_k^* - c_0/\theta$, we get our claim. □

Let us make one comment regarding the proof of Theorem 3.6, which we cannot bring here. The fact that Theorem 3.6 is much more precise than Theorem 3.2 hinges on the fact that the information as to the satisfiability of 2-SAT for various densities is more detailed than for k-SAT in general. In fact, [Bollobás et al., 2001] proved that there exist some $0 < \varepsilon_0 < 1$ and $\lambda_0 > 0$ such that the satisfiability probability of a random 2-SAT instance $\mathcal{I}$ with $m = n \cdot (1 + \varepsilon)$ clauses is

$$P\left(\mathcal{I} \text{ is SAT}\right) = \begin{cases} 1 - \Theta\big(\dfrac{1}{n\left|\varepsilon\right|^3}\big), & -\varepsilon_0 \leq \varepsilon \leq \frac{-\lambda_0}{n^{1/3}}, \\ \Theta\left(1\right), & \frac{-\lambda_0}{n^{1/3}} < \varepsilon < \frac{\lambda_0}{n^{1/3}}, \\ \exp\left(-\Theta\left(n\varepsilon^3\right)\right), & \frac{\lambda_0}{n^{1/3}} \leq \varepsilon \leq \varepsilon_0. \end{cases}$$

(Actually, we use only the first two cases.) Note that in the case $m = n \cdot (1 - \varepsilon)$ with $\lambda_0 n^{-1/3} \leq \varepsilon \leq \varepsilon_0$, we have

$$1 - \frac{\theta_1}{n \cdot \varepsilon^3} \leq P\left(\mathcal{I} \text{ is SAT}\right) \leq 1 - \frac{\theta_2}{n \cdot \varepsilon^3}$$

for some constants $\theta_1, \theta_2 > 0$.

5 Conclusions

We have dealt with the satisfiability threshold of a particular model of SAT. This model highlights one of the features in which so-called industrial SAT instances differ from classical SAT instances. Namely, the set of variables decompose into several disjoint subsets-communities.The significance of these communities stems from the fact that clauses tend to contain variables from the same community. We have shown, roughly speaking, that the satisfiability threshold of such instances tends to be lower than for regular instances. Moreover, if the communities are very small, the threshold may even vanish.

The paper leaves a lot to study for industrial SAT instances. To begin with, there are other features considered in the literature as being characteristic of industrial instances. For example, in the scale free structure the variables are selected by some heavy-tailed distribution. Moreover, even regarding the issue of communities, there is more to be done. We assumed here that all communities are of the same size. Obviously, there is no justification for this assumption beyond the fact that it simplifies significantly the analysis of the model. What can be said about the threshold if there are both small and large communities? Even prior to that, what would be reasonable to assume regarding the probability of a variable to be selected from each of the communities?

Acknowledgments

We would like to express our gratitude to David Wilson for helpful information regarding the topic of this paper, as well as to the referees for their comments on the first version of the paper.

References

[Achlioptas and Peres, 2004] Dimitris Achlioptas and Yuval Peres. “The threshold for random k-SAT is $2^k \ln 2 - O(k)$”.
Journal of the American Mathematical Society 17.4 (2004), 947–973.

[Ansótegui et al., 2009] Carlos Ansótegui, María L. Bonet and Jordi Levy. “Towards industrial-like random SAT instances”. Proc. of the 21st International Joint Conference on Artificial Intelligence, IJCAI 2009 (2009).

[Ansótegui et al., 2012] Carlos Ansótegui, Jesús Giráldez-Cru and Jordi Levy. “The community structure of SAT formulas”. Proc. of Theory and Applications of Satisfiability Testing–SAT 2012: 15th International Conference, Trento, Italy, June 17–20, 2012, A. Cimatti, R. Sebastiani, Eds., 410–423.

[Bollobás et al., 2001] Béla Bollobás, Christian Borgs, Jennifer T. Chayes, Jeong H. Bim and David B. Wilson. “The scaling window of the 2-SAT transition”. Random Structures & Algorithms 18 (2001), 3, 201–256.

[Chvátal and Reed, 1992] Vaclav Chvátal and B. Reed, “Mick gets some (the odds are on his side)”, Proc. 33rd Symp. Foundations of Computer Science (1992), 620–627.

[Cook, 1971] Stephen A. Cook. “The complexity of theorem proving procedures”. Proc. of the Third Annual ACM STOC (1971), 151–158.

