Computers and Artificial Intelligence, Vol. 20, 2001, 1–17

# ON DIFFERENT MODELS FOR GENERATING RANDOM SAT PROBLEMS

Predrag Janičić
Nenad Dedić
Goran Terzić

*Faculty of Mathematics, University of Belgrade,*
*Studentski trg 16, 11 000 Belgrade, Yugoslavia*
*email:* {janicic, dedic, terzic}@matf.bg.ac.yu

**Abstract.** In the last decade a lot of effort has been invested into both theoretical and experimental analysis of SAT phase transition. However, a deep theoretical understanding of this phenomenon is still lacking. Besides, many of experimental results are based on some assumptions that are not supported theoretically. In this paper we introduce the notion of SAT–equivalence and we prove that some restrictions often used in SAT experiments don't make an impact on location of a crossover point. We consider several fixed and random clause length SAT models and relations between them. We also discuss one new SAT model and report on a detected phase transition for it.

**Keywords.** Phase transition, SAT problems, NP–complete problems

## 1 INTRODUCTION

In recent years, phase transition for many NP-hard problems has been the subject of both theoretical and experimental consideration (some of the first were the influential papers by Cheesman *et. al.* and by Mitchell *et. al.* published in early 1990s [2, 15]). A prototypical example of such problems is propositional satisfiability problem — SAT (SAT is the problem of deciding if there is an truth assignment for which a given propositional formula is evaluated to true; it was shown by Cook that SAT is NP-complete problem [3]). We focus on SAT problems in conjunctive normal form: $(N, L)$-SAT problem consists of $L$ clauses over the set of $N$ variables and their negations (in the rest of the paper, by SAT problem we mean problem of this form). By $M_i(N, L)$ we denote sets of $(N, L)$ formulae satisfying some additional syntactical restriction (say, with no multiple occurrences of some clause). Many experiments (over problem sets with different additional syntactical constraints) suggest that there is

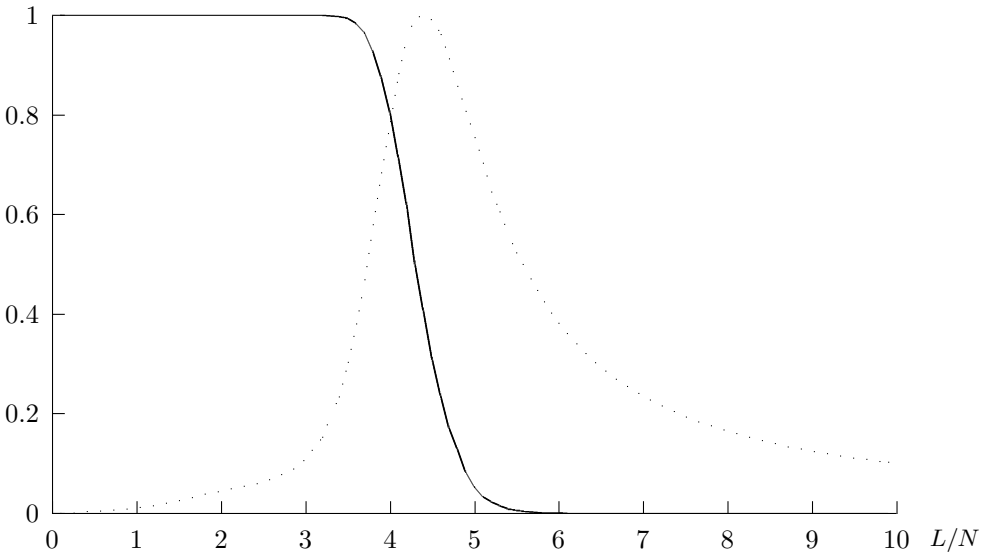


Fig. 1. Satisfiability and computational cost function for 3-SAT problem set (the solid line represents the satisfiability function, and the dashed line represents the (normalized) computational cost)

a phase transition in SAT problems between satisfiability and unsatisfiability as the ratio $L/N$ is increased. For different types of problem sets $M(N, L)$, it is conjectured that there is a critical value $c_0$ of $L/N$, which we call a *crossover point* (or a *phase transition point*) such that:

$$\lim_{N\to\infty} s(M(N, [cN])) = \begin{cases} 1, & \text{for } c < c_0 \\ 0, & \text{for } c > c_0 \end{cases}$$

where $s$ is a *satisfiability function* that maps sets of propositional formula into the segment $[0, 1]$ and corresponds to a percentage of satisfiable formulae. The value of the crossover point might be (and often is) different for different types of problem sets. For a fixed problem set, according to the properties of the crossover point, the sequence of points $L/N$ in which the satisfiability function is (approximately) equal $p$ (where $0\% < p < 100\%$), converges to the crossover point as $N$ increases; in most of the experiments, crossover points for different SAT problems is estimated (usually using $p = 50\%$) on the basis of this fact. Additionally, experimental results suggest that at the crossover point approximately the same percentage of formulae is satisfiable for all large of $N$ (while that percentage depends on SAT model examined) [14, 7].

For most of SAT problem sets $M(N, L)$ it is easy to show that the function $s(M(N, L))$ is strictly decreasing in its second argument and $\lim_{L\to\infty} s(M(N, L)) = 0$. Thus, clearly, if the crossover point exists, there is only one such point (for one SAT problem set). As yet, for none of SAT problem sets the crossover point has been theoretically computed nor even proved that it exists (with the only exception of 2-SAT problem). However, recent Friedgut's results [6] serve as a major step towards solving this problem: he proved that the transition region for $k$-SAT problems narrows as the number of variables increases (despite that, as Friedgut says, it is still feasible

that, though there is a swift transition of the satisfiability function, the critical value does not converge to any given value).

Experimental results also suggest that in all SAT problems there is a typical easy-hard-easy pattern as the ratio $L/N$ is increased. Indeed, for small values of $L/N$, problems are *under-constrained* and (relatively) easy for all propositional decision procedures because there are many satisfying assignments; for large values of $L/N$, problems are *over-constrained*, thus (relatively) easily shown to be unsatisfiable. Interestingly, the most difficult SAT problems for all decision procedures for propositional logic are those in the crossover region (see Fig. 1; Figure 1 shows experimental results for SAT formulae with all lengths of clauses equal 3; we call that model 3-SAT model and the estimated value of the corresponding crossover point is 4.25 [4]). All known decision procedures for propositional logic are of exponential worst-case complexity. Decision procedures most often considered in SAT experiments are Davis-Logemann-Loveland's procedure (often misattributed to Davis and Putnam), resolution based procedures and tableau based procedures. This paper mostly discusses satisfiability function and we are not much concerned by behavior of particular decision procedures.

