Rigorous Results for Random $(2+p)$-SAT

Dimitris Achlioptas[*§]
(optas@cs.toronto.edu)

Lefteris M. Kirousis[†¶]
(kirousis@ceid.upatras.gr)

Evangelos Kranakis[‡‖]
(kranakis@scs.carleton.ca)

Danny Krizanc[‡‖]
(krizanc@scs.carleton.ca)

Abstract

Consider random k-SAT instances with rn clauses over n variables, where each clause is chosen uniformly and independently from the set of all clauses of length k. It is widely believed that for all $k \geq 2$ there exists r_k such that random instances are almost surely satisfiable for $r < r_k$ and almost surely unsatisfiable for $r > r_k$. For $k = 2$ this was proven to be true in [6, 13] and moreover it was shown that $r_2 = 1$. For $k > 2$ only lower and upper bounds are known for almost sure (un)satisfiability are known. Experimental evidence suggests that the value of r_3 is around 4.2. As an attempt to "interpolate" between $k = 2$ and $k = 3$ the $(2+p)$-SAT model was proposed in [22]. In this model, as in random k-SAT, there are rn clauses over n variables, but now prn clauses have length 3 and $(1-p)rn$ have length 2, where $p \in [0, 1]$. In this paper we prove the first rigorous results for $(2+p)$-SAT and they turn out to provide a very tight "envelope" for the heuristic results of [22]. Among others, we show that if $p \leq 2/5$ then a random instance of $(2+p)$-SAT is almost surely (un)satisfiable if r is smaller (greater) than $1/(1-p)$, i.e., for $p \leq 2/5$ the clauses of length 3 are "irrelevant".

[*]Department of Computer Science, University of Toronto, Toronto, ON M5S 3G4, Canada.
[†]Department of Computer Engineering and Informatics, University of Patras, Rio, 265 00 Patras, Greece.
[‡]School of Computer Science, Carleton University, Ottawa, ON K1S 5B6, Canada.
[§]Supported by a PGS B Scholarship from NSERC (Natural Sciences and Engineering Research Council) of Canada.
[¶]Partially supported by the EU ESPRIT Long-term Research Project ALCOM-IT (Project Nr. 20244).
[‖]Supported in part by an NSERC grant.

1 Introduction

The problem of determining the satisfiability of boolean formulae in conjuctive normal form has played a central role in the understanding of the complexity of computations. In addition, it is of tremendous practical interest, arising naturally in a variety of applications. Given its NP-hardness, practitioners seek heuristic solutions to the problem of verifying the satisfiability of large formulae, or enumerating all satisfying truth assignments.

In the last few years there has been increased interest in the analysis of the behaviour of random formulae. Typically, it is assumed that the formula is in CNF and the clauses are chosen uniformly and independently. Mitchell et al [19] pointed out that some commonly used distributions tend to generate random formulae that are uninteresting since they are almost always very easy to satisfy. They also reported that the model where each formula has precisely $k \geq 3$ literals per clause, can generate some very hard instances. In particular, they reported that for $k = 3$ the following remarkable behaviour is observed in experiments: let r denote the ratio of clauses to variables. For $r < 4$, almost all formulae are satisfiable and a satisfying truth assignment is easy to find. For $r > 4.5$ almost all formulae are unsatisfiable. Finally, a "50%-point" seems to appear around 4.2, the same point where the computational complexity of finding a satisfying truth assignment is maximized.

More precisely, we will consider random k-SAT formulae with rn clauses over n variables, where each clause is chosen uniformly and independently from the space of all clauses of length k. Note that, in general, the problem is solvable in linear time for $k = 2$, but it is NP-complete for $k \geq 3$. The main driving force for studying this problem has been the following conjecture: *For all $k \geq 2$ there exists a critical value r_k such that for any $\epsilon > 0$, a random instance of k-SAT with $(r_k - \epsilon)n$ clauses is almost surely* (a.s.) satisfiable while a random instance with $(r_k + \epsilon)n$ clauses is a.s. unsatisfiable.* The following summarizes our state of knowledge regarding the problem:

For $k = 2$, Chvatál and Reed [6], Goerdt[13] and Fernandez de la Vega [9] independently proved the conjecture, in fact showing $r_2 = 1$. Note that a problem being solvable in polynomial time is equivalent to saying that we have a characterization of the witnesses of (un)satisfiability. This fact was of great help in determining r_2 since one can "look" for the emergence of specific structures in the evolution of a random formula.