[Coja-Oghlan and Panagiotou, 2016] Amin Coja-Oghlan and Konstantinos Panagiotou. “The asymptotic k-SAT threshold”. Advances in Mathematics 288 (2016), 985–1068.

[Ding et al., 2015] Jian Ding, Allan Sly and Nike Sun. “Proof of the satisfiability conjecture for large k”. (English summary) STOC’15—Proc. of the 2015 ACM Symposium on Theory of Computing (2015), 59–68.

[Dubois et al., 2000] Olivier Dubois, Yacine Boufkhad, and Jacques Mandler. “Typical random 3-SAT formulae and the satisfiability threshold”. Proc. 11th ACM-SIAM Symp. on Discrete Algorithm (2000), 126–127.

[Fernandez de la Vega, 1992] Wenceslas Fernandez de la Vega, On random 2-SAT, unpublished manuscript (1992).

[Franco and Paull, 1983] John V. Franco and Marvin C. Paull. “Probabilistic analysis of the Davis–Putnam procedure for solving the satisfiability problem”. Discrete Applied Mathematics 5(1) (1983), 77–87.

[Giráldez-Cru and Levy, 2015] Jesús Giráldez-Cru and Jordi Levy. “A modularity-based random SAT instances generator”. Proc. of the 24th International Joint Conference on Artificial Intelligence, IJCAI’15 (2015), 1952–1958.

[Giráldez-Cru and Levy, 2016] Jesús Giráldez-Cru and Jordi Levy. “Generating SAT instances with community structure”. Artificial Intelligence 238 (2016), 119–134.

[Giráldez-Cru and Levy, 2017] Jesús Giráldez-Cru, Jordi Levy. “Locality in Random SAT Instances”. Proc. of the 26th International Joint Conference on Artificial Intelligence, IJCAI’17 (2017), 638–644.

[Greenberg and Mohri, 2014] Spencer Greenberg and Mehryar Mohri. “Tight lower bound on the probability of a binomial exceeding its expectation”. Statistics & Probability Letters 86 (2014), 91–98.

[Goerdt, 1992] Andreas Goerdt. “A threshold for unsatisfiability”, Proc. Mathematical Foundations of Computer Science 1992: 17th International Symp. Prague, Czechoslovakia, August 24–28, 1992. 629 Springer, Berlin (1992) 264–274.

[Kaporis et al., 2002] Alexis C. Kaporis, Lefteris M. Kirousis, and Efthimios G. Lalas. “The probabilistic analysis of a greedy satisfiability algorithm”. Proc. of the 10th Annual European Symp. on Algorithms, Lecture Notes in Computer Science 2461, Springer-Verlag, Berlin (2002), 574–585.

[Kirousis et al., 1998] Lefteris M. Kirousis, Evangelos Konstantinou Kranakis, Danny Krizanc, and Y. C. Stamatiou. “Approximating the unsatisfiability threshold of random formulas”. Random Structures & Algorithms 12(3) (1998), 253–269.

[Kolchin et al., 1978] Valentin F. Kolchin, Boris A. Sevastýanov and Vladimir P. Chistyakov. “Random allocations”. V. H. Winston & Sons, Washington, D.C., (1978).

[Mertens et al., 2006] Stephan Mertens, Marc Mézard and Riccardo Zecchina. “Threshold values of random k-SAT from the cavity method”. Random Structures & Algorithms 28 (2006), 340–373.

[Mézard and Zecchina, 2002] Marc Mézard and Riccardo Zecchina. “Random k-satisfiability problem: From analytic solution to an efficient algorithm”. Physical Review E 66 (2002), 056126.

[Newman, 2006] Mark E. J. Newman. “Finding community structure in networks using the eigenvectors of matrices”. Physical Review E 74(3) (2006), 036104.

[Newsham et al., 2014] Zack Newsham, Vijay Ganesh, Sebastian Fischmeister, Gilles Audemard, and Laurent Simon. "Impact of community structure on SAT solver performance", In Theory and applications of satisfiability testing—SAT 2014, Lecture Notes in Comput. Sci., 8561, Springer (2014), 252– 268.

[Park and Newman, 2003] Juyong Park and Mark E. J. Newman, Physical Review E 68 (2003), 026112.

[Raab and Steger, 1998] Martin Raab and Angelika Steger. “Balls into bins–a simple and tight analysis”. J. D. P. Rolim, M. Serna and M. Luby, Eds., Randomization and Approximation Techniques in Computer Science, 1518, Springer, Berlin (1998), 159–170.

[Wilson, 1998] David B. Wilson, http://dbwilson.com/2sat-data/ (1998).