

# It's all about the support: a new perspective for the satisfiability problem

Dan Vilenchik

School of Computer Science, Sackler Faculty of Exact Sciences,  
Tel-Aviv University, Tel-Aviv, Israel.  
email: vilenchi@post.tau.ac.il.

**Abstract.** In this paper we study a new approach to the satisfiability problem, which we call the *Support Paradigm*. Given a CNF formula  $F$  and an assignment  $\psi$  to its variables, we say that a literal  $x$  *supports* a clause  $C$  in  $F$  w.r.t.  $\psi$  if  $x$  is the *only* literal that evaluates to true in  $C$ . Our focus in this work will be simulated-annealing SAT heuristics – that is, a heuristic which starts at some assignment to the variables, and iteratively, using some predefined greedy rule, tries to minimize the number of unsatisfied clauses until a satisfying assignment is reached. We say that such a heuristic is part of the Support Paradigm if the greedy rule uses the support as its main criterion. We present a new algorithm in the Support Paradigm and rigorously prove its effectiveness on a certain distribution over satisfiable 3CNF formulas, known as the planted distribution. One motivation for this work is recent experimental results showing that some simple variants of the RWalkSAT algorithm, which base their greedy rules on the notion of support, remain effective for random 3CNF formulas in the “hard” near-threshold regime, while for example RWalkSAT, which disregards the support, is already inefficient in a much earlier stage.

## 1 Introduction and Results

Given a computational problem it is desirable to have algorithms that produce optimal results, are efficient (polynomial time), and work on every input instance. For many combinatorial problems, amongst which is SAT, this goal is too ambitious, as shown by the theory of NP-completeness. Hence one should set goals that are more modest. In this work we shall be interested in the heuristical approach, which relaxes the universality requirement. Here we define a heuristic to be a polynomial time algorithm that produces optimal results on typical inputs. The notion of a typical input, however, is rather fuzzy. One possibility to define typical instances is the use of random models. One such popular model is the following, which we denote by  $\mathcal{P}_{n,m}$ : fix  $c, n > 0$  ( $c$  may depend on  $n$ ), choose  $m = cn$  clauses uniformly at random out of  $8\binom{n}{3}$  possible ones. Despite its simplicity, many essential properties of this model are yet to be understood. In particular, the hardness of deciding if a random formula is satisfiable, and finding a satisfying assignment for a random formula, are both major open problems [12, 23].

Remarkable phenomena occurring in the random model  $\mathcal{P}_{n,m}$  are **phase transitions**. With respect to the property of being satisfiable, such a phase transition takes place too [17]. More precisely, there exists a threshold  $d = d(k, n)$  such that a  $k$ -CNF formula with clause-variable ratio greater than  $d$  is not satisfiable *whp*<sup>1</sup>, while one with ratio smaller than  $d$  is. For  $k = 3$  the threshold is known to be at least 3.42 [19] and at most 4.506 [20].

One way of evaluating and comparing heuristics is by running them on a collection of input instances (“benchmarks”), and checking which heuristic usually gives better results.

---

<sup>1</sup> When writing *whp* we mean with probability tending to 1 as  $n$  goes to infinity.

Empirical results are sometimes informative, but we also seek more rigorous measures of evaluating heuristics. In this paper we study a new heuristical approach to the satisfiability problem, which we call the *Support Paradigm*.

**Definition 1.** (*support*) *Given a 3CNF formula  $F$  and some assignment  $\psi$  to its variables, we say that a literal  $x$  supports a clause  $C$  (in which it appears) w.r.t.  $\psi$  if  $x$  is the only literal that evaluates to true in  $C$  under  $\psi$ .*

Our focus in this work will be simulated-annealing SAT heuristics – that is, a heuristic which starts at some assignment to the variables, and iteratively, using some predefined greedy rule, tries to minimize the number of unsatisfied clauses until a satisfying assignment is reached (or failure due to exceeding some maximal number of allowed steps, if the heuristic is to remain polynomial). We say that such a heuristic is part of the Support Paradigm if the greedy rule relies on the support as its main criterion. In this work we present a new algorithm in the **Support Paradigm** and rigorously prove its effectiveness on a certain distribution over satisfiable 3CNF formulas, known as the planted distribution.

One motivation for this work is recent experimental results showing that some simple variants of the well known RWalkSAT algorithm [25], which base their greedy rules on the notion of support, remain effective for random 3CNF formulas in the “hard” near-threshold regime [27, 5]. Specifically, these algorithms are efficient in finding a satisfying assignments for random instances in  $\mathcal{P}_{n,m}$  with clause-variable ratio of about 4.21 (while the conjectured satisfiability threshold for 3SAT is roughly 4.26). In contrast, the “original” RWalkSAT heuristic, which is not part of the Support Paradigm, consumes super-polynomial time already for clause-variable ratio of 2.65 [26].

## 2 Our Contribution

Motivated by a search of a unifying rule that explains this phenomenon, we define the Support Paradigm. We present a new algorithm which is part of the Support Paradigm and rigorously show its effectiveness for the Planted 3SAT distribution with clause-variable ratio greater than some constant (an exact definition of the Planted 3SAT distribution is given in Section 2.1). Therefore in some exact sense we provide evidence to the usefulness of the notion of support and a possible explanation for the clear advantage that [27, 5] possesses over RWalkSAT.

The notion of support also has a constructive interpretation when considering things from the statistical-mechanics point of view. In this discipline, the combinatorial object 3CNF is a diluted 3-spin spin glass system. Every assignment to the variables corresponds to an energy level of the system, where the free energy of the system in a certain state is the number of clauses not satisfied by the given assignment. Thus, the question of whether the 3CNF is satisfiable or not is equivalent to the question whether the ground state energy of this diluted 3-spin spin glass system is zero. One of the main theoretical bases, at least from a physical point of view, underlying Survey Propagation [8] is the structure of the energy states for near-threshold random 3CNF formulas.

Having said that, one immediately notices that the notion of support is tightly connected to the notion of free energy. For example, flipping the assignment of the variable with the lowest (maybe zero) support corresponds to making a move which incurs the least increase in the free energy of the system; or, lowering the energy of the system (by flipping the assignment of a variable which appears in at least one unsatisfied clause) corresponds to increasing the number of clauses that belong to the support of some variable, and so forth.

