Journal on Satisfiability, Boolean Modeling and Computation 3 (2007) 125-139

It's All about the Support: A New Perspective on the Satisfiability Problem

Dan Vilenchik vilenchi@post.tau.ac.il
School of Computer Science
Faculty of Exact Sciences, Tel-Aviv University

Abstract

We study a new approach to the satisfiability problem, which we call the *Support Paradigm*. Given a CNF formula F and an assignment ψ to its variables we say that a literal x *supports* a clause C in F w.r.t. ψ if x is the *only* literal that evaluates to true in C. Our focus in this work will be on heuristics that obey the following general template: start at some assignment to the variables, then iteratively, using some predefined (greedy) rule, try to minimize the number of unsatisfied clauses (or the distance from some satisfying assignment) 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 for a certain distribution over satisfiable k-CNF 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 rule on the support, seem to 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.

Keywords: *average case analysis, random k-SAT, efficient heuristics, computational complexity*

Submitted May 2007; revised October 2007; published November 2007

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 k-SAT, this goal is too ambitious as shown by the theory of NP-completeness. 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 $k, c, n > 0$ (c may depend on n, k), choose $m = cn$ clauses uniformly at random out of $2^k\binom{n}{k}$ possible ones. Despite its simplicity, many essential properties of this model are yet to be understood (for $k \geq 3$). 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 [13, 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 [18]. 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* while one with ratio smaller than d is (when writing *whp* we mean with probability tending to 1 as n goes to infinity). For $k = 3$ the threshold is known to be at least 3.42 [20] and at most 4.506 [12].

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

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

Our focus in this work will be on heuristics that obey the following general template: start at some assignment to the variables, then iteratively, using some predefined (greedy) rule, try to minimize the number of unsatisfied clauses (or the distance from some satisfying assignment) until a satisfying assignment is reached (or failure due to exceeding some maximal number of allowed steps). We say that such a heuristic is part of the Support Paradigm if the greedy rule uses the support as its main criterion. In this work we present a new algorithm in the Support Paradigm and rigorously prove its effectiveness for a certain distribution over satisfiable k-CNF formulas known as the planted distribution.

Part of the motivation for this work comes from recent experimental results [27, 5] showing that some simple variants of the well known RWalkSAT algorithm [25], which base their greedy rule on the support (although the notion of support is not referred to explicitly in any of these works), seem to be effective for solving random 3SAT formulas in the “hard” near-threshold regime. Specifically, the experimental results suggest that these algorithms may be efficient in finding satisfying assignments for random 3SAT instances in $\mathcal{P}_{n,m}$ with m/n as large as 4.21 (the conjectured satisfiability threshold for 3SAT is roughly 4.26). In contrast, the “original” RWalkSAT heuristic, which is not part of the Support Paradigm, seems to consume super-polynomial already for $m/n = 2.65$ [26].

2. Our Contribution

Motivated by a search of a unifying rule that may contribute to the understanding of 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 k-SAT distribution with clause-variable ratio greater than some constant (an exact definition of the Planted k-SAT distribution is given in Section 2.1). Our results are thus in line with the experimentally-observed advantage of the algorithms in [27, 5] over RWalkSAT.

To keep the presentation simple we shall confine ourselves to the “canonical” case $k = 3$, and just point out that our result extends to any fixed k.

One disclaimer is due before we proceed. We do not claim that our result provides a direct explanation as to why certain algorithms seem to perform well in the below-threshold regime. For one, we deal with higher densities in which the instances are typically (by typically we mean *whp* over formulas from the concerned distribution) more structured. Nevertheless, in a recent work [1] concerning the below-threshold regime of $\mathcal{P}_{n,m}$ (for k some sufficiently large constant) a structure quite similar in spirit to our notion of core (Section 5.1) is proved to typically exist for such sparse formulas and is responsible for the existence of frozen variables in that regime. In fact, this structure is also defined using the notion of support (although this notion is not referred to explicitly in [1]).