Most of SAT experiments assume that some additional assumptions on a set of formulae examined can not significantly change the location of a crossover point. In all related papers we are aware of, a problem of different restrictions for generating random SAT problems is just tackled: for instance, in experiments conducted using the most often model it is not checked if generated problems do have $N$ variables and do have $L$ different clauses; usually, it is assumed that related restrictions could not have significant impact on final results of experiments. In this paper, we discuss this issue. We introduce the notion of SAT-*equivalence* as a methodology for exploring if two models lead to the same crossover point. It could be a step closer to deeper understanding of SAT phase transition phenomenon. We also give one sufficient condition for two models to be SAT-equivalent and illustrate it on some typical examples. This methodology can be important both for theoretical and practical purposes. For instance, given a proof that some value is the crossover point for one type of SAT problem set, then there are proofs for crossover points for all types of problems sets that are SAT-equivalent with the first one. This methodology can also be used to support some assumptions in experiments made so far. Besides, some known experimental results would be useful for different real–world problems satisfying different constraints.

Another problem which we discuss in this paper is relation between different fixed and random clause length SAT models. We briefly discuss one new SAT model and report on a detected phase transition. We also discuss relationship between the fixed and the random clause length SAT models.

*Overview of the paper.* In Sect. 2 we discuss most often models used in random generation of SAT problems and in Sect. 3, we describe different types of SAT problem sets (sets of SAT problems meeting different syntactical restrictions). In Sect. 4 we introduce the notion of SAT–equivalence and in Sect. 5 we prove that some often used types of SAT problem sets are SAT–equivalent. In Sect. 6 we introduce one new random clause length SAT model and report on the phase transition detected. In Sect. 7

we discuss the relation between the fixed and random clause length SAT models. In Sect. 8 we discuss further work and in Sect. 9 we draw some final conclusions.

## 2 MODELS FOR GENERATING RANDOM SAT PROBLEMS

Most of SAT experiments are conducted in the following way: for some $N$ and for $L/N$ varied by some constant (usually between 0.01 and 0.1), generate randomly (large) number (usually between 1000 and 100000) of formulae of some SAT corpora $M(N, L)$; for large samples of formulae, a percentage of satisfiable formulae approximates the satisfiability function $s(M(N, L))$. Usually, it is not checked if some formula occurs more than once. The crossover point for the model $M$ is usually determined in the following way: for each $N$ there is approximated a critical point at which there are 50% satisfiable formulae from the set $M(N, L)$ (actually, instead of 50%, it can be taken any percentage other than 0% and 100%); the sequence of these critical points converges to the crossover point as $N$ is increased.

The following models are used most often (the first one is a fixed clause length model, while the remaining three are random clause length models):

**Random $k$-SAT model (Fixed clause length model):** For given values $N$ and $L$, an instance of random $k$-SAT formula is produced by randomly generating $L$ clauses of length $k$. Each clause is produced by randomly choosing $k$ distinct variables from the set of $N$ available variables, and negating each with probability 0.5 [15]. It is known that $k$-SAT is NP–complete for natural numbers $k$ such that $k > 2$. There is a polynomial decision procedure for 2-SAT problem (i.e., 2-SAT $\in P$), but still there is a phase transition as for $k$-SAT problems for $k > 2$. It is proved that the crossover point for 2-SAT problems is 1 [8]. For random 3-SAT the phase transition occurs at $L/N \approx 4.25$ [4]. For random 4-SAT the phase transition occurs at $L/N \approx 9.76$ [7]. For large $k$, Kirkpatrick and Selman estimate the crossover points for $k$-SAT at $L/N = -1/\log_2(1 - 1/2^k)$ [13]. It has been shown theoretically that the crossover point for 3-SAT is (if it exists) between 3.003 and 4.87 [12]. Friedgut proved that the transition region for $k$-SAT problems narrows as the number of variables increases [6].

**Constant probability model:** In this model [9], given $N$ variables and $L$ clauses, each clause is generated so that it contains each of $2N$ different literals with probability $p$. Some experiments use a variant of this model: if an empty clause or a unit clause is generated, it is discarded and another clause is generated in its place. Parameter $p$ can be chosen such that $2Np = 3$ and then the mean clause length remains approximately constant as $N$ varies [7]. It is shown that there is a phase transition between satisfiability and unsatisfiability for constant probability model as $L/N$ is varied and for $2Np = 3$, the crossover point is approximated as $L/N \approx 2.80$ [7].

**Random mixed SAT:** In this model [7], each clause is generated as in random $k$-SAT except that $k$ (the length of clauses), is chosen randomly according to a

finite probability distribution $\phi$ on integers. For instance, if $\phi(2) = 1/3$ and $\phi(4) = 2/3$, clauses of length 2 appear with the probability 1/3 and clauses of length 4 with the probability 2/3 (this problem is then called $2, 4, 4$-SAT). For random $2, 4, 4$-SAT, the phase transition occurs at $L/N \approx 2.74$ [7].

$2 + p$-SAT **model** In this model [16], a formula with $L$ clauses has (approximately) $(1 - p)L$ clauses of the length 2 and $pL$ clauses of the length 3 (this model is closely related to the random mixed SAT and can be considered as its special case). Hence, a model smoothly interpolates between 2-SAT and 3-SAT model. Crossover points are approximated for different values between 0 and 1. For $p \leq 0.4$ it has been proved that the crossover point is at $L/N = 1/(1 - p)$ [1]. In addition, $2 + p$-SAT behaves as 2-SAT for $p \leq 0.4$ and as 3-SAT for $p > 0.4$.

## 3 TYPES OF CORPORA OF SAT PROBLEMS

In this section we discuss several different restrictions on corpora of SAT problems. Some of these restrictions are used in most of SAT experiments. These restrictions are rather general and applicable to different classes of SAT problems. The corpora of $(N, L)$–SAT problems are considered ($N$ is a number of variables, $L$ is a number of clauses in problem). We discuss the following restrictions:

1. only formulae with all clauses having exactly $k$ literals are considered ($k$ is a fixed value);
2. only formulae with all clauses different are considered;
3. only formulae with all $N$ variables are considered;
4. only formulae with all clauses not containing multiple occurrences of some variable (either in positive or in negative form) are considered.