Another exciting area where the notion of support plays a role is the following. Using partially non-rigorous analytical tools from statistical mechanics [24], the following structure of the solution space of  $k$ -CNF formulas with clause-variable ratio just below the satisfiability threshold was suggested. Typically such formulas have an exponential number of **clusters** of satisfying assignments. Any two assignments in distinct clusters disagree on a fair number of variables and any two assignments within one cluster coincide on many variables. Furthermore, each cluster has a linear number of **frozen** variables whose assignment coincide in **all** satisfying assignments within that cluster. This picture was recently proved rigorously for  $k$ -SAT with  $k \geq 8$  in [1]. This structure is believed by many to be responsible for the “hardness” of the near-threshold formulas.

In this work we prove that the notion of support plays a crucial role in explaining the existence of frozen variables, at least for the Planted 3SAT distribution with clause-variable ratio greater than some constant. For example, if a variable has zero-support w.r.t. some satisfying assignment, then it cannot be frozen – flip its assignment and the new assignment, which lies in the same cluster, remains satisfying. The other direction is more tricky (that is what happens when a variable has a large support w.r.t. some satisfying assignment) – and the argument is more involved.

## 2.1 The Planted Model

In this work we consider the regime of satisfiable 3CNF formulas with clause-variable ratio which is some sufficiently large constant above the satisfiability threshold. In this regime almost all formulas are not satisfiable, and therefore  $\mathcal{P}_{n,m}$  is not suitable for the study of satisfiability heuristics. We choose to consider the **Planted 3SAT distribution**, which we shall denote by  $\mathcal{P}_{n,p}^{\text{plant}}$ . A formula in the planted distribution is chosen by first fixing an assignment, and then including every one of  $7\binom{n}{3}$  clauses that are satisfied by it with probability  $p = p(n)$ . This of course guarantees that non-zero probability is assigned only to satisfiable 3CNF formulas.

Planted distributions provide a good model for many real world statistical problems. They correspond to situations where the constraints are correlated in such a way as to be consistent (or statistically correlated) with a pre-specified assignment of the variables  $x_1, \dots, x_n$ . Planted models are also of interest in computational complexity [12]

It is favored by many researchers in the context of SAT [16, 6, 21], and also for other optimization problems such as max clique, min bisection, and coloring [3, 4, 7, 18, 13] to mention just a few. Another nice feature of planted-solution distributions is the fact that they are efficiently sampleable.

Furthermore, as recent results in [10, 11] imply, our results can be reproved in the uniform setting as well (to be specific, the uniform distribution over satisfiable 3CNF formulas with  $m = cn$  clauses,  $c$  greater than some sufficiently large constant).

We now **formally** state our result. We state it for 3SAT though it generalizes to  $k$ -SAT for any fixed  $k$ .

**Theorem 1.** *Let  $F$  be random formula from  $\mathcal{P}_{n,p}^{\text{plant}}$ , with  $n^2p \geq C_0$ ,  $C_0$  a sufficiently large constant. Then whp the algorithm  $\text{SupportSAT}(F)$  finds a satisfying assignment of  $F$  using polynomial time.*

The algorithm  $\text{SupportSAT}(F)$ , which belongs to the Support Paradigm, is described in Figure 2 of Section 4. The proof of Theorem 1 also reveals an interesting connection between the notion of support and the notion of frozeness (as defined above). Details in Sections 5.1 and 8.

## 2.2 Paper’s Structure

We proceed with some more background and related work, Section 3. In Section 4 we present our algorithm and analyze its performance in Sections 5–7, which contain all the technical details. Concluding remarks are given in Section 8.

## 3 Related Work

The seminal work of Alon and Kahale [3] paved the road towards average case analysis of algorithms on sparse (constant average degree) random instances. [3] present an algorithm that *whp*  $k$ -colors planted  $k$ -colorable graphs (the distribution of graphs generated by partitioning the  $n$  vertices into  $k$  equally-sized color classes, and including every edge connecting two different color classes with probability  $p = p(n)$ ) with a sufficiently large constant expected degree. Building upon their techniques, Flaxman [16] presents an efficient algorithm for  $\mathcal{P}_{n,p}^{\text{plant}}$  (the same regime that we study). The algorithm in [16] also proceeds in steps, starting with a spectral step, which typically obtains a fair approximation of the planted assignment, and ending with an exhaustive search. In addition some other algorithms were analyzed when the input is sampled according to  $\mathcal{P}_{n,p}^{\text{plant}}$  [15, 14, 22].

In [9] a simulated annealing algorithm for satisfiability is studied. In fact, this algorithm is also part of the Support Paradigm, and in some parts our algorithm is inspired by [9]. The algorithm in [9] is *exponential*-time and is analyzed for the uniform distribution over satisfiable 3CNF formulas with a sufficiently large, yet constant, clause-variable ratio. The authors leave as an open question whether one can find a polynomial time algorithm that solves *whp* such instances. This question was recently answered, positively, in [11] (though the algorithm described in [11] is not a simulated-annealing one, and is not part of the Support Paradigm). The analysis in [11] implies that Theorem 1 can be restated for the uniform distribution as well (with clause-variable ratio greater than some constant). Therefore our result settles the open question in [9] to show that a similar, arguably simpler, simulated-annealing algorithm solves the uniform distribution in *polynomial* time.

Our main goal in this work is not to show how one can solve planted instances, as this problem has been already settled in many papers as we just pointed out. Our starting point in this work is an experimentally observed phenomenon – the remarkable success of simple algorithms that use the notion of support for near-threshold random 3CNF formulas, while many other algorithms fail in that regime. Then we ask whether these experimental results can be justified rigorously. In search of a unifying rule that explains this phenomenon we define the Support Paradigm and use the planted distribution as a case study for a rigorous analysis of an algorithm in that paradigm.

This is similar in spirit to the work in [2] where it is shown that  $\text{RWalkSat}$  is efficient for  $\mathcal{P}_{n,m}$  with  $m/n \leq 1.63$ , and is inefficient for  $\mathcal{P}_{n,p}^{\text{plant}}$  with  $n^2p$  greater than some constant; to the work in [15] where a certain version of the  $k$ -opt heuristic is shown to work on  $\mathcal{P}_{n,p}^{\text{plant}}$ ; and

the work in [14] where the behavior of the well known Warning Propagation message passing algorithm was analyzed for  $\mathcal{P}_{n,p}^{\text{plant}}$  (in both the latter,  $\mathcal{P}_{n,p}^{\text{plant}}$  is in the same regime that we consider).