For $k \geq 3$ much less progress has been made. Not only has r_k not been determined, but its existence has not been established either. In the following we will provide "upper and lower bounds for r_k", meaning that for r smaller (greater) than each bound random formulae of k-SAT formulae are a.s. (un)satisfiable. Naturally, the case $k = 3$ has received most attention. In [10] it was observed that for $r > 5.191$ the expected number of satisfying truth assignments tends to 0 and hence, so does the probability of satisfiability†. El-Maftouhi and Fernandez de la Vega [8] lowered this to 5.08. Kamath et al. [14] lowered 5.08 to 4.758 by proving that if a random 3-SAT formula is satisfiable then, with extremely high probability, it has an exponential number of satisfying truth assignments. Hence, providing some account for the "wastefulness" of the first moment method. Finally, Kirousis et al [16] introduced a general refinement of the first moment method whereby one counts not all the satisfying truth assignments ("bad" objects) but only those that specify a local maximality condition. We will use their technique in section 2, and elaborate on it therein. With this method they showed [16] that $r_3 < 4.598\ldots$, which is the best bound currently known. (Independently, Dubois et al. [7] using a similar method obtained $r_3 < 4.64$.)

*We will say that a sequence of events $\mathcal{E}_n$ occurs almost surely if $\lim_{n\to\infty} \Pr[\mathcal{E}_n] = 1$.

†This idea of using the expectation of a non-negative variable, to bound its probability of being positive is known as the *first moment method*.

Unlike upper bounds for r_3 which are probabilistic counting arguments the lower bounds are, so far, entirely algorithmic. A first bound of $r_3 > 2/3$ follows from the analysis of [4] and we will see more on this bound in section 3, since it turns out to be relevant to $(2+p)$-SAT. Chao and Franco [4] analyzed the UNIT CLAUSE (UC) algorithm for 3-SAT (with and without a "majority" rule) and showed that for $r < 2.9$ and $r < 8/3$, respectively, it finds a satisfying truth assignment with positive probability. Note, that this does not imply 2.9 as a lower bound for r_3 since the algorithm succeeds only with positive probability instead of a.s. Broder et al [3] analyzed the "pure literal" rule and showed that it a.s. succeeds for $r < 1.63$, establishing $r_3 \geq 1.63$. Finally, Frieze and Suen [12], using some very sophisticated analysis, analyzed a generalization of UC and showed that it succeeds a.s. for $r < 3.003$. This gives the currently best known lower bound for r_3.

For general k, an upper bound of $r_k = O(2^k)$ can be proved by the first moment method while the technique of [16] gives a constant factor improvement for this bound. On the other hand, a lower bound of $r_k = \Omega(2^k/k)$ is given in [6, 5] and it is improved by a constant factor in [12]. Determining the asymptotic order of r_k seems an interesting, and challenging, open problem.

A recent and very important development pertaining to monotone properties in general (such as satisfiability) is due to Friedgut [11] who provided a characterization of properties with sharp thresholds. In particular for k-SAT he shows the following: let $f_p(n,r)$ denote the probability that a random instance of k-SAT with n variables and rn clauses has a satisfying truth assignment. There exists $r_k(n)$ such that for any $\epsilon > 0$,

$$\lim_{n\to\infty} f_p(n, r_k(n) - \epsilon) = 1 \quad \text{and} \quad \lim_{n\to\infty} f_p(n, r_k(n) + \epsilon) = 0 \ .$$

Note, that r_k is a function of n and need not converge. Hence, the above does not establish the existence of r_k (which would be $\lim_{n\to\infty} r_k(n)$). On the other hand it does imply that if for some value of r the probability of satisfiability for a random k-SAT instance is $\epsilon > 0$, where ϵ is independent of n then, in fact, the probability is $1 - o(1)$. Hence, the values of [4] are in fact lower bounds for r_3 (slightly superceded, though, by the value provided in [12]).