We will by $M_1(N, L)$ a corpus meeting restriction 1, by $M_{1,2}(N, L)$ a corpus meeting restrictions 1 and 2 etc. For instance, if $d = (2N)^k$ ($k$ is a fixed value), then there are $d^L$ formulae in $M_1(N, L)$; $d \cdot (d - 1) \cdots (d - L + 1)$ formulae in $M_{1,2}(N, L)$ etc.

In every fixed clause length problem all clauses contain the same number of literals (so the first restriction is met). The random $k$-SAT model corresponds to the corpus $M_{1,4}(N, L)$; note that it is not checked if generated formula does have $N$ variables and $L$ literals. In other words, in the set of formulae generated in this way, there are some formulae which are not from the corpus $M_{1,2,3,4}(N, L)$. The question is whether these two corpora lead to the same crossover point. We discuss this question in the further text.

The random clause length problems are based on some distribution of clause lengths. In this paper we consider both fixed clause length problems and random clause length problems and relations between them.

## 4 SAT-EQUIVALENCE BETWEEN DIFFERENT TYPES OF CORPORA

In this section, we introduce a SAT-*equivalence* relation over the set of different types of corpora of SAT problems. Two types of corpora are in the same class if their satisfiability functions are asymptotically (as the number of variables increases) the same. In this section we also give one sufficient condition for two types of corpora to be in the same class.

**Definition 4.1.** We say that the value $c_0$ is a *crossover* point for a corpora of the type $M$ if it holds that

$$\lim_{N\to\infty} s(M(N,[cN])) = \begin{cases} 1, & \text{for } c < c_0 \\ 0, & \text{for } c > c_0 \end{cases}$$

($c$ ranges over the set of positive reals numbers; $N$ ranges over the set of natural numbers).

**Definition 4.2.** We say that two types of corpora $M'$ and $M''$ of SAT problems are SAT-*equivalent* (and we write $M' \sim_{SAT} M''$) if it holds that

$$(\forall c) \lim_{N\to\infty} |s(M'(N,[cN])) - s(M''(N,[cN]))| = 0$$

($c$ ranges over the set of positive reals numbers).

It is easy to prove the following two statements:

**Theorem 4.1.** The relation $\sim_{SAT}$ is an equivalence relation.

**Theorem 4.2.** If $M'$ and $M''$ are two types of corpora of SAT problems such that $M' \sim_{SAT} M''$ and if the value $c'$ is the crossover point for a corpora of the type $M'$ then it is also the crossover point for a corpora of the type $M''$.

Thus, all types of corpora in one class of SAT-equivalence relation lead to the same crossover point. The following theorem gives one sufficient condition for two types of corpora of SAT problems to be SAT-equivalent.

**Theorem 4.3.** If $M'$ and $M''$ are two types of corpora of SAT problems such that $M'(N,L) \subseteq M''(N,L)$ and such that for every positive real number $c$ it holds that $\lim_{N\to\infty} \frac{|M'(N,[cN])|}{|M''(N,[cN])|} = 1$, then $M' \sim_{SAT} M''$.

**Proof.** Let us denote by $M(N,L)$ the set $M''(N,L) \setminus M'(N,L)$. It holds that:

$|M''(N,L)| = |M'(N,L)| + |M(N,L)|$, and for the number of satisfiable formulae in $M''(N,L)$ (there are $s(M''(N,L))|M''(N,L)|$ of them) it holds that:

$$s(M'(N,L))|M'(N,L)| \leq s(M''(N,L))|M''(N,L)| =$$

$$= s(M'(N,L))|M'(N,L)| + s(M(N,L))(|M''(N,L)| - |M'(N,L)|)$$

and

$$s(M'(N,L))\frac{|M'(N,L)|}{|M''(N,L)|} \leq s(M''(N,L)) =$$

$$= s(M'(N,L))\frac{|M'(N,L)|}{|M''(N,L)|} + s(M(N,L))\left(1 - \frac{|M'(N,L)|}{|M''(N,L)|}\right) .$$

Let $c$ and $\epsilon$ be arbitrary positive real numbers. From

$$\lim_{N\to\infty} |M'(N,[cN])|/|M''(N,[cN])| = 1 ,$$

since $|M'(N,L)| \leq |M''(N,L)|$, it follows that there is a value $N_0$ such that

$$N > N_0 \Rightarrow 1 - \epsilon \leq |M'(N,[cN])|/|M''(N,[cN])| \leq 1 .$$

For $N > N_0$, since $s(M'(N,[cN])) \leq 1$ and $s(M(N,[cN])) \leq 1$, we have

$$s(M'(N,[cN])) - \epsilon \leq s(M'(N,[cN]))(1-\epsilon) \leq$$

$$\leq s(M'(N,[cN]))\frac{|M'(N,[cN])|}{|M''(N,[cN])|} \leq s(M''(N,[cN])) =$$

$$= s(M'(N,[cN]))\frac{|M'(N,[cN])|}{|M''(N,[cN])|} + s(M(N,[cN]))\left(1 - \frac{|M'(N,[cN])|}{|M''(N,[cN])|}\right) \leq$$

$$\leq s(M'(N,[cN])) + s(M(N,[cN]))\epsilon \leq s(M'(N,[cN])) + \epsilon$$

i.e.

$$s(M'(N,[cN])) - \epsilon \leq s(M''(N,[cN])) \leq s(M'(N,[cN])) + \epsilon$$

and

$$|s(M''(N,[cN])) - s(M'(N,[cN]))| \leq \epsilon .$$

Therefore,

$$(\forall c)(\forall \epsilon)(\exists N_0)(\ N > N_0 \ \Rightarrow\ |s(M''(N,[cN])) - s(M'(N,[cN]))| \leq \epsilon) ,$$

i.e.

$$(\forall c) \lim_{N\to\infty} |s(M''(N,[cN])) - s(M'(N,[cN]))| = 0 .$$

which yields $M' \sim_{SAT} M''$. □

## 5 FIXED CLAUSE LENGTH PROBLEMS AND SAT-EQUIVALENCE

In this section, we show that the given sufficient condition for two types of corpora of SAT problems (with fixed clause length) to be SAT-*equivalent* holds in some important cases. We also discuss some other cases in which this condition does not hold.