## 4 The Algorithm

We start with a high-level description of the algorithm. Given a formula  $F$  and an assignment  $\psi$  to its variables, we say that a variable  $x$  is “suspicious” if it supports very few clauses w.r.t.  $\psi$  (we soon quantify “very few” in an exact sense). Our algorithm is composed of two parts. The main part is a simple simulated annealing procedure in which we iteratively flip the assignment of suspicious variables. From a physical point of view, this part can be viewed as a fast cool-down process. When reaching low temperature, a large portion of the formula is already satisfied. If the remaining part of the formula is “simple”, then one can find a satisfying assignment for it using some simple of-the-shelf heuristic.

The fast cool-down process is done mainly via a procedure that we call a **Directed Walk** (inspired by the work of [9]), and then another procedure with a more refined criterion. This corresponds to Steps 1 and 2 in the description of SupportSAT below (Figure 2). Step 2 typically ends up with an assignment which is very close to a satisfying one. Step 3 completes the job using a simple exhaustive search. As typically the unsatisfied part left at the end of Step 2 is “simple”, the exhaustive search takes polynomial time (by typically we mean *whp* over formulas from the planted distribution).

*Remark 1.* The reader may wonder at this point if one can do without Step 3 of SupportSAT. Namely, can one push the greedy part of the algorithm (Steps 1 and 2) to typically find a satisfying assignment? The answer is probably no, at least not in our planted setting. To see this observe that every variable is expected to appear in  $7\binom{n}{2}p = \Theta(n^2p)$  clauses, which may be constant in our case. Further, the number of clauses in which a variables appears is binomially distributed. Thus, with constant probability some variables never appear, or appear very scarcely (say once or twice). Therefore in some sense they don't show enough structure to allow a greedy procedure of the sort we presented to set them “correctly”. On the other hand, these variables induce a “simple” formula, for which exhaustive search is efficient for example.

We now introduce the subprocedure Directed Walk (the name will become clear soon), which is possibly of its own interest. The input to Directed Walk is a 3CNF formula  $F$  and a number  $\varepsilon \in [0, 1]$ .

### Directed Walk( $F, \varepsilon$ )

1.  $\psi_1 \leftarrow$  an arbitrary assignment to the variables.
2. **for**  $i = 2$  to  $3/\varepsilon$ 
  - $\psi_i \leftarrow \psi_{i-1}$  with the assignment of the  $\varepsilon n$  variables with the lowest support in  $F$  w.r.t.  $\psi_{i-1}$  – flipped.
3. **return**  $\psi_{3/\varepsilon}$ .

Fig. 1. Directed Walk

The name “Directed Walk” comes from the fact that as opposed to RWalkSat (‘R’ stands for Random), where the choice of which variables to flip blends randomness, here one employs a decisive deterministic rule, thus the walk is in some sense directed. Directed Walk can be used with other measurements – e.g. flip the assignment of the  $\varepsilon n$  variables whose flipping will gain the maximal number of satisfied clauses, and so forth. In fact, using the last rule with  $\varepsilon = 1/n$  is exactly the algorithm in [21]. Actually, one can generalize Directed Walk to receive the “directing rule” as an argument, and then have a general template (and analysis) for such algorithms. More details in Section 6.

We are now ready to present our main algorithm.

**Notations.** We let  $Support(x, \psi)$  be the support of a variable  $x$  w.r.t. an assignment  $\psi$ , where the formula is clear from the context. We let  $\psi^{(x)}$  be the assignment  $\psi$  with the assignment of  $x$  flipped. For a formula  $F$  and a subset  $U$  of the variables we denote by  $F[U]$  the subformula containing all clauses with some variable in  $U$ . By *partial assignment* we mean an assignment where some variables may take the value UNASSIGNED. For a partial assignment  $\psi$ ,  $Support(x, \psi)$  counts only clauses where all variables are assigned.

**SupportSAT( $F$ )**  
*Step 1: Directed Walk*  
 1.  $\psi_1 \leftarrow \text{Directed Walk}(F, 10^{-4})$ .  
*Step 2: refining the assignment*  
 2. **for**  $i = 1$  **to**  $\log n$   
 3.   **for all**  $x \in V$   
 4.     **if**  $Support(x, \psi_i) \leq n^2 p / 10$  **then**  $\psi_{i+1} \leftarrow \psi_i^{(x)}$   
 5.   **end for.**  
 6. **end for.**  
 7. **let**  $\tau$  be the final assignment.  
*Step 3: the exhaustive search*  
 8. **set**  $\tau_1 = \tau$ ,  $j = 1$ .  
 9. **while**  $\exists x$  s.t.  $Support(x, \tau_j) \leq n^2 p / 10$   
 10.   **set**  $\tau_{j+1} \leftarrow \tau_j$  **with**  $x$  **unassigned.**  
 11.    $j \leftarrow j + 1$ .  
 12. **end while.**  
 13. **Let**  $\xi$  be the final partial assignment.  
 14. **let**  $U$  be the set of unassigned variables in  $\xi$ .  
 15. exhaustively search  $F[U]$ , separately in every connected component.

Fig. 2. SupportSAT

Lines 8-12 in SupportSAT are intended to isolate the part of the formula which is not satisfied by the first two steps (maybe with some additional satisfied clauses) and is to be exhaustively searched. To do so we chose an unassignment procedure, rather than just take all clauses which  $\tau$  fails to satisfy. For a variable to survive the unassignment step it has to show large support with other assigned variables. Throughout we proclaim “large support” to be synonymous with “high-certainty” for being correctly assigned – and indeed, the variables that survive the unassignment step are typically correctly assigned (namely, their assignment agrees with the planted one), and in fact only few variables get unassigned.

## 5 Properties of a Random Instance from $\mathcal{P}_{n,p}^{\text{plant}}$

In this section we analyze the structure of a typical formula in  $\mathcal{P}_{n,p}^{\text{plant}}$ . These properties will come handy when analyzing the algorithm **SupportSAT** in Section 7. One consequence of the discussion in this section is showing how the notion of support plays a crucial role in the existence of frozen variables in  $\mathcal{P}_{n,p}^{\text{plant}}$  (see Section 5.1).