The $(2+p)$-SAT model was introduced by Monasson et al. [22] in order to illuminate the study of the transition from 2-SAT to 3-SAT and is defined as follows. Consider n Boolean variables and a random CNF formula $F = F_2 \wedge F_3$ consisting of m clauses such that F_2 (resp., F_3) is a random 2-CNF (resp., 3-CNF) with $(1-p)m$ (resp., pm) clauses. F_2 and F_3 are drawn independently from each other and p is a real number $0 \leq p \leq 1$. Notice that the case $p = 0$ corresponds to 2-SAT while the case $p = 1$ corresponds to 3-SAT.

In [22] methods from statistical mechanics are used to derive predictions for the value of the threshold for $(2+p)$-SAT. Monasson et al. observe that the known bounds for 2 and 3-SAT imply the critical value for $(2+p)$-SAT is bounded above by $\min\{1/(1-p), r_3/p\}$. Using the "replica method" [18, 22] they predict that for $p < .413...$, $1/(1-p)$ is also a lower bound, i.e., for values of p in the range 0 to .413..., $(2+p)$-SAT behaves in a similar manner as 2-SAT.

In this paper we provide a rigorous proof of this assertion for $p \leq 2/5$ by showing that the threshold for values of p between 0 and $2/5$ is indeed $1/(1-p)$, i.e., the clauses of length 3 are "irrelevant" to the critical behaviour of $(2+p)$-SAT in this range. We also provide improved upper and lower bounds for the threshold of $(2+p)$-SAT for values of $p > 2/5$. Our results are summarized in the following theorem:

Theorem 1 *Let Π be a random instance of $(2+p)$-SAT.*

a. If $p \leq 2/5$ then Π is a.s. satisfiable if $r < 1/(1-p)$ and a.s. unsatisfiable if $r > 1/(1-p)$.

b. If $p > 2/5$ then Π is a.s. satisfiable if $r < 24p/(p+2)^2$ and a.s. unsatisfiable if

$$(7/8)^{rp}(3/4)^{r(1-p)}(2 - e^{-2/3(1-5p/14)}) < 1$$

In the next section we discuss bounds for almost sure satisfiability establishing the upper bound of part b of the theorem. In section 3, the lower bound of the theorem is derived. Finally we provide an intuitive explanation of where the value 2/5 comes from and suggest possible future work.

2 Almost sure unsatisfiability

For a set of n variables V, let $\mathcal{B}_i(V)$ denote the set of all $2^i\binom{n}{i}$ possible clauses of length i, on the variables of V. Unless, otherwise stated we will consider $V = \{x_1, \ldots, x_n\}$ and write $\mathcal{B}_i$ for $\mathcal{B}_i(V)$. For a literal l, $v(l)$ will denote the corresponding variable.

From [6, 16] we know that a random instance of $(2+p)$-SAT is a.s. unsatisfiable for $r > \min\{1/(1-p), 4.598/p\}$ simply by combining the bounds for 2- and 3-SAT. It is easily seen that for $p \leq 0.821$ we have that $\min\{1/(1-p), 4.598/p\} = 1/(1-p)$. In the following we will devise a bound that is tighter for values of $p \geq 0.6945....$

Let Π be a random instance of $(2+p)$-SAT with n variables and rn clauses and let $\mathcal{S}_n$ denote the random set of solutions of Π. It is straightforward to see that

$$\mathbf{E}[|\mathcal{S}_n|] = 2^n(7/8)^{rpn}(3/4)^{r(1-p)n} = c(r,p)^n \ .$$

Hence, by the first moment method, Π is a.s. unsatisfiable if $c(r,p) < 1$. That is, if

$$r > \frac{\ln 2}{\ln 4/3 - p \ln 7/6} \ .$$

The right hand side above is less than $1/(1-p)$ for $p \geq 0.752\ldots$

The price paid for the simplicity of the first moment method is that instances of $(2+p)$-SAT with a very large number of solutions, although they may occur with very small probability, contribute substantially to $E[|\mathcal{S}_n|]$. In [16], a refinement of the first moment method was introduced which amounts to "compressing" $\mathcal{S}_n$ by requiring truth assignments not only to be solutions of Π but to also satisfy a certain "local maximality" condition. The underlying intuition is that for a random solution of Π, if we choose a variable at random and change its value, the probability we will end up with a non-solution is rather small. Consequently, solutions (when they exist) tend to appear in large "clusters". The crux of the technique is that instead of counting all the solutions in a cluster we need only count a representative one (locally maximum).