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 of 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 the following structure of the solution space of random k-CNF formulas with clause-variable ratio just below the satisfiability threshold was suggested [24]. 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 assignments coincide in **all** satisfying assignments within that cluster. Part of this picture was recently rigorously proved in [1] for k-SAT with $k \geq 8$.

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 with respect to 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 less obvious (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 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 $\mathcal{P}^{\text{plant}}_{n,p}$. 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 models are of interest in computational complexity [13] and are favored by many researchers in the context of SAT [17, 6, 21], and also for other optimization problems such as max clique, min bisection, and coloring [3, 4, 7, 19, 14] to mention just a few. Another nice feature of planted-solution distributions is the fact that they are efficiently sampleable. Planted distributions may also provide a good model for statistical problems where the constraints are correlated in such a way as to be consistent (or statistically correlated) with a pre-specified assignment of the variables.

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

The algorithm SupportSAT(F), which belongs to the Support Paradigm, is described in Figure 2 in Section 4. The proof of Theorem 1 also reveals an interesting connection between the notion of support and the notion of frozen variables. Details in Section 5.1.

Combining our result with the work in [2] draws the following interesting picture. We show that SupportSAT, which can be viewed as a variation on the classical RWalkSAT, succeeds *whp* in finding a satisfying assignment for sufficiently dense $\mathcal{P}^{\text{plant}}_{n,p}$ formulas. For the same clause-density regime it is shown in [2] that RWalkSAT, which disregards the support, fails[1] *whp* to find a satisfying assignment, and not even an assignment which is closer than say $n/3$ to the planted one. This mirrors nicely the near-threshold picture: experiments predict that RWalkSAT fails to find a satisfying assignment for random 3SAT instances with clause-variable ratio greater than 2.65 [26], while variants of RWalkSAT which take the support into account succeed as far as 4.21 [27, 5].

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 first to address planted instances in the constant average degree regime (some sufficiently large constant) were Alon and Kahale in the context of k-colorable graphs [3].

[1] Throughout when we say that a heuristic fails to produce an optimal solution we always mean that it fails when spending polynomial time (and similarly for success).

Building upon their techniques Flaxman [17] presents an efficient algorithm for $\mathcal{P}_{n,p}^{\text{plant}}$ (the same regime that we study). The algorithm in [17] 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 $\mathcal{P}_{n,p}^{\text{plant}}$ [16, 15, 22].

In [9] the *uniform* distribution over satisfiable 3CNF formulas with a linear number of clauses is studied (again, the average degree is some sufficiently large constant). [9] describe an exponential time algorithm which *whp* solves such instances (in fact this algorithm is also part of the Support Paradigm, and in some parts our algorithm is inspired by [9]). [9] 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 "purely" part of the Support Paradigm). The analysis in [11] actually implies that Theorem 1 can be restated and reproved for the uniform distribution as well.

4. The Algorithm

We start with a high-level description of our algorithm. Given a formula F and an assignment ψ to its variables, we say that the assignment of x is *suspicious* in ψ if it supports very few clauses w.r.t. ψ. The main part of the algorithm is a simple greedy procedure in which iteratively the assignment of suspicious variables is flipped. 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 is "simple", one can find a satisfying assignment using some of-the-shelf heuristic.

The fast cool-down process is implemented via a procedure that we call a **Directed Walk** (inspired by the work of [9]), and then another procedure with a more refined flipping 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.