All the propositions in this section appear also in [3, 16] for example (maybe stated a bit differently). We therefore only state the propositions and give an outline of the proof (which can be easily reconstructed into a full proof).

The following proposition excludes the case where  $F$  contains a small subformula in which many variables appear many times (much more than expected).

**Proposition 1.** *Let  $F$  be distributed according to  $\mathcal{P}_{n,p}^{\text{sat}}$ , with  $n^2p \geq C_0$ ,  $C_0$  some sufficiently large constant. Then whp there exists no subset of variables  $U$  s.t.*

- $|U| \leq n/10^4$ ,
- There are  $n^2p|U|/500$  clauses in  $F$  that contain two variables from  $U$ .

**Proof.** The proof uses the union bound technique. For a fixed set  $U$  of  $k$  variables, the number of clauses containing two variables from  $U$  is

$$\binom{k}{2}(n-2)2^3 \leq 4k^2n.$$

Each of these clauses is included independently w.p.  $p$ . Thus, the probability that  $n^2pk/500$  of them are included is at most

$$\binom{4k^2n}{n^2pk/500} p^{n^2pk/500} \leq \left( \frac{2000 \cdot e \cdot k}{n} \right)^{n^2pk/500}.$$

Summing over all possible sets  $U$  of size up to  $n/10^4$ , one obtains that the probability for such a “bad” set  $U$  in  $F$  is at most

$$\sum_{k=1}^{n/10^4} \binom{n}{k} \left( \frac{2000 \cdot e \cdot k}{n} \right)^{n^2pk/500} = \dots = o(1).$$

The “...” can easily be filled with standard calculations. ■

### 5.1 The Core Variables

We describe a subset of the variables, referred to as the *core variables*, which plays a crucial role in the analysis of the algorithm and in the understanding of  $\mathcal{P}_{n,p}^{\text{plant}}$ . Recall that a variable is said to be frozen in  $F$  if in every satisfying assignment it takes the same value. The notion of core captures this phenomenon, while the main property defining the core has to do with the notion of support. In addition, a core typically contains all but a small (though constant) fraction of the variables.

**Definition 2.** (core) A set of variables  $\mathcal{H}$  is called a  $t$ -core of  $F$  w.r.t. to a satisfying assignment  $\psi$ , if the following two properties hold:

- Every variable  $v \in \mathcal{H}$  supports least  $t$  clauses in  $F$  w.r.t.  $\psi$ , where all variables in these clauses belong to  $\mathcal{H}$ .
- $v$  appears in at most  $t/10$  clauses in  $F$  where not all variables belong to  $\mathcal{H}$ .

We proceed by asserting some relevant facts that such a core typically possesses.

**Proposition 2.** Let  $F$  be distributed according to  $\mathcal{P}_{n,p}^{\text{plant}}$ , with  $n^2p \geq C_0$ ,  $C_0$  some sufficiently large constant. Then whp there exists a  $t$ -core  $\mathcal{H}$  with  $t = n^2p/4$ , w.r.t. the planted assignment. Furthermore, whp  $|\mathcal{H}| \geq (1 - e^{-\Theta(n^2p)})n$ .

Note that  $t$  is chosen to be half of the expected support of a variable.

**Proof.** Consider the following procedure, which we prove – defines a core. We use  $\varphi$  to denote the planted assignment.

Let  $B$  be the set of variables whose support in  $F$  w.r.t.  $\varphi$  is at most  $n^2p/3$ .

1. set  $H_0 = V \setminus B$ .
2. **while** there exists a variable  $a_i \in H_i$  for which one of the following holds:
  - it supports less than  $n^2p/4$  clauses where all variables belong to  $H_i$ .
  - it appears in more than  $n^2p/40$  clauses where not all variables belong to  $H_i$ .**do** define  $H_{i+1} = H_i \setminus \{a_i\}$ .
3. let  $a_m$  be the last variable removed in step 2. Define  $\mathcal{H} = H_{m+1}$ .

The set  $\mathcal{H}$  which this procedure outputs is a  $t$ -core (according to Definition 2) by its construction, with  $t = n^2p/4$ .

Let  $\bar{\mathcal{H}} = V \setminus \mathcal{H}$ , and set  $\delta = e^{-\Theta(n^2p)}$ . Partition the variables in  $\bar{\mathcal{H}}$  into variables that belong to  $B$ , and variables that were removed in the iterative step,  $\bar{H}^{it} = H_0 \setminus \mathcal{H}$ . If  $|\bar{\mathcal{H}}| \geq \delta n$ , then at least one of  $B$ ,  $\bar{H}^{it}$  has cardinality at least  $\delta n/2$ . Consequently,

$$Pr[|\bar{\mathcal{H}}| \geq \delta n] \leq \underbrace{Pr[|B| \geq \delta n/2]}_{(a)} + \underbrace{Pr[|\bar{H}^{it}| \geq \delta n/2 \mid |B| \leq \delta n/2]}_{(b)}.$$

To bound (a), we use the following lemma, whose proof consists of standard probabilistic arguments; details omitted.

**Lemma 1.** Let  $F$  be random formula from  $\mathcal{P}_{n,p}^{\text{plant}}$ ,  $n^2p \geq C_0$ ,  $C_0$  a sufficiently large constant, and let  $F_{\text{SUPP}}$  be a random variable counting the number of variables in  $F$  whose support w.r.t.  $\varphi$  is less than  $n^2p/3$ . Then, whp  $F_{\text{SUPP}} \leq e^{-\Theta(n^2p)}n$ .

To bound (b), observe that every variable that is removed in iteration  $i$  of the iterative step (Step 2), supports at least  $(n^2p/3 - n^2p/4) = n^2p/12$  clauses in which at least another variable belongs to  $\{a_1, a_2, \dots, a_{i-1}\} \cup B$ , or appears in  $n^2p/40$  clauses each containing at least one of the latter variables. Consider iteration  $\delta n/2$ . Assuming  $|B| \leq \delta n/2$ , by the end



of this iteration there exists a set containing at most  $\delta n$  variables, and there are at least  $n^2 p / 40 \cdot \delta n / 2 \cdot 1/3 = n^2 p / 240 \cdot \delta n$  clauses containing at least two variables from it (we divide by 3 as every clause might have been counted 3 times). This however contradicts Proposition 1 as  $\delta = e^{-\Theta(n^2 p)} \ll 10^{-4}$ . ■

The next two propositions are given without a proof as their proofs are quite technical and can be found in complete in [16] and [11] respectively.