### 5.1 $M_1 \sim_{SAT} M_{1,2}$ and $M_{1,3} \sim_{SAT} M_{1,2,3}$

In all experiments that followed [15] (which is the dominant model) it is not checked if generated formulae do indeed have all clauses different — actually, some of the formulae don't and therefore they don't belong to the set $M_{1,2}$. Usually, just some informal argument is given saying that there are not too many such formulae and that this assumption probably will not have an impact on the location of the crossover point. As we are aware, the formal proof for this has not been given as yet. We prove that this assumption is valid, i.e., we show that it holds that $M_1 \sim_{SAT} M_{1,2}$.

**Lemma 5.1.** $M_1 \sim_{SAT} M_{1,2}$ .

**Proof.** Let us denote by $d$ the total number of different clauses in $M_1$ and $M_{1,2}$ ($d = (2N)^k$; $k$ is a fixed natural number and $k > 2$). In the corpus $M_1(N, L)$ there is a finite number of $k-$SAT formulae: $|M_1(N, L)| = d^L$. Some of them contain only different clauses (these formulae make corpus $M_{1,2}(N, L)$) and some of them contain multiple occurrences of some clauses (thus $M_{1,2}(N, L) \subset M_1(N, L)$). There are $d(d-1)(d-2)\cdots(d-L+1)$ formulae in $M_{1,2}(N, L)$ (it has to be $L < d$). For $L \geq 2$ it holds:

$$1 > \frac{|M_{1,2}(N,L)|}{|M_1(N,L)|} = \frac{d(d-1)(d-2)\cdots(d-L+1)}{d^L} =$$

$$= \left(1-\frac{1}{d}\right)\left(1-\frac{2}{d}\right)\cdots\left(1-\frac{L-1}{d}\right) \geq$$

$$\geq \left(1-\frac{L-1}{d}\right)^{L-1} \geq 1-\frac{(L-1)^2}{d} \geq 1-\frac{L^2}{d}$$

Let $c$ and $\epsilon$ be arbitrary positive real numbers.

$$1 > \frac{|M_{1,2}(N,[cN])|}{|M_1(N,[cN])|} \geq 1-\frac{[cN]^2}{d} = 1-\frac{[cN]^2}{(2N)^k} > 1-\frac{([c+1]N)^2}{(2N)^k} > 1-\frac{[c+1]^2}{N^{k-2}} .$$

Thus, for $N_0 = \left[\frac{[c+1]^2}{\epsilon}\right] + 1$ and for $N > N_0$ it holds that

$$1 > \frac{|M_{1,2}(N,[cN])|}{|M_1(N,[cN])|} \geq 1-\epsilon .$$

Therefore,

$$(\forall c) \lim_{N\to\infty} \frac{|M_{1,2}(N,[cN])|}{|M_1(N,[cN])|} = 1 ,$$

and by Theorem 4.3 it follows $M_1 \sim_{SAT} M_{1,2}$. □

Therefore, in the random $k$-SAT model, formulae that contain less than $L$ different clauses don't make an impact on the location of the crossover point.

Now we prove that the corpora $M_{1,3}$ and $M_{1,2,3}$ are SAT–equivalent, i.e., we prove that in the corpus of SAT formulae with all $N$ variables the restriction on formulae having all clauses different does not change the location of the crossover point.

**Lemma 5.2.** $M_{1,3} \sim_{SAT} M_{1,2,3}$ .

**Proof.** Let us denote by $M_{1,3}^{(m)}(N,L)$ the set of formulae from $M_{1,3}$ such that they have exactly $m$ different clauses. Note that $M_{1,3}^{(L)}(N,L) = M_{1,2,3}(N,L)$. Since all formulae from $M_{1,3}^{(m)}(N,L)$ contain all $N$ variables, it follows that $km \geq N$, i.e., all sets $M_{1,3}^{(m)}(N,L)$ are empty for $m < N/k$.

Let us associate to each formula $F$ from $M_{1,3}^{(m)}(N,L)$ $(m < L)$ a class of formulae from $M_{1,3}^{(m+1)}(N,L)$ such that they differ from $F$ only in its last duplicate clause and let $F'$ be one of these formulae. The corresponding clause in a formula $F'$ from $M_{1,3}^{(m+1)}(N,L)$ is not a duplicate clause (otherwise it would have less than $m+1$ different clauses), so the formula $F'$ occurs in at most $(m+1)\cdot m$ classes (a critical clause in the formula $F$ has to be from the set of $m$ different clauses). In one such class, there are $(2N)^k - m$ formulae from $M_{1,3}^{(m+1)}(N,L)$ (all these formulae indeed include all $N$ variables, because the formula $F$ includes all variables with or without the critical clause). Therefore, it holds $|M_{1,3}^{(m+1)}(N,L)| \geq \frac{(2N)^k-m}{(m+1)\cdot m}|M_{1,3}^{(m)}(N,L)| \geq \frac{(2N)^k-L}{L^2}|M_{1,3}^{(m)}(N,L)|$. It can be easily shown by mathematical induction that it holds $|M_{1,3}^{(L)}(N,L)| \geq \left(\frac{(2N)^k-L}{L^2}\right)^{L-m}|M_{1,3}^{(m)}(N,L)|$ for $m \leq L$. Thus,

$$\sum_{m=1}^{L}|M_{1,3}^{(m)}(N,L)| \leq \sum_{m=1}^{L}\frac{1}{(((2N)^k-L)/L^2)^{L-m}}|M_{1,3}^{(L)}(N,L)|$$

$$\leq |M_{1,3}^{(L)}(N,L)|\frac{1}{1-L^2/((2N)^k-L)}\ .$$

Union of the sets $|M_{1,3}^{(m)}(N,L)|$ $(L \geq m \geq 1)$ is the set $M_{1,3}(N,L)$ and thus we have

$$\frac{|M_{1,2,3}(N,L)|}{|M_{1,3}(N,L)|} = \frac{|M_{1,3}^{(L)}(N,L)|}{\sum_{m=1}^{L}|M_{1,3}^{(m)}(N,L)|} \geq$$

$$\geq \frac{|M_{1,3}^{(L)}(N,L)|}{\frac{1}{1-L^2/((2N)^k-L)}|M_{1,3}^{(L)}(N,L)|} = 1 - \frac{L^2}{(2N)^k-L}$$

Therefore, it holds

$$1 \geq \frac{|M_{1,2,3}(N,[cN])|}{|M_{1,3}(N,[cN])|} \geq 1 - \frac{[cN]^2}{(2N)^k - [cN]}$$

Since $k$ is greater than 2, for every $\epsilon > 0$ there is sufficiently large value $N_0$ such that it holds $|M_{1,2,3}(N,[cN])|/|M_{1,3}(N,[cN])| > 1 - \epsilon$. Thus,

$$\lim_{N\to\infty} |M_{1,2,3}(N,[cN])|/|M_{1,3}(N,[cN])| = 1$$

and by Theorem 4.3 it follows $M_{1,3} \sim_{SAT} M_{1,2,3}$. □

### 5.2 $M_1 \sim_{SAT} M_{1,3}$? and $M_1 \sim_{SAT} M_{1,2,3}$, $M_{1,4} \sim_{SAT} M_{1,2,3,4}$?

In the random $k$-SAT experiments (that followed [15]) it is not checked if the formulae generated do indeed have all $N$ variables. We question this assumption (this issue is discussed from the probability point of view in [11]).