Definition 1 *The set $\mathcal{S}_n^\sharp$ is defined as the random set of truth assignments A such that*

1. *A is a solution of Π, and*
2. *any truth assignment obtained from A by changing the value of exactly one variable from 0 to 1, is not a solution of Π.*

For a truth assignment A, and a variable x assigned value 0 by A, $A(x)$ will denote the truth assignment obtained from A, by changing the value of x from 0 to 1. The number of possible such changes for a truth assignment A (i.e., the number of 0s A contains) will be denoted by $sf(A)$ (for "single flips"). From [16] we have that

$$\Pr[\Pi \text{ has a solution}] \leq E[|\mathcal{S}_n^\sharp|] \tag{1}$$

and

$$E[|\mathcal{S}_n^\sharp|] = \Pr[A \in \mathcal{S}_n] \sum_{A \in \mathcal{A}_n} \Pr[A \in \mathcal{S}_n^\sharp \mid A \in \mathcal{S}_n] \ . \tag{2}$$

As we saw in the application of the first moment method,

$$\Pr[A \in \mathcal{S}_n] = (7/8)^{rpn}(3/4)^{r(1-p)n} \ . \tag{3}$$

We will now compute an upper bound to $\Pr[A \in \mathcal{S}_n^\sharp \mid A \in \mathcal{S}_n]$.

Fix a truth assignment $A \in \mathcal{S}_n$. Conditioning on $A \in \mathcal{S}_n$ exclude $\binom{n}{3}$ clauses from $\mathcal{B}_3$ and $\binom{n}{2}$ clauses from $\mathcal{B}_2$ from the conjucts of Π. Now, consider changing the value of variable x from 0 to 1. The event $A(x) \notin \mathcal{S}_n$ occurs iff among the rn clauses in Π, there is a clause that contains $\overline{x}$ and is not satisfied by $A(x)$. In particular,

$$\begin{aligned}
\Pr[A(x) \notin \mathcal{S}_n] &= 1 - \left(1 - \frac{\binom{n-1}{2}}{7\binom{n}{3}}\right)^{rpn} \left(1 - \frac{\binom{n-1}{1}}{3\binom{n}{2}}\right)^{r(1-p)n} \\
&= 1 - \left(1 - \frac{3}{7n}\right)^{rpn} \left(1 - \frac{2}{3n}\right)^{r(1-p)n} \\
&= 1 - e^{-2/3r(1-5p/14)} + o(1) \ .
\end{aligned}$$

Let $\mathcal{A}_i$ denote the event $A(x_i) \notin \mathcal{S}_n$ and let $\mathcal{C}(x_i)$ denote the set of clauses associated with (that can cause) $\mathcal{A}_i$. The events $\mathcal{A}_i$ are not independent but $\mathcal{C}(x_i) \cap \mathcal{C}(x_j) = \emptyset$ for $i \neq j$. Intuitively, the occurence of $\mathcal{A}_i$ "exposes" only clauses that cannot contribute to any $\mathcal{A}_j$, $i \neq j$. Moreover, it exposes that at least one of the rn clauses was "used" to cause $\mathcal{A}_i$. Hence, the events $\mathcal{A}_i$ are in fact negatively correlated. Formally, this fact follows from the main Theorem of [17]. This implies,

$$\Pr[A \in \mathcal{S}_n^\sharp \mid A \in \mathcal{S}_n] \leq \left(1 - e^{-2/3(1-5p/14)} + o(1)\right)^{sf(A)} \ . \tag{4}$$

Hence, from (1),(2),(3),(4) we have

$$\begin{aligned}
\Pr[\Pi \text{ has a solution}] &\leq (7/8)^{rpn}(3/4)^{r(1-p)n} \sum_{A \in \mathcal{A}_n} \Pr[A \in \mathcal{S}_n^\sharp \mid A \in \mathcal{S}_n] \\
&\leq (7/8)^{rpn}(3/4)^{r(1-p)n} \sum_{k=0}^{n} \binom{n}{k} \left(1 - e^{-2/3r(1-5p/14)} + o(1)\right)^k \\
&= \left((7/8)^{rp}(3/4)^{r(1-p)}(2 - e^{-2/3r(1-5p/14)} + o(1))\right)^n \ .
\end{aligned}$$

Thus, if

$$(7/8)^{rp}(3/4)^{r(1-p)}(2 - e^{-2/3r(1-5p/14)}) < 1 \ ,$$

then Π is a.s. unsatisfiable.
For $p \geq 0.6945\ldots$ the solution of

$$(7/8)^{rp}(3/4)^{r(1-p)}(2 - e^{-2/3r(1-5p/14)}) = 1$$

is less than $1/(1-p)$.