**Proposition 3.** *Let  $F$  be distributed according to  $\mathcal{P}_{n,p}^{\text{plant}}$ , with  $n^2 p \geq C_0$ ,  $C_0$  some sufficiently large constant. Let  $\mathcal{H}$  be a  $t$ -core of  $G$ , and let  $F[V \setminus \mathcal{H}]$  be the formula induced by the non-core variables. If  $|\mathcal{H}| \geq (1 - e^{-\Theta(n^2 p)})n$  then whp the graph induced by  $F[V \setminus \mathcal{H}]$  contains no connected component of size greater than  $\log n$ .*

“The graph induced by a CNF formula” means the graph whose vertices are the variables, and two variables share an edge if there exists some clause containing them both.

The following fact establishes the “frozenness” of the core variables. Its proof can be found in [11], we just note that the first property in Definition 2 (the support) plays a major role in the proof.

**Proposition 4.** *Let  $F$  be distributed according to  $\mathcal{P}_{n,p}^{\text{plant}}$ , with  $n^2 p \geq C_0$ ,  $C_0$  some sufficiently large constant. Let  $\mathcal{H}$  be the core promised in Proposition 2. Then whp  $\mathcal{H}$  is uniquely satisfiable*

## 6 Analysis of Directed Walk

In this section we analyze the typical behavior of Directed Walk for  $\mathcal{P}_{n,p}^{\text{plant}}$ ,  $n^2 p$  greater than some sufficiently large constant. Directed Walk, as defined in Figure 1, uses the measure of support to determine which variables flip their assignment in every round. Nevertheless, one can use other measures – such as the number of unsatisfied clauses in which a variable appears, the number of satisfied clauses gained by flipping the variable, and so on. Our analysis can be easily fit to other measures that satisfy some sufficient conditions, which are implicit in Lemma 2 (and stated explicitly in Remark 3). In fact, our analysis implies the main result in [21]. The following proposition summarizes the main quality of Directed Walk.

**Proposition 5.** *Let  $F$  be distributed according to  $\mathcal{P}_{n,p}^{\text{plant}}$ ,  $n^2 p \geq C_0$ ,  $C_0$  some sufficiently large constant. Then whp Directed Walk( $F, 10^{-4}$ ) enjoys the following property: after at most  $3 \cdot 10^4$  rounds, the output assignment differs from the planted one on at most  $n/10^4$  variables.*

*Remark 2.* The constant  $10^4$  is arbitrary. In fact, one can show that whp Directed Walk( $F, \varepsilon$ ) finds after  $3/\varepsilon$  rounds an assignment at distance at most  $\varepsilon n$  from the planted one, for  $\varepsilon$  as small as  $e^{-\Theta(n^2 p)}$ . This is (up to a constant in the exponent that does not depend on  $n, p$ ) exactly the approximation ratio of the Majority Vote [15, 6]. Therefore, if one considers  $\mathcal{P}_{n,p}^{\text{plant}}$  with  $n^2 p \geq C_0 \log n$ , then whp Directed Walk finds the planted assignment. Indeed, this is what [21] proves, though the directing measure is not the support.

Before proving the proposition we need some further observations.

**Definition 3.** (*misleading assignments*) Let  $F$  be a satisfiable CNF formula and  $\varphi$  a satisfying assignment of  $F$ . We call an assignment  $\psi$   $k$ -misleading w.r.t.  $\varphi$  if there exists a set of  $2k$  variables  $t_1, \dots, t_k, f_1, \dots, f_k$  s.t. for every  $i, j = 1, \dots, k$ :

- $\varphi(t_i) = \psi(t_i), \varphi(f_i) \neq \psi(f_i)$ ,
- $Support(t_i, \psi) \leq Support(f_j, \psi)$ .

**Definition 4.** ( $\varepsilon$ -directable) We say that  $F$  is  $\varepsilon$ -directable w.r.t. a satisfying assignment  $\varphi$  of  $F$  if there exists no  $k$ -misleading assignments w.r.t.  $\varphi$ ,  $k = \varepsilon n/3$ , at Hamming distance greater than  $\varepsilon n$  from  $\varphi$

**Proposition 6.** Let  $F$  be distributed according to  $\mathcal{P}_{n,p}^{\text{plant}}$ ,  $n^2 p \geq C_0$ ,  $C_0$  some sufficiently large constant, and let  $\varphi$  be its planted assignment. Then whp  $F$  is  $10^{-4}$ -directable w.r.t.  $\varphi$ .

The proof of Proposition 6 is given by the following lemma.

**Lemma 2.** Fix  $\varepsilon \in (0, 1)$  and let  $\varphi, \psi$  be two assignments at distance  $\geq \varepsilon n$  from each other. Let  $F$  be distributed according to  $\mathcal{P}_{n,p}^{\text{plant}}$ ,  $n^2 p \geq C_0$ ,  $C_0$  some sufficiently large constant, with  $\varphi$  its planted assignment. Then the probability that  $\psi$  is  $\varepsilon n/3$ -misleading w.r.t.  $\varphi$  is at most  $3^{-n}$ .

The union bound guarantees that whp no misleading  $\psi$  exists, as there are at most  $2^n$  possible ways to choose  $\psi$ . To prove Lemma 2 we need the following easy fact whose proof consists of standard probabilistic arguments.

**Lemma 3.** Let  $\beta \in (0, 1)$ . Let  $B_1 = Binom(p, \binom{N}{2}), B_2 = Binom(p, \binom{N}{2} - \binom{\beta N}{2})$ . Then  $Pr[B_1 \leq B_2] \leq e^{-g(\beta)pN^2}$ , where  $g: (0, 1) \rightarrow (0, \infty)$  is a monotonically increasing function.