Remark 1. *The reader may wonder at this point if one can extend the greedy part of the algorithm (Steps 1 and 2) to find a satisfying assignment (and let go of step 3)? The answer is probably no, at least not in our planted setting. Every variable is expected to appear in $7\binom{n}{2}p = \Theta(n^2p)$ clauses (which we think of as constant in our analysis). Further, the number of clauses in which a variables appears is binomially distributed. Thus, with constant probability some appear very scarcely (say once or twice). Therefore in some sense such variables don't show enough structure to allow a greedy procedure of the sort we use to set them correctly. On the other hand, these variables induce a "simple" formula for which exhaustive search is efficient for example.*

4.1 The Directed Walk

We now introduce the sub-procedure Directed Walk 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($F, \varepsilon$)
  $\psi_0 \leftarrow$ an arbitrary assignment to the variables.
  for $i = 1$ 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.
  return $\psi_{3/\varepsilon}$.
```

Figure 1. Directed Walk

While RWalkSat ('R' stands for Random) uses randomness when deciding which variable to flip, the "Directed Walk" employs a deterministic directing rule. Directed Walk can be used with other measurements – e.g. flip the assignment of the εn 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 (the notations we use are clarified right after the following figure).

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

Figure 2. SupportSAT

Notations. $Support_F(x, \psi)$ is the support of a variable x in F w.r.t. an assignment ψ; $\psi^{(x)}$ is the assignment ψ 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 ψ, $Support_F(x, \psi)$ counts only clauses where all variables are assigned.

The last step of SupportSAT (Step 3) is an exhaustive search. The variables that we want to exhaustively search are those which are still suspicious after the greedy step ends (Steps 1 and 2). To separate the suspicious variables form the reliable ones we employ a careful unassignment step (lines 8-12) which leaves assigned only variables with large support.

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 (Section 5.1). All propositions in this section appear also in [3, 17] for example (maybe stated differently); therefore we only state the propositions and give an outline of the proof (which can be easily reconstructed into a full proof).

The following is a discrepancy property which excludes the existence of a small subformula of F in which variables appear too many times (much more than expected).

Proposition 1. *Let F be a random formula 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 *nonempty 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 at most

$$\binom{k}{2}(n-2)(2^3-1) \leq 4k^2n.$$

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

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

The "..." can be easily filled using 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 $\mathcal{P}_{n,p}^{\text{plant}}$. Amongst other things, the core captures the notion of frozen variables. Also observe that the main ingredient which is used in the definition of a core is the support.

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

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

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

Proposition 2. *Let F be a random formula distributed according to $\mathcal{P}_{n,p}^{\text{plant}}$ with $n^2p \geq C_0$, C_0 some sufficiently large constant. Then whp there exits a t-core $\mathcal{H}$ with $t = n^2p/4$ w.r.t. the planted assignment φ. 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 w.r.t. the planted assignment.

Proof. Consider the following procedure which we prove defines a core.

```
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,

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

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

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

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 $\{a_1, a_2, ..., a_{i-1}\} \cup B$, 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 δn variables, and there are at least $n^2p/40 \cdot \delta n/2 \cdot 1/3 = n^2p/240 \cdot \delta n$ 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 $\delta = e^{-\Theta(n^2p)} << 10^{-4}$.

■

The next two propositions are given without a proof as their proofs are quite technical and can be found in complete in [17] and [11] respectively.

Proposition 3. *Let F be a random formula distributed according to $\mathcal{P}^{\text{plant}}_{n,p}$ with $n^2p \geq C_0$, C_0 some sufficiently large constant. Let $\mathcal{H}$ be a maximum size $n^2p/4$-core of F, and let $F[V \setminus \mathcal{H}]$ be the formula induced by the non-core variables. 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 a random formula distributed according to $\mathcal{P}^{\text{plant}}_{n,p}$, with $n^2p \geq C_0$, 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 the Directed Walk

In this section we analyze a typical execution of Directed Walk for $\mathcal{P}^{\text{plant}}_{n,p}$, n^2p 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 a random formula distributed according to $\mathcal{P}^{\text{plant}}_{n,p}$, $n^2p \geq C_0$, C_0 some sufficiently large constant. Then whp* Directed Walk$(F, 10^{-5})$ *enjoys the following property: after at most $3 \cdot 10^5$ rounds the output assignment differs from the planted one on at most $n/10^5$ variables.*

Remark 2. The constant 10^5 is arbitrary. In fact, one can show that *whp* Directed Walk(F, ε) finds after $3/\varepsilon$ rounds an assignment at distance at most εn from the planted one for ε 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 [16, 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's implicitly proved in [21], though the directing measure is not the support.