3 Almost sure satisfiability

Let $f_p(n,r)$ denote the probability that a random instance of $(2+p)$-SAT with n variables and rn clauses has a satisfying truth assignment.

Lemma 1 *For all $p \in [0,1]$, there exists $r_p(n)$ such that for any $\epsilon > 0$,*

$$\lim_{n\to\infty} f_p(n, r_p(n) - \epsilon) = 1 \quad \textit{and} \quad \lim_{n\to\infty} f_p(n, r_p(n) + \epsilon) = 0 \ .$$

Hence, for all $r \geq 0, p \in [0,1]$ and $\epsilon > 0$,

$$\textit{if} \ \lim_{n\to\infty} f_p(n,r) \geq \epsilon \ \textit{then} \ \lim_{n\to\infty} f_p(n,r) = 1 - o(1) \ .$$

Proof The proof is essentially identical to the proof of the existence of a sharp threshold for random k-SAT [11] so we only present an outline, highlighting the differences.

Let U_p denote the property of unsatisfiability for $(2+p)$-SAT, i.e., the set of all $(2+p)$-SAT instances that are unsatisfiable. For $\mathcal{L} = \Pi_1, \ldots, \Pi_l$ a fixed list of $(2+p)$-SAT formulae, let $C_{\mathcal{L}}$ denote the set of all $(2+p)$-SAT formulae that contain as a subformula (a copy of) some formula from $\mathcal{L}$. Using theorem 5.1 of [11] it follows that the lemma is false iff: for every $\epsilon > 0$ there exists a constant $l = l(\epsilon)$ and a list $\Pi_1, \ldots, \Pi_l$ such that for a random formula Π of $(2+p)$-SAT

$$\Pr[\Pi \in U_p \mid \Pi \in C_{\mathcal{L}}] \geq 1 - \epsilon \ .$$

In other words, unsatisfiability can be "approximated" by the property of containing some formula from a short, fixed list of formulae. To prove that this last statement leads to a contradiction it suffices to observe that all analogous steps of the proof in [11] can be repeated verbatim, up to the final lemma 5.7. For this final step, by analogy, we should take the codimension to be 3 with probability p and 2 with probability $1-p$, in each repetition. Taking the codimension to be 3 in all repetitions allows the lemma to be applied verbatim. Since by doing this, the expected number of clauses needed for coverage is increased only by a constant factor, the lemma still holds. □

By lemma 1, in order to prove that for some fixed r, p, random instances of $(2+p)$-SAT with rn clauses are a.s. satisfiable, it suffices to show that they are satisfiable with probability $\epsilon > 0$, where ϵ is independent of n. To show the latter, we will in fact show how to find a satisfying truth assignment, in polynomial time, with probability $\epsilon > 0$.

The algorithm we will use, Unit Clause (uc), was introduced and analyzed for random 3-SAT by Chao and Franco [4]. Solely for the sake of self-containment we present the algorithm in the following. The algorithm makes n iterations, setting permanently one variable in each one. In the following, "at time t" will mean after t such iterations have been performed (i.e., after t variables have been set). We will let $\mathcal{C}_i(t)$ denote the set of clauses of length i at time t.

Unit Clause (without "majority")

- $V \leftarrow \{x_1, ..., x_n\}$.
- `For` $t = 0, \ldots, n-1$
  1. `If` $\mathcal{C}_1(t) \neq \emptyset$ `then` choose a random clause (literal) $\ell \in \mathcal{C}_1(t)$
     `else` choose a random variable $v \in V$ and
     take $\ell = v, \ell = \overline{v}$ with equal probability.

2. Set $v(\ell)$ so as to satisfy ℓ.
3. $V \leftarrow V - v(\ell)$.
4. Remove all clauses containing ℓ.
5. Remove $\overline{\ell}$ from all clauses (i.e., "shrink" all clauses containing $\overline{\ell}$).