Theorem 4.3 gives a sufficient condition for two types of SAT corpora to be SAT-equivalent. Let us check if the preconditions of Theorem 4.3 are fulfilled for the corpora $M_1$ and $M_{1,3}$.

There are $(2N-2)^k$ clauses that do not contain one variable from the set of $N$ given variables; thus, there are $((2N-2)^k)^L$ formulae that do not contain one variable. Having chosen one formula from $M_1(N,L)$, the probability that it does not contain one variable is $((2N-2)^k)^L/((2N)^k)^L = (N-1)^{kL}/N^{kL}$. Therefore, the probability that it does contain this variable is $1-(N-1)^{kL}/N^{kL}$. There are $N$ variables, so the probability that the chosen formula contain each of $N$ variables is

$$\left(1 - \frac{(N-1)^{kL}}{N^{kL}}\right)^N .$$

It holds $M_{1,3}(N,L) \subseteq M_1(N,L)$ and

$$\frac{|M_{1,3}(N,[cN])|}{|M_1(N,[cN])|} = \left(1 - \frac{(N-1)^{k[cN]}}{N^{k[cN]}}\right)^N = \left(1 - \left(1 - \frac{1}{N}\right)^{k[cN]}\right)^N .$$

Therefore, the preconditions of Theorem 4.3 are not fulfilled. In addition, the ratio $|M_{1,3}(N,[cN])|/|M_1(N,[cN])|$ is asymptotically the same as $(1-e^{-ck})^N$ and it does not converges to 1 (for $N \to \infty$), but to 0. However, for $k=3$, $c=4.25$ and for $N=10$, the ratio $|M_{1,3}(N,[cN])|/|M_1(N,[cN])|$ is approximately equal to 0.999986, for $N=100$ to 0.999728, for $N=1000$ to 0.99712. Thus, in random $k$-SAT model, formulae containing less than $L$ different clauses don't make an *significant* impact on the location of the crossover point at least for small values of $N$ (say, for $N<1000$); this means that results in [15] can be considered as correct. Additionally, most of formulae in $M_1(N,[cN])$ contain *almost* $N$ variables, so it is still possible that $M_1 \sim_{SAT} M_{1,3}$ and $M_{1,4} \sim_{SAT} M_{1,3,4}$, but some deeper knowledge on nature of SAT problems has to be used to prove it. In any case, it seems that this usual assumption has to be reconsidered.

We have proved $M_{1,3} \sim_{SAT} M_{1,2,3}$, (and $M_1 \sim_{SAT} M_{1,2}$). We have also shown that it can be considered that it holds $M_1 \sim_{SAT} M_{1,3}$ (or, more precisely, it can be considered that satisfiability functions for $M_1$ and $M_{1,3}$ are the same at least for small values for $N$ — say, for $N < 1000$). With this assumption, since SAT-equivalence is equivalence relation, it holds that $M_1 \sim_{SAT} M_{1,2,3}$.

By analogy with $M_1 \sim_{SAT} M_{1,2,3}$ it can be proved (with similar assumptions) that it holds $M_{1,4} \sim_{SAT} M_{1,2,3,4}$. This means that restrictions 2 and 3 make no difference in location of a crossover point. This also supports experiments made following [15].

## 6 GD-SAT MODEL

In this section we briefly discuss one new model for generating SAT problems with random clause length. We report on results showing that there is a phase transition for this model too.

Definitions of random SAT problems include information on the distribution of clause lengths. For instance, the constant probability model [7] has a limiting distribution on clause lengths determined by the Poisson distribution with parameter $2Np$ (adjusted for the omission of clauses of length 0 and 1); random mixed SAT has a discrete distribution on clause lengths. We consider SAT model based on a geometric distribution, and hence denote it by GD-SAT. In this model, generating of clauses over the set of $N$ variables, for the probability parameter $p$ ($0 < p \leq 1$), is specified by the stochastic context–free grammar given in Table 1 (a stochastic context–free grammar is a context–free grammar with a stochastic component which attaches a probability to each of the production rules and controls its use).

| # | Rule | Probability |
|---|---|---|
| 1. | $\langle clause \rangle := \langle literal \rangle \vee \langle literal \rangle$ | $p$ |
| 2. | $\langle clause \rangle := \langle clause \rangle \vee \langle literal \rangle$ | $1 - p$ |
| 3. | $\langle literal \rangle := \langle variable \rangle \mid \neg \langle variable \rangle$ | 0.5 |
| 4. | $\langle variable \rangle := v_1 \mid v_2 \mid \ldots \mid v_N$ | $1/N$ |

Table 1. Stochastic grammar for generating GD-SAT clauses

We point out that we do not perform a check whether some variable occurs more times in one clause, whether in some clause there is both a variable and its negation or whether there are multiple occurrences of some clause in a formula generated. These questions we have discussed in detail in the previous sections and we can obtain similar results for the GD-SAT model. Thus, we won't discuss these variants of the GD-SAT model.

By the given stochastic grammar, only clauses of length equal or greater than 2 can be generated (therefore there is no need for discarding any of generated clauses so the original distribution on clause lengths is kept intact, which is not the case in the constant probability model). Lengths of clauses in the GD-SAT model have a geometric distribution. The most probable clause length is 2 (with the probability $p$), while the expected clause length is $1 + 1/p$. For $p = 1$, the GD-SAT model is exactly 2-SAT model (and, hence, it belongs to the class P). For $p < 1$, GD-SAT problem is NP-complete. As $p$ decreases, GD-SAT problems smoothly interpolate between 2-SAT and NP-complete GD-SAT problems. This makes the GD-SAT model convenient for exploring a computational cost for directly related P and NP-complete problems (in a similar manner as in $2 + p$-SAT model).