**Proof.**(Lemma 2) Suppose that  $\psi$  is at distance  $\beta n$  from  $\varphi$ . Let  $t, f$  be two variables s.t.  $\psi(t) = \varphi(t), \psi(f) \neq \varphi(f)$ .  $t$  supports  $\binom{n}{2}$  clauses w.r.t.  $\psi$ , and every one of them could have been included in  $F$  (since  $\varphi$  satisfies all of them). On the other hand,  $f$  supports  $\binom{n}{2}$  clauses w.r.t.  $\psi$ , but clauses where both  $\varphi$  and  $\psi$  agree on the assignment of the other two variables, cannot be included in  $\varphi$ , as they are not satisfied by  $\varphi$  – and there are  $\binom{(1-\beta)n}{2}$  of them. Therefore we get:

$$Pr[Sup(t, \psi) \leq Sup(f, \psi)] \leq Pr[Bin(p, \binom{n}{2}) \leq Bin(p, \binom{n}{2} - \binom{(1-\beta)n}{2})] \leq e^{-g(1-\beta)n^2 p}. \quad (1)$$

Further observe that the sets of clauses that any two variables support w.r.t. to some assignment are always disjoint (since the supporting variable is unique, by definition). If  $\psi$  is  $k$ -misleading w.r.t.  $\varphi$ , then in particular there exist  $k$  pairs of variables  $(t_1, f_1), \dots, (t_k, f_k)$  s.t.  $Support(t_i, \psi) \leq Support(f_i, \psi)$ . The probability for this is at most

$$e^{-g(1-\beta)n^2 p \cdot \varepsilon n/3} \leq e^{-(g(1-\beta)C_0 \cdot \varepsilon/3)n} = \left( e^{-g(1-\beta)C_0 \cdot \varepsilon/3} \right)^n \leq 12^{-n}.$$

The last inequality is due to  $1 - \beta \leq 1 - \varepsilon$  and therefore  $g(1 - \beta) \leq g(1 - \varepsilon)$ , where  $\varepsilon$  is some fixed number, while  $C_0$  can be an arbitrarily large. Finally observe that there are at most  $2^n \cdot 2^n = 4^n$  ways to choose the sets of  $t_i$ 's and  $f_j$ 's. Now apply the union bound to receive the lemma. ■

*Remark 3.* In order for a measure function  $M$  to fit the proof of Proposition 5 it suffices for  $M$  to obey Equation (1) (maybe with some other function  $g'$ ), and also

$$\Pr[M(t) \leq M(f) | M(t_{i_1}) \leq M(f_{i_1}), \dots, M(t_{i_r}) \leq M(f_{i_r})] \leq \Pr[M(t_i) \leq M(f_i)],$$

for every  $r$ -subset of the variables ( $i, r \leq k$ ).

**Proof.** (Proposition 5) The proof we give here shares some ideas with the analysis in [9]. We prove that Proposition 5 holds with probability 1 for  $F$  s.t.  $F$  is  $10^{-4}$ -directable w.r.t.  $\varphi$ . Since this is the case *whp*, as asserted by Proposition 6, Proposition 5 follows. For two assignments  $\psi, \varphi$ , define  $T(\psi, \varphi)$  to be the set of variables on which  $\psi$  and  $\varphi$  agree, and  $F(\psi, \varphi)$  the set of variables on which they disagree. Let  $E_\varepsilon(\psi)$  be the set of  $\varepsilon n$  variables with lowest support w.r.t.  $\psi$ . Observe that if  $\varphi$  is some satisfying assignment of  $F$ , and  $\psi$  is the current assignment that Directed Walk( $F, \varepsilon$ ) considers, then the variables in  $T(\psi, \varphi) \cap E_\varepsilon(\psi)$  will be “wrongly” flipped. Our goal is then to show that  $|T(\psi, \varphi) \cap E_\varepsilon(\psi)|$  cannot be too large.

Set  $\varepsilon = 10^{-4}$  (as required by Proposition 5). Suppose at first that for every  $\psi$  at distance  $\geq \varepsilon n$  from  $\varphi$ ,  $|T(\psi, \varphi) \cap E_\varepsilon(\psi)| \leq \varepsilon n/3$ . If so, then  $|F(\psi, \varphi) \cap E_\varepsilon(\psi)| \geq 2\varepsilon n/3$ . Thus, in every iteration of Directed Walk the distance from  $\varphi$  is decreased by at least  $2\varepsilon n/3 - \varepsilon n/3 = \varepsilon n/3$ . The initial distance is at most  $n$ . Hence, after at most  $n/(\varepsilon n/3) = 3/\varepsilon = 3 \cdot 10^4$  rounds, an assignment  $\psi'$  at distance at most  $\varepsilon n$  from  $\varphi$  is reached.

It remains to prove that the above picture is indeed the case. To this end, consider a “bad” assignment  $\psi$  at distance  $> \varepsilon n$  from  $\varphi$  but for which  $|T(\psi, \varphi) \cap E_\varepsilon(\psi)| \geq \varepsilon n/3$ . This implies that  $|F(\psi, \varphi) \cap E_\varepsilon(\psi)| \leq 2\varepsilon n/3$ . Since the distance between  $\psi$  and  $\varphi$  is  $\geq \varepsilon n$ , it holds that  $|F(\psi, \varphi)| \geq \varepsilon n$ . The two last observations imply that  $|F(\psi, \varphi) \setminus E_\varepsilon(\psi)| \geq \varepsilon n/3$ .

Set  $k = \varepsilon n/3$ . Let  $f_1, f_2, \dots, f_k$  be variables in  $F(\psi, \varphi) \setminus E_\varepsilon(\psi)$ , and  $t_1, t_2, \dots, t_k$  be variables in  $T(\psi, \varphi) \cap E_\varepsilon(\psi)$ . For every  $t_i, f_j$ ,  $\text{Support}(t_i, \psi) \leq \text{Support}(f_j, \psi)$  (by the definition of  $E_\varepsilon(\psi)$  and the choice of the  $t_i$ 's and the  $f_j$ 's). However this means that  $\psi$  is  $k$ -misleading w.r.t.  $\varphi$ , and the Hamming distance between  $\psi$  and  $\varphi$  is greater than  $\varepsilon n$ . This however contradicts that fact that  $F$  is  $\varepsilon$ -directable w.r.t.  $\varphi$ . ■

## 7 Algorithm's Analysis – Proof of Theorem 1

We say that a formula  $F$  is *typical* if Propositions 1, 2, 3 and 6 hold for  $F$ . The discussion in Sections 5 and 6 guarantees that indeed *whp* a formula sampled according to  $\mathcal{P}_{n,p}^{\text{plant}}$ ,  $n^2 p$  greater than some sufficiently large constant, is typical. Thus, proving Theorem 1 reduces to proving that SupportSAT finds a satisfying assignment in polynomial time for typical formulas. Therefore, in all the propositions below we assume that  $F$  is typical. We let  $\mathcal{H}$  be the core promised in Proposition 2, and  $\varphi$  – the planted assignment of  $F$ .