At each substep of type (4) UC might generate a clause of length 0 and, clearly, such a clause will never be removed. On the other hand, if this never happens, i.e., $\mathcal{C}_0(n) = \emptyset$, then UC finds a satisfying truth assignment, in which case we say that it *succeeds*.

Let $C_i(t) = |\mathcal{C}_i(t)|$. Lemmata 2,3, below, were proven in [4] and regard the behaviour of UC while processing a random instance Π of 3-SAT. We will provide the intuition behind the proof for both of them, making it easy to see that they generalize to $(2+p)$-SAT.

Lemma 2 [4] *Let $V(t)$ denote the set of unset variables at time t. For all $0 \le t \le n$, $\mathcal{C}_i(t)$ is distributed as a set of $C_i(t)$ clauses chosen uniformly and independently from $\mathcal{B}_i(V(t))$.*

Intuition: The lemma is true at $t = 0$ by the construction of Π. At each step, every clause c either becomes satisfied and is removed, or precisely one literal is removed from c, or c is not affected. Hence, after a step where variable v is set, the only new thing exposed about the remaining clauses is that they do not contain $v, \overline{v}$. The lemma follows inductively.

Observation 1 *Lemma 2 remains true as long as each $\mathcal{C}_i(0)$ is distributed as a set of $C_i(0)$ clauses chosen uniformly and independently from $\mathcal{B}_i(V)$. In particular, it is true if Π is a random instance of $(2+p)$-SAT.*

Using lemma 2 it can be shown that

Lemma 3 [4] *Let $\Delta C_i(t) = C_i(t+1) - C_i(t)$. For all $0 \le t \le n-3$,*

$$\begin{aligned}
\mathbf{E}[\Delta C_3(t)] &= -\frac{3}{n-t}\, C_3(t) \\
\mathbf{E}[\Delta C_2(t)] &= \frac{3}{2(n-t)}\, C_3(t) - \frac{2}{n-t}\, C_2(t)
\end{aligned}$$

.

Intuition: For each term we apply linearity of expectation to multiply a probability (that of containing the variable being set) with the number of choices (clauses of length i). Each negative term represents the expected number of clauses leaving $\mathcal{C}_i(t)$ during step t, either as satisfied or as "shrunk". The positive term denotes the fact that half the clauses leaving $\mathcal{C}_3(t)$, leave it as "shrunk" and hence end up in $\mathcal{C}_2$. The negative terms follow by applying lemma 2 to $\mathcal{C}_1$ and noting that any choices made regarding which variable to set are random ones. Hence, in every step a random variable is set. For the positive term it suffices to add that to the previous argument that the choice of which "way" to set a variable, if it exists, is also random.

Lemma 4, to follow, is based on lemma 3 and it provides us with two functions that "approximate" $C_3(t), C_2(t)$ very well and a.s. . The underlying idea is to use $\mathbf{E}[\Delta C_3(t)], \mathbf{E}[\Delta C_2(t)]$ as the derivatives of two "deterministic" functions, solve the corresponding differential equations and use the solutions as approximations of $C_3(t), C_2(t)$. This use of differential equations to describe the "mean path" of Markov chains (in this case defined implicitly) has found many applications including [1, 4, 5, 12, 15, 20, 23, 24]. Although the step is quite intuitive, the justification of the fact that the chain a.s. stays "close" to its mean path, and hence the solutions of the differential equations provide a good approximation, is highly nontrivial. A rigorous, unifying framework regarding the applicability of this technique was developed by Wormald in [24]. The following lemma follows directly from the main theorem of [24], lemmata 2,3 and observation 1. A similar, but less precise, statement for 3-SAT ($p = 1$) appears in [4].

Lemma 4 *Assume that functions $c_3(x), c_2(x)$ satisfy the following differential equations and boundary conditions.*

$$c_3'(x) = -\frac{3}{1-x} c_3(x) \quad , \quad c_3(0) = rp \ . \tag{5}$$

$$c_2'(x) = \frac{3}{2(1-x)} c_3(x) - \frac{2}{1-x} c_2(x) \quad , \quad c_2(0) = r(1-p) \ . \tag{6}$$

Let Π be a random instance of $(2+p)$-SAT with n variables and rn clauses. If UC *is applied to Π then a.s. for all $0 \leq t \leq n$, $i = 2, 3$*

$$C_i(t) = c_i(t/n)\, n + o(n) \ . \tag{7}$$

The following lemma states that if the rate at which "forced" variables are generated is low enough (less than one per step) then UC has a positive probability of success.