We performed the following series of experiments for $p = 0.5$ and for $N = 25, 50, 75, 100, 200, 300, 400, 500$ we generated 10000 formulae in the GD-SAT model for values from $L/N$ varying from 0.1 to 10.0 by the step 0.1. For checking satisfiability we used Davis–Logemann–Loveland's procedure [5] with the following simple heuristic: when using a *split* rule, we apply it on a variable with most occurrences. We measured the percentage of satisfiable formulae and the number of branches made by the decision procedure. (The programs were written in C; experiments were ran on a PC Dual Pentium 256Mb running under Linux.)

The results we obtained show that there is a typical phase transition in the GD-SAT model (see Fig. 2). There is also a typical easy-hard-easy pattern concerning the computational cost for Davis–Logemann–Loveland's procedure as the ratio $L/N$ is increased. The most difficult GD-SAT problems are those in the crossover region (see Fig. 3). The expected clause length in GD-SAT for $p = 0.5$ is 3, but the average difficulty of the generated problems was less than for 3-SAT problems; this is due to the clauses of the length 2 which make GD-SAT problems (for $p = 0.5$) easier (the same behaviour is reported for the constant probability model [7]).

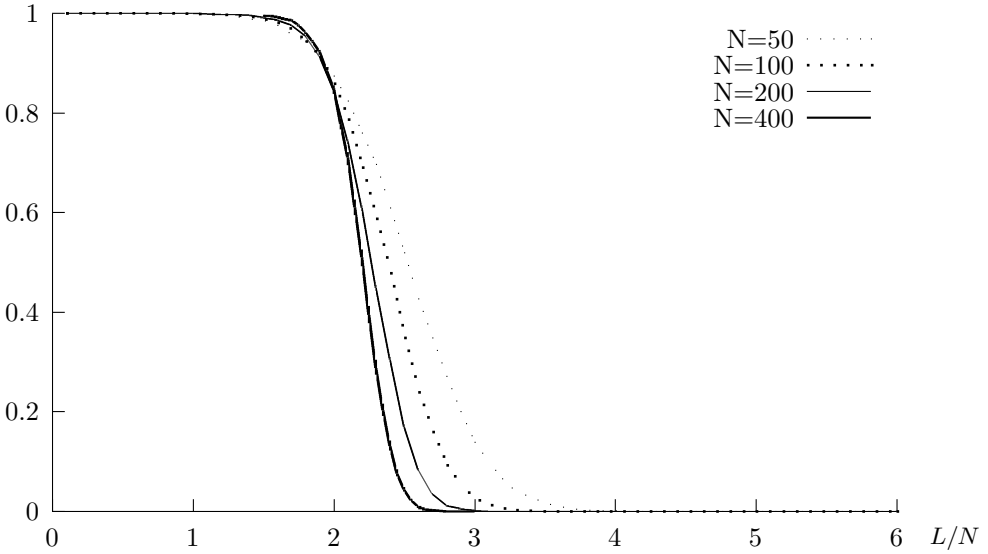


Fig. 2. Satisfiability function for the GD-SAT model as a function of ratio $L/N$ (for $N = 50, 100, 200, 400$).

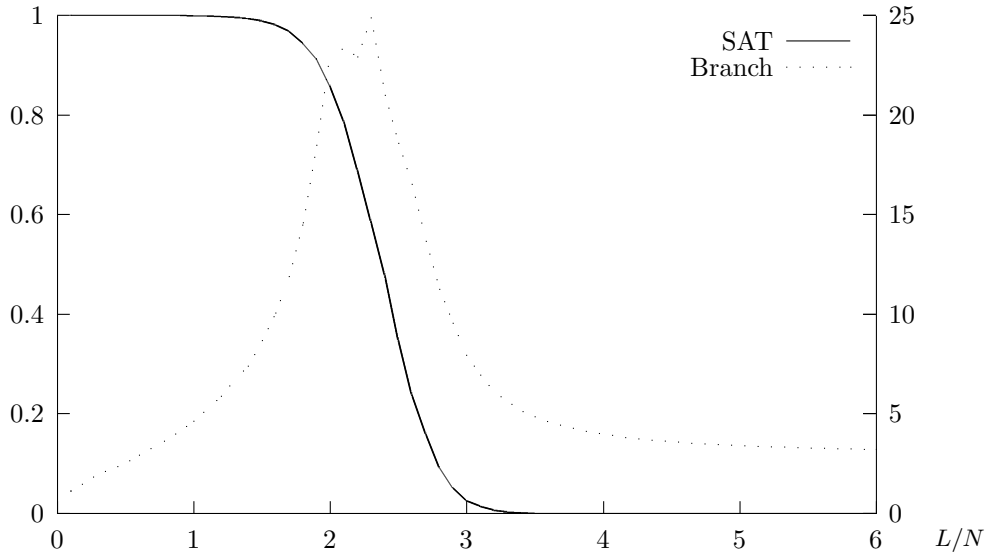


Fig. 3. Satisfiability function and a branching factor function for the GD-SAT model as functions of ratio $L/N$ (for $N = 100$).

| $N$ | 25 | 50 | 75 | 100 | 200 | 300 | 400 | 500 |
|---|---|---|---|---|---|---|---|---|
| critical value $L/N$ | 2.740 | 2.521 | 2.432 | 2.380 | 2.268 | 2.229 | 2.203 | 2.179 |

Table 2. Experimental estimates of the crossover point for the GD-SAT problem for $p = 0.5$