**Proposition 7.** *Let  $\tau$  be the assignment defined in line 7 of SupportSAT. Then  $\tau$  agrees with  $\varphi$  on the assignment of all variables in  $\mathcal{H}$ .*

**Proof.** Let  $B_i$  be the set of core variables whose assignment in  $\psi_i$  disagrees with  $\varphi$  at the beginning of the  $i^{\text{th}}$  iteration of the main for-loop – line 2 in SupportSAT. It suffices to prove that  $|B_{i+1}| \leq |B_i|/2$  (if this is true, then after  $\log n$  iterations  $B_{\log n} = \emptyset$ ). Observe

that by Proposition 5,  $|B_0| \leq n/10^4$ . By contradiction, assume that not in very iteration  $|B_{i+1}| \leq |B_i|/2$ , and let  $j$  be the first iteration violating the inequality –  $|B_{j+1}| \geq |B_j|/2$ . Consider a variable  $x \in B_{j+1}$ . If also  $x \in B_j$ , this means that  $x$ 's assignment was not flipped in the  $j^{\text{th}}$  iteration, and therefore,  $x$  supports at least  $n^2p/10$  clauses w.r.t.  $\psi_j$ . By the second item in the definition of a core, at least  $n^2p/10 - n^2p/40 \geq n^2p/20$  of these clauses contain only core variables. Since the literal of  $x$  is true in all these clauses, but in fact should be false under  $\varphi$ , each such clause must contain another variable on which  $\varphi$  and  $\psi_j$  disagree, that is another variable from  $B_j$ . If  $x \notin B_j$ , this means that  $x$ 's assignment was flipped in the  $j^{\text{th}}$  iteration. This is because  $x$  supports less than  $n^2p/10$  clauses w.r.t.  $\psi_j$ . Since  $x$  supports at least  $n^2p/4$  clauses w.r.t.  $\varphi$ , it must be that in at least  $n^2p/4 - n^2p/10 \geq n^2p/8$  of them, the literal of some other core variable evaluates to true (rather than false, as it should be w.r.t.  $\varphi$ ). For conclusion, let  $U = B_j \cup B_{j+1}$ . Then there are at least  $n^2p/20 \cdot |B_{j+1}|$  clauses containing at least two variables from  $U$ . Now if  $|B_{j+1}| \geq |B_j|/2$ , then  $n^2p/20 \cdot |B_{j+1}| \geq n^2p/30 \cdot |U|$ , contradicting Proposition 1. ■

**Proposition 8.** *Let  $\xi$  be the partial assignment defined in line 13 of *SupportSAT*. Then all assigned variables in  $\xi$  are assigned according to  $\varphi$ , and all the variables in  $\mathcal{H}$  are assigned.*

**Proof.** The core variables are assigned according to  $\varphi$  when the unassignment begins (Proposition 7). Therefore, by the definition of core, every core variable supports at least  $n^2p/4$  clauses w.r.t.  $\varphi$ , and also w.r.t.  $\tau_1$  (the assignment at hand before the unassignment step begins). Therefore all core variables survive the first round of unassignment. By induction it follows that the core variables survive all rounds. Now suppose by contradiction that not all assigned variables are assigned according to  $\varphi$  when the unassignment step ends. Let  $U$  be the set of variables that remain assigned when the unassignment step ends, and whose assignment disagrees with  $\varphi$ . Every  $x \in U$  supports at least  $n^2p/10$  clauses w.r.t. to  $\psi$ , but each such clause must contain another variable on which  $\psi$  and  $\varphi$  disagree (since the clause is satisfied by  $\varphi$ , and  $\varphi(x) = \text{false}$ ). Thus, we have  $n^2p|U|/10$  clauses each containing at least two variables from  $U$ . Since  $U \cap \mathcal{H} = \emptyset$  (by the first part of this argument) it follows that  $|U| \leq e^{-\Theta(n^2p)}n < n/10^4$ , contradicting Proposition 1. ■

**Proposition 9.** *The exhaustive search in Step 3 of *SupportSAT* completes in polynomial time with a satisfying assignment of  $F$ .*

**Proof.** By Proposition 8, the partial assignment at the beginning of the exhaustive search step is partial to some satisfying assignment of the entire formula. Therefore the exhaustive search will succeed. Further observe that the unassigned variables are a subset of the non-core variables (Proposition 8). Proposition 3 then guarantees that the running time of the exhaustive search will be at most polynomial. ■

Theorem 1 then follows from Propositions 7-9.

## 8 Discussion

Our starting point in this work is an experimentally observed phenomenon – the remarkable success of simple algorithms that use the notion of support for near-threshold random 3CNF formulas, while heuristics of similar flavor that don't use the support are far less effective. In search of a unifying rule that rigorously explains this phenomenon, we define the Support Paradigm. We describe an algorithm which is part of this paradigm and rigorously show its effectiveness for some natural distribution over satisfiable 3CNF formulas ( $\mathcal{P}_{n,p}^{\text{plant}}$  with  $n^2p$  greater than some sufficiently large constant).

Combining our result with the work in [2] draws the following interesting picture. We show that some variation of RWalkSAT that takes into account the notion of support succeeds *whp* in finding a satisfying assignment for sufficiently “dense”  $\mathcal{P}_{n,p}^{\text{plant}}$  formulas. For the same clause-density regime, [2] shows that RWalkSAT, which disregards the notion of support, fails *whp* to find a satisfying assignment, and not even an assignment which is closer than say  $n/3$  to the planted one. This mirrors the near-threshold picture in the following sense. Experiments predict that RWalkSAT fails to find a satisfying assignment for random 3SAT instances with already  $2.65n$  clauses [26], while variants of RWalkSAT which take the support into account succeed as far as  $4.21n$  clauses [27, 5]. Thus in some rigorous sense we provide evidence for the fact that using the notion of support gives a clear edge.

To prove Theorem 1 we actually prove the following:

- The simulated-annealing step in SupportSAT (Steps 1 and 2) set (almost all) **frozen** variables in  $F$  correctly (that is, according to the planted assignment).
- The exhaustive search completes the assignment of the rest.