Before proving the proposition we make some further observations.

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

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

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

Proposition 6. *Let F be a random formula distributed according to $\mathcal{P}_{n,p}^{\text{plant}}$, $n^2 p \geq C_0$, C_0 some sufficiently large constant, and let φ be its planted assignment. Then whp F is 10^{-5}–directable w.r.t. φ.*

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

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

The union bound then guarantees that *whp* no misleading ψ exists as there are at most 2^n possible ways to choose ψ. 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 $Pr[B_1 \leq B_2] \leq e^{-g(\beta) p N^2}$, where $g : (0,1) \to (0, \infty)$ is a monotonically increasing function.*

Proof.(Lemma 2) Consider some assignment ψ at distance βn from φ, $\beta \geq \varepsilon$. 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. ψ, and every one of them could have been included in F (since φ satisfies all of them). On the other hand, f supports $\binom{n}{2}$ clauses w.r.t. ψ, but clauses where both φ and ψ agree on the assignment of the other two variables cannot be included in φ as they are not satisfied by φ – and there are $\binom{(1-\beta)n}{2}$ of them. Therefore we get:

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

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 ψ is k-misleading w.r.t. φ then in particular there exist k pairs of variables $(t_1, f_1), \ldots, (t_k, f_k)$ s.t. $\mathit{Support}(t_i, \psi) \leq \mathit{Support}(f_i, \psi)$. 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 ε 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. The lemma follows by applying the union bound. ■

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*

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

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^{-5}–directable w.r.t. φ. Since this is the case *whp*, as asserted by Proposition 6, Proposition 5 follows. For two assignments ψ, φ, define $T(\psi, \varphi)$ to be the set of variables on which ψ and φ agree, and $F(\psi, \varphi)$ the set of variables on which they disagree. Let $E_\varepsilon(\psi)$ be the set of εn variables with lowest support w.r.t. ψ. Observe that if φ is some satisfying assignment of F, and ψ is the current assignment that Directed Walk(F, ε) 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 $\varepsilon = 10^{-5}$ (as required by Proposition 5). Suppose at first that for every ψ at distance $\geq \varepsilon n$ from φ, $|T(\psi, \varphi) \cap E_\varepsilon(\psi)| \leq \varepsilon n/3$. 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 φ 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 $n/(\varepsilon n/3) = 3/\varepsilon = 3 \cdot 10^5$ rounds an assignment ψ' at distance at most εn from φ is reached.

It remains to prove that the above picture is indeed the case. To this end, consider a "bad" assignment ψ at distance $> \varepsilon n$ from φ 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 ψ and φ 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 $f_1, f_2, ..., f_k$ be variables in $F(\psi, \varphi) \setminus E_\varepsilon(\psi)$, and $t_1, t_2, ..., t_k$ be variables in $T(\psi, \varphi) \cap E_\varepsilon(\psi)$. For every t_i, f_j, $\mathit{Support}(t_i, \psi) \leq \mathit{Support}(f_j, \psi)$ (by the definition of $E_\varepsilon(\psi)$ and the choice of the t_i's and the f_j's). However this means that ψ is k-misleading w.r.t. φ (as the Hamming distance between ψ and φ is greater than εn), which contradicts that fact that F is ε-directable w.r.t. φ. ■

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 $\mathcal{P}^{\text{plant}}_{n,p}$, n^2p greater than some sufficiently large constant, is typical. Thus proving Theorem 1 reduces to proving that SupportSAT (always) finds a satisfying assignment in polynomial time for typical formulas. In all the propositions below we assume that F is typical; we let $\mathcal{H}$ be the core promised in Proposition 2 and φ the planted assignment of F.