For each $N$ we located a point $L/N$ with 50% satisfiable formulae via linear interpolation between the two closest points with the satisfiability percentage determined experimentally (these values are given in Table 2). These values converge to the crossover point for GD-SAT for $p = 0.5$, but this converging is very slow (much slower than in 3-SAT model) and without giving clear estimate for the crossover point even after $N = 400$. Thus, instead, we tried to locate a point with approximately the same percentage of satisfiable formulae for all values of $N$. The satisfiability function is approximately constant for $L/N = 2.0$, and thus we estimate the crossover point at $L/N = 2.0$. In order to check this estimate, we additionally generated 1000 formulae in the GD-SAT model in the points $L/N = 1.9,\ 2.0,\ 2.1$ for $N = 1000$ and $N = 2000$ and measured the percentage of satisfiable formulae (for $N = 2000$ at both $L/N = 2.0$ and $L/N = 2.1$ there were respectively 16 and 7 formulae not decided within the $1hr$ time limit; we didn't count these formulae). Figure 4 shows the satisfiability functions as functions of $N$ for different values of $L/N$. For $L/N = 1.9$ (and hence for $L/N \leq 1.9$) the satisfiability function (as a function of $N$) is (slightly) increasing, for $L/N = 2.1$ (and hence for $L/N \geq 2.1$) it is decreasing, so the crossover point is between these two values. The satisfiability function at $L/N = 2.0$ is approximately 85% for all larger values $N$ and this supports the estimate that the crossover point is close to $L/N = 2.0$. (All these percentages of formulae satisfiable at $L/N = 1.9, 2.0, 2.1$ for different values of $N$ are given in Ta-

| $N$ | 25 | 50 | 75 | 100 | 200 | 300 | 400 | 500 | 1000 | 2000 |
|---|---|---|---|---|---|---|---|---|---|---|
| $L/N = 1.9$ | 0.917 | 0.911 | 0.914 | 0.912 | 0.912 | 0.924 | 0.923 | 0.924 | 0.930 | 0.951 |
| $L/N = 2.0$ | 0.889 | 0.874 | 0.871 | 0.858 | 0.846 | 0.850 | 0.849 | 0.839 | 0.854 | 0.846 |
| $L/N = 2.1$ | 0.866 | 0.824 | 0.814 | 0.787 | 0.743 | 0.727 | 0.711 | 0.683 | 0.620 | 0.513 |

Table 3. Experimental estimates of the satisfiability function at $L/N = 1.9, 2.0, 2.1$ for GD-SAT problem for $p = 0.5$

ble 3; note that the critical point with 50% satisfiable formulae is greater than 2.1 even for $N = 2000$).

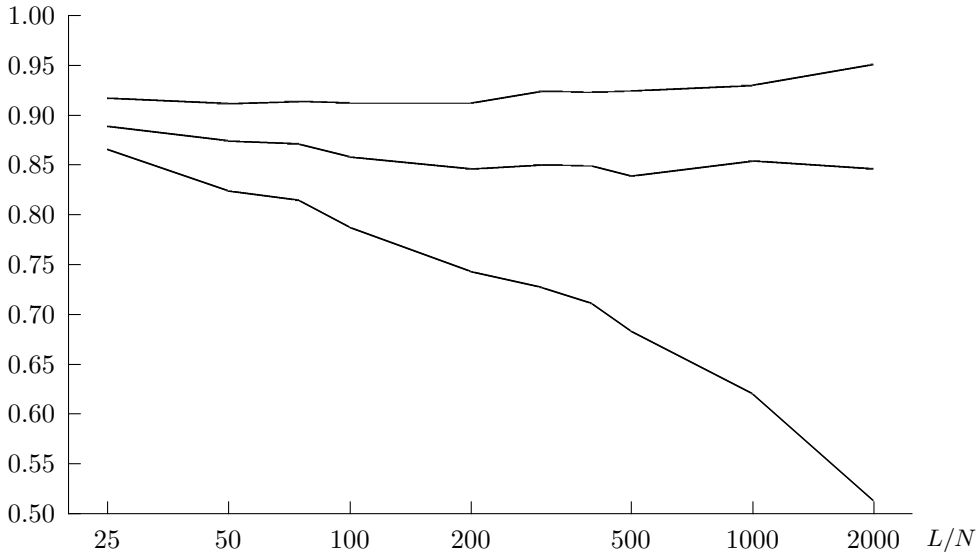


Fig. 4. Satisfiability function as a function of $N$ for different values of $L/N$ — 1.9 (top), 2.0 (middle), 2.1 (bottom).

As the parameter $p$ varies, the crossover point for GD-SAT changes its value and it is reasonable to expect that small changes of $p$ lead to small changes of corresponding crossover point. Consider a function $c$ such that $c(1/p)$ is equal to a crossover point for GD-SAT with a parameter $p$. It appears that this function $c$ (defined for values $\geq 1$) is continuous and it determines a curve which we will call a *crossover curve* for the GD-SAT model. This curve passes through the point $(1, 1)$ (because the crossover point for 2-SAT problem is 1) and through the point $(2, c(2.0))$. More extensive experiments [10] also suggest that $c(2.0)$ is between 1.9 and 2.0 and, moreover, they suggest that the crossover curve is linear (i.e., $c(1/p) = \alpha/p + \beta$). This elegant result further suggests that there is a unique parameter for all GD-SAT problems — instead of the parameter $L/N$, we can consider a parameter which gathers all satisfiability functions for GD-SAT problems into one such function. However, these results go beyond the scope of this paper and we won't discuss them further.

The results we obtained by the described experiments show that SAT problems with geometric distribution on clause lengths behave like other SAT problems known from the literature. Moreover, this distribution seems more elegant than one in CP model: in the GD-SAT model, for each $N$ there is geometric distribution on clause lengths, while in CP model there is Poisson distribution only as a limit and for each particular value $N$ this distribution is different. On the other hand, the GD-SAT model is also convenient for exploring a behavior of computational cost for solving P and NP-complete GD-SAT problems.

## 7 RELATION BETWEEN FIXED AND RANDOM CLAUSE LENGTH MODELS

The relation between fixed and random clause length models is not trivial. These two class of models can be related via different parameters: the most probable clause length, the expected clause length or via density (as introduced in [7]). However, neither of these gives the desired result, i.e., fixed and random clause length models with one of above parameters in common are still not SAT-equivalent. Indeed, in the GD-SAT model (with $p = 0.5$) the most probable clause length is 2, but the crossover point is close to 2 and is not 1 as for 2-SAT model. In the GD-SAT model (with $p = 0.5$) the expected clause length is 3, but the crossover point is not at 4.25 as in 3-SAT model. In [7] it was shown that there are models with the same density which do not have the same crossover point. Thus, the relation between fixed and random clause length models must be more involved. In [7] there is an interesting conjecture on this relation: let $c_k$ be the crossover point for $k$-SAT model (for $k = 2, 3, 4, \ldots$) and let $c_\phi$ be a crossover point for some random clause length model with the distribution $\phi$ on clause lengths, thus it holds:

$$\frac{1}{c_\phi} = \frac{\phi(2)}{c_2} + \frac{\phi(3)}{c_3} + \frac{\phi(4)}{c_4} + \ldots$$