Lemma 5 [4] *Let $U(t)$ be the random variable defined as the number of clauses that shrink to unit length, during step t. Let $g(t) = \mathbf{E}[U(t)]$.*
(i) For all $0 \leq t \leq n-2$,

$$g(t) = \frac{1}{n-t} C_2(t) \ .$$

(ii) If there is $\delta > 0$, such that for all $0 \leq t \leq n-2$,

$$g(t) \leq 1 - \delta$$

then UC *succeeds with probability $\rho = \rho(\delta) > 0$.*

The unique solution of differential equations (and boundary conditions) (5),(6) of lemma 4 is

$$\begin{aligned} c_3(x) &= rp(1-x)^3 \\ c_2(x) &= 1/2r(3px - 2p + 2)(1-x)^2 \ . \end{aligned} \tag{8}$$

Combining (7),(8) and lemma 5 we get that for a random instance Π of $(2+p)$-SAT with rn clauses, UC succeeds with positive probability if for all $x \in [0, 1]$,

$$1/2r(3px - 2p + 2)(1-x) < 1 \ . \tag{9}$$

For $p \leq 2/5$, the left hand side of (9) is non-increasing with x and hence (9) holds iff

$$r < \frac{1}{1-p} \ . \tag{10}$$

For $p > 2/5$, the left hand side of (9) is unimodal for $x \in [0, 1]$, the unique maximum occuring for $x = (5p-2)/6$. Substituting this value of x in (9) we get

$$r < \frac{24p}{(p+2)^2} \ .$$

The discussion above completes the proof of Theorem 1.

4 Discussion

Let S_p denote the set of all satisfiable $(2+p)$-SAT formulae and $S2_p$ denote the set of all $(2+p)$-SAT formulae whose 2-SAT subformula is satisfiable. Clearly, $S_p \subset S2_p$. From (10) we see that for a *randomly chosen* $(2+p)$-SAT formula on n vertices, where $p \leq 2/5$,

$$\Pr[\Pi \in S_p \mid \Pi \in S2_p] = 1 - o(1) \ .$$

In the following we try to give an "intuitive explanation" of why this is so.

Given a set of clauses $\mathcal{C} \subseteq \mathcal{B}_3$, consider the 3-uniform hypergraph on vertices $x_1, \ldots, x_n$ formed by introducing for each clause $c_i = (l_{i1} \vee l_{i2} \vee l_{i3}) \in \mathcal{C}$, a hyperedge $\{v(l_{i1}), v(l_{i2}), v(l_{i3})\}$. Hence, for a random 3-SAT formula with rn clauses the corresponding hypergraph is the random 3-uniform hypergraph $G(n, m = rn)$. It is well-known [2] that for a random k-uniform hypergraph $G(n,m)$ the birth of a "giant component" occurs precisely at $m = (k-1)/k\,n$. That is, for any $\epsilon > 0$, if $m = (1-\epsilon)(k-1)/k\,n$ then a.s. all the components of $G(n,m)$ have size logarithmic in n and moreover they have at most one cycle. Translating this last statement back to 3-SAT we get that if $r < 2/3$ then a.s. in each clause c of length 3 there is at least one variable v such that c is the unique clause that v appears in. Such a 3-SAT formula is trivial to satisfy and this gives the first lower bound for r_3, as we saw in section 1. Note that such a 3-SAT formula imposes very "local" constraints on the variables, since all connected components are very small.

Note, now, that for $p \leq 2/5$, having $r < 1/(1-p)$ implies $rp < 2/3$, while $rp = 2/3$ occurs *precisely* when $p = 2/5$ and $r = 1/(1-p)$. In other words: the "breakpoint" for UC occurs exactly when p is such that the 3-SAT clauses can form a "giant component". We feel that determining whether there exists $p > 2/5$, ex. $0.413\ldots$, such that for $r < 1/(1-p)$, random $(2+p)$-SAT formulae are a.s. satisfiable, is a worthwhile goal for future work.