The latter implies that the frozen variables embed enough “support structure” so that a support-based simulated-annealing heuristic sets their assignment correctly. This asserts an interesting connection between the clustering phenomena, the notion of frozen variables and the success of support-based simulated annealing.

Recall the definition of a core (Definition 2) and Proposition 2 which claims that typically all but a small fraction of the variables belong to the core. The main ingredient in the definition of a core is the notion of support. This is also the key ingredient in the proof of Proposition 4 (given in complete in [11]) – which claims that the core variables are typically frozen. This observation strengthens the connection between the clustering phenomenon and the notion of support. Thus our result suggests the notion of support as a central piece in understanding the structure of typical instances in different SAT distributions, and in turn their average-case complexity.

Hopefully, our new approach will encourage the development of further heuristics that prove useful in practice – based on the notion of support. As part of this line of research, it will be interesting to check experimentally whether the subprocedures that compose SupportSAT (Directed Walk, or the refinement step, Step 2) are effective in other settings as well, for example the below-threshold random 3SAT regime.

### Acknowledgement

We would like to thank Michael Krivelevich for his helpful remarks.

## References

1. D. Achlioptas and F. Ricci-Tersenghi. On the solution-space geometry of random constraint satisfaction problems. In *STOC '06: Proceedings of the thirty-eighth annual ACM symposium on Theory of computing*, pages 130–139, 2006.
2. M. Alekhnovich and E. Ben-Sasson. Linear upper bounds for random walk on small density random 3cnfs. *Electronic Colloquium on Computational Complexity (ECCC)*, (016), 2004.
3. N. Alon and N. Kahale. A spectral technique for coloring random 3-colorable graphs. *SIAM J. on Comput.*, 26(6):1733–1748, 1997.
4. N. Alon, M. Krivelevich, and B. Sudakov. Finding a large hidden clique in a random graph. *Random Structures and Algorithms*, 13(3-4):457–466, 1998.
5. J. Ardelius and E. Aurell. Behavior of heuristics on large and hard satisfiability problems. *Phys. Rev.*, (E 74, 037702), 2006.
6. E. Ben-Sasson, Y. Bilu, and D. Gutfreund. Finding a randomly planted assignment in a random 3CNF. *manuscript*, 2002.
7. A. Blum and J. Spencer. Coloring random and semi-random  $k$ -colorable graphs. *J. of Algorithms*, 19(2):204–234, 1995.
8. A. Braunstein, M. Mezard, and R. Zecchina. Survey propagation: an algorithm for satisfiability. *Random Structures and Algorithms*, 27:201–226, 2005.
9. H. Chen. An algorithm for sat above the threshold. In *SAT*, pages 14–24, 2003.
10. A. Coja-Oghlan, M. Krivelevich, and D. Vilenchik. Why almost all  $k$ -colorable graphs are easy. In *24th International Symposium on Theoretical Aspects of Computer Science*, volume 4393 of *Lecture Notes in Comput. Sci.*, pages 121–132, 2007.
11. A. Coja-Oghlan, M. Krivelevich, and D. Vilenchik. Why almost all satisfiable  $k$ -cnf formulas are easy. In *13th conferene on Analysis of Algorithms*, 2007. to appear.
12. U. Feige. Relation between average case complexity and approximation complexity. In *STOC '02: Proceedings of the thirty-fourth annual ACM symposium on Theory of computing*, 2002.
13. U. Feige and R. Krauthgamer. Finding and certifying a large hidden clique in a semirandom graph. *Random Structures and Algorithms*, 16(2):195–208, 2000.
14. U. Feige, E. Mossel, and D. Vilenchik. Complete convergence of message passing algorithms for some satisfiability problems. In *RANDOM'06*, volume 4110 of *Lecture Notes in Comput. Sci.*, pages 339–350. Springer, Berlin, 2006.
15. U. Feige and D. Vilenchik. A local search algorithm for 3SAT. Technical report, The Weizmann Institute of Science, 2004.
16. A. Flaxman. A spectral technique for random satisfiable 3CNF formulas. In *Proc. 14th ACM-SIAM Symp. on Discrete Algorithms*, pages 357–363, 2003.
17. E. Friedgut. Sharp thresholds of graph properties, and the  $k$ -sat problem. *J. Amer. Math. Soc.*, 12(4):1017–1054, 1999.
18. C. Hui and A. Frieze. Coloring bipartite hypergraphs. In *Proceedings of the 5th International IPCO Conference on Integer Programming and Combinatorial Optimization*, pages 345–358, 1996.
19. A. C. Kaporis, L. M. Kirousis, and E. G. Lalas. The probabilistic analysis of a greedy satisfiability algorithm. In *Proc. 10th Annual European Symposium on Algorithms*, volume 2461 of *Lecture Notes in Comput. Sci.*, pages 574–585. Springer, Berlin, 2002.
20. A. C. Kaporis, L. M. Kirousis, Y. C. Stamatiou, M. Vamvakari, and M. Zito. Coupon collectors,  $q$ -binomial coefficients and the unsatisfiability threshold. In *Proc. Theoretical Computer Science: 7th Italian Conference*, volume 2202 of *Lecture Notes in Comput. Sci.*, pages 328–338. Springer, Berlin, 2001.
21. E. Koutsoupias and C. H. Papadimitriou. On the greedy algorithm for satisfiability. *Info. Process. Letters*, 43(1):53–55, 1992.
22. M. Krivelevich and D. Vilenchik. Solving random satisfiable 3cnf formulas in expected polynomial time. In *Proc. 17th ACM-SIAM Symp. on Discrete Algorithms*, pages 454–463, 2006.
23. L. Levin. Average case complete problems. *SIAM J. on Comput.*, 15:285–286, 1986.
24. M. Mezard, T. Mora, and R. Zecchina. Clustering of solutions in the random satisfiability problem. *Physical Review Letters*, 94:197–205, 2005.
25. C.H. Papadimitriou. On selecting a satisfying truth assignment. In *Proceedings of the 32th Annual Symposium on Foundations of Computer Science (FOCS '91)*, 1991.
26. A. J. Parke. Scaling properties of pure random walk on random 3-sat. In *n Proceedings of the 8th International Conference on Principles and Practice of Constraint Programming*, 2002.
27. S. Seitz, M. Alava, and P. Orponen. Focused local search for random 3-satisfiability. *Journal of Statistical Mechanics*, (P06006):1–27, 2005.