The above conjecture gives good estimates for random mixed SAT model and for constant probability model [7]. We used the given conjecture for the GD-SAT model with $p = 0.5$ having geometric distribution $\phi(2) = 0.5$, $\phi(3) = 0.25$, $\phi(4) = 0.125$, .... We took the well known estimates for the crossover points for $k$-SAT ($k = 2, 3, 4$) problems ($c_2 = 1$, $c_3 = 4.25$, $c_4 = 9.76$) and we used used Kirkpatrick/Selman's approximations for crossover points for $k$-SAT ($k > 4$): $-1/\log_2(1 - \frac{1}{2^k})$ [13]. We computed the given sum in iterations and the values for the first eight iteration are given in Table 4. After thirty iterations the first three decimal places remained the same as in the sixth iteration and it suggests, by Gent/Walsh conjecture, that the crossover point for GD-SAT problem for $p = 0.5$ is $\approx 1.738$. However, the experimental results suggest that the crossover point is close to 2.0. This shows that, even the above conjecture is valid, it is not always of practical use since the Kirkpatrick/Selman's approximations are good only for the large values of $k$.

| number of summands | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| estimated crossover point | 2.000 | 1.789 | 1.749 | 1.741 | 1.739 | 1.738 | 1.738 | 1.738 |

Table 4. Estimates of crossover point for GD-SAT problem for $p = 0.5$ based on Gent/Walsh conjecture (using the approximations for crossover points for $k$-SAT based on Kirkapatrick/Selman's estimates).

## 8 FUTURE WORK

We have shown that it holds (or at least it can be considered that it holds) $M_1 \sim_{SAT} M_{1,2,3}$ and $M_{1,4} \sim_{SAT} M_{1,2,3,4}$. However, we haven't discussed if it holds $M_1 \sim_{SAT} M_{1,4}$ (as it can't be established by Theorem 4.3). Experiments made following [15] use only clauses with no multiple occurrences of one variable (e.g., they meet restriction 4). In future work we will try to investigate whether it holds $M_1 \sim_{SAT} M_{1,4}$.

In future work, we will try to define classes of SAT problem sets with more relaxed conditions: we are planning to explore relations between satisfiability functions for types of corpora that are not SAT–equivalent. These attempts rely on the hypothesis that there are some more involved parameter (different than $L/N$) made links between different types of SAT corpora. Some of these attempts will be based on the GD-SAT model which we also further investigate in much more details. We will also investigate behavior of the computational cost for directly related P and NP-complete GD-SAT problems.

## 9 CONCLUSIONS

In this paper we have introduced the notion of SAT–equivalence relation which links different types of SAT problem sets that have same crossover points. We have given one sufficient condition for two types of SAT corpora to be SAT–equivalent. By using this condition we proved that some restrictions often used in SAT experiments (such as restriction on $(N, L)$–SAT formulae that contain $L$ different clauses) don't make an impact on location of a crossover point. These results support many of experiments made so far (which used these conjectures as assumptions).

We have briefly discussed one new random clause length SAT model — a model with geometry distribution on clause lengths (denoted GD-SAT). We performed experiments that showed the typical phase transition behavior in the GD-SAT model. We have also discussed the relationship between the fixed and the random clause length SAT models.

We are planning to further investigate relations between different types of SAT-corpora and between crossover points of corpora that are not SAT–equivalent. We are also planning to further investigate the GD-SAT model and the behavior of the computational cost for P and NP GD-SAT problems.

## REFERENCES


[1] D. Achlioptas, L. M. Kirousis, E. Kranakis, and D. Krizanc. Rigorous results for random $(2 + p)$-SAT. In Proceedings of RALCOM' 97, 1997.

[2] P. Cheeseman, B. Kanefsky, and W. M. Taylor. Where the really hard problems are. In Proceedings of the 12th International Joint Conference on Artificial Intelligence, 1991.

[3] S. A. Cook. The complexity of theorem proving procedures. In *Proceedings of the 3rd Annual ACM Symposium on the Theory of Computation*, pages 151–158, 1971.

[4] M. James Crawford and D. Larry Auton. Experimental results on the crossover point in random 3-sat. Artificial Intelligence, 81:31–57, 1996.

[5] M. Davis, G. Logemann, and D. Loveland. A machine program for theorem-proving. *Communications of the Association for Computing Machinery*, 5:394–397, 1962.

[6] E. Friedgut. Sharp threshold for graph properties and the $k$-sat problem. *Journal of the American Mathematical Society*, 12:1017–1054, 1999.

[7] Ian P. Gent and Toby Walsh. The SAT phase transition. In *Proceedings of ECAI-94*, pages 105–109, 1994.

[8] A. Goerdt. A treshold for unsatisfiability. In *Proceedings of the 17th International Symposium on Mathematical Foundations of Computer Science*, 1992.

[9] J.N. Hooker and C. Fedjki. Branch-and-cut soultion of inference problems in propositional logic. *Ann. Math. Artif. Intell.*, 1:123–139, 1990.

[10] Predrag Janičić. gd-sat model and crossover line. Journal of Experimental and Theoretical Artificial Intelligence, 13(3):181–198, 2001.

[11] Frank Jeremy. The coupon collector's complaint — a cautionary note on random problem instance generation. In Proceedings of IJCAI-97 Workshop: Empirical AI, 1997.

[12] A. Kamath, R. Motwani, K. Palem, and P. Spirakis. Tail bounds for occupancy and the satisfiability treshhold conjecture. In *Proceedings 35th Symposium on Foundation of Computer Science*, pages 592–603, 1994.

[13] S. Kirkpatrick and B. Selman. Critical behaviour in the satisfiability of random boolean expressions. *Science*, 264:1297–1301, 1994.

[14] T. Larrabee and Y. Tsuji. Evidence for a satisfiability threshold for random 3cnf formulas. Technical Report UUCSC-CRL-92-42, University of California, Santa Cruz, 1992.

[15] G. David Mitchell, Bart Selman, and J. Hector Levesque. Hard and easy distributions of sat problems. In *Proceedings AAAI-92*, pages 459–465, San Jose, CA, 1992. AAAI Press/The MIT Press.

[16] R. Monasson, R. Zecchina, S. Kirkpatrick, B. Selman, and L. Troyansky. Phase transition and search cost in the $2 + p$-sat problem. In Proceedings of the Fourth Workshop on Physics and Computation, pages 229–232. Boston University, 1996.