References

[1] D. Achlioptas, and M. Molloy, *The analysis of a list-coloring algorithm on a random graph*, submitted.

[2] N. Alon, and J.H. Spencer, and P. Erdös, **The Probabilistic Method,** J. Wiley, New York (1992).

[3] A.Z. Broder, and A.M. Frieze, and E. Upfal, *On the satisfiability and maximum satisfiability of random 3-CNF formulae*, Proceedings of the 4th ACM-SIAM Symposium on Discrete Algorithms, (1993) 322–330.

[4] M.-T. Chao, and J. Franco, *Probabilistic analysis of two heuristics for the 3-satisfiability problem*, SIAM Journal on Computing, **15** (1986), 1106–1118.

[5] M.-T. Chao, and J. Franco, *Probabilistic analysis of a generalization of the unit-clause literal selection heuristic for the k-satisfiability problem*, Information Science, **51** (1990), 289–314.

[6] V. Chvátal, and B. Reed, *Mick gets some (the odds are on his side)*, Proceedings of 33rd IEEE Symposium on Foundations of Computer Science, (1992) 620–627.

[7] O. Dubois, and Y. Boutkhad, *A General Upper Bound for the Satisfiability Threshold of Random r-SAT Formulas*, Preprint, LAFORIA, CNRS-Université Paris 6, 1996.

[8] A. El Maftouhi and W. Fernandez de la Vega, *On random 3-SAT*, Combinatorics, Probability and Computing, **4** (1995), 189–195.

[9] W. Fernandez de la Vega, *On random 2-SAT*, manuscript (1992).

[10] J. Franco, and M. Paull, *Probabilistic analysis of the Davis Putnam procedure for solving the satisfiability problem*, Discrete Applied Mathematics, **5** (1983), 77–87.

[11] E. Friedgut, *Necessary and sufficient conditions for sharp thresholds*, preprint (1997).

[12] A. Frieze, and S. Suen, *Analysis of two simple heuristics on a random instance of k-SAT*, Journal of Algorithms, **20** (1996), 312–355.

[13] A. Goerdt, *A threshold for unsatisfiability*, Journal of Computer and System Sciences, **53** (1996), 469–486.

[14] A. Kamath, R. Motwani, K. Palem, and P. Spirakis, *Tail bounds for occupancy and the satisfiability threshold conjecture*, Random Structures and Algorithms, **7** (1995), 59–80.

[15] R. M. Karp, M. Sipser, *Maximum matchings in sparse random graphs*, Proceedings of the 22nd IEEE Symposium on Foundations of Computer Science, (1981) 364–375.

[16] L. M. Kirousis, and E. Kranakis, and D. Krizanc, *Approximating the unsatisfiability threshold of random formulae*, Proceedings of the 4th European Symposium on Algorithms, (1992) 27–38.

[17] C. McDiarmid, *On a correlation inequality of Farr*, Combinatorics, Probability and Computing, **1** (1992), 157–160.

[18] M. Mézard, G. Parisi, and M. Virasoro, **Spin Glass Theory and Beyond**, World Scientific, Singapore, 1987.

[19] D. Mitchell, and B. Selman, and H. Levesque, *Generating hard satisfiability problems*, Artificial Intelligence, **81** (1996), 17–29.

[20] M. Molloy, and B. Reed, *The dominating number of a random cubic graph*, Random Structures and Algorithms, **7** (1995), 209–221.

[21] R. Monasson, and R. Zecchina, *Statistical Mechanics of the random K-SAT model*, to appear in Physical Review E.

[22] R. Monasson, and R. Zecchina, and S. Kirkpatrick, and B. Selman, and L. Troyansky, *Phase Transition and Search Cost in the* $2+p$*-SAT problem*, Presented in 4th Workshop on Physics and Computation, Boston University 22-24 November 1996, (PhysComp96).

[23] B. Pittel, J. Spencer, N. Wormald, *Sudden emergence of a giant k-core in a random graph*, Journal of Combinatorial Theory Series B, **67** (1996), 111–151.

[24] N. C. Wormald, *Differential equations for random processes and random graphs*, Annals of Applied Probability, **5** (1995), 1217–1235.