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

Proof. Let B_i be the set of core variables whose assignment in ψ_i disagrees with φ at the beginning of the i^{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 $B_{\log n} = \emptyset$). 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 (that is, $|B_{j+1}| \geq |B_j|/2$). 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^{th} iteration, and therefore, x supports at least $n^2p/10$ clauses w.r.t. ψ_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 φ, each such clause must contain another variable on which φ and ψ_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^{th} iteration. This is because x supports less than $n^2p/10$ clauses w.r.t. ψ_j. Since x supports at least $n^2p/4$ clauses w.r.t. φ 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. φ). For conclusion, let $U = B_j \cup B_{j+1}$; 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 $n^2p/20 \cdot |B_{j+1}| \geq n^2p/30 \cdot |U|$, and to begin with (by Proposition 5) $|B_0| \leq n/10^5$ (we can assume w.l.o.g. that $|B_{j+1}| \leq n/10^5$, otherwise just take the first $n/10^5$ variables in B_{j+1} and therefore $|U| \leq 2n/10^5 \leq n/10^4$). This contradicts Proposition 1. ■

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

Proof. The core variables are assigned according to φ 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. φ, and also w.r.t. τ_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 φ when the unassignment step ends. Let U be the set of variables that remain assigned when the unassignment step ends, and whose assignment disagrees with φ. Every $x \in U$ supports at least $n^2p/10$ clauses w.r.t. to ψ, but each such clause must contain another variable on which ψ and φ disagree (since the clause is satisfied by φ, and $\varphi(x) = false$). 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 patrial 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.

Taking a closer look at the proof of Theorem 1 it turns out that we actually prove the following:

- The support-based greedy step of 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" to allow a support-based greedy heuristic to set their assignment correctly. This asserts an interesting connection between the clustering phenomenon, the notion of frozen variables and the success of support-based greedy heuristics.

8. Discussion

In this work we introduce a new approach to SAT-solving heuristics. The main building stone of our approach is the notion of support. As a case study, we rigorously show the effectiveness of a simple support-based algorithm for the Planted 3SAT distribution with clause-variable ratio some sufficiently large constant.

The notion of support seems to be useful in other contexts as well. The experimental results in [27, 5] seem to indicate that simple variations on the classical RWalkSAT (which now take into account the support) outperforms RWalkSAT by far, and seem to remain efficient way into the "hard" near threshold regime. Also in [1] the existence of separated clusters and frozen variables is shown for $\mathcal{P}_{n,m}$ (with m/n some suitably chosen parameter below the threshold, and k some large enough constant) – the notion of support plays a major role in showing the existence of frozen variables (in fact a similar structure to our core is shown to exist *whp*).

Therefore there is hope that our new approach will be applicable in various distributions and clause-density regimes, and 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 in below-threshold random 3SAT formulas.

Acknowledgement

We thank Michael Krivelevich and the anonymous referees for their 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 satifiable k-CNF formulas are easy. In *13th conferene on Analysis of Algorithms*, 2007. to appear.

[12] O. Dubois, Y. Boufkhad, and J. Mandler. Typical random 3-SAT formulas and the satisfiability threshold. In *Proc. 11th ACM-SIAM Symp. on Discrete Algorithms*, pages 126–127, 2000.

[13] U. Feige. Relation between average case complexity and approximation complexity. In *STOC '02: Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*, 2002.

[14] 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.

[15] 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.

[16] U. Feige and D. Vilenchik. A local search algorithm for 3SAT. Technical report, The Weizmann Institute of Science, 2004.

[17] A. Flaxman. A spectral technique for random satisfiable 3CNF formulas. In *Proc. 14th ACM-SIAM Symp. on Discrete Algorithms*, pages 357–363, 2003.

[18] E. Friedgut. Sharp thresholds of graph properties, and the k-SAT problem. *J. Amer. Math. Soc.*, **12**(4):1017–1054, 1999.

[19] 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.

[20] 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.

[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 3SAT. 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.