A Sharp Threshold for the Phase Transition of a restricted Satisfiability Problem for Horn Clauses

Paul E. Dunne and Trevor J.M. Bench-Capon

Dept. of Computer Science; University of Liverpool; Liverpool, L69 7ZF, Great Britain

Abstract: In this paper we examine a variant, k-*HSAT*, of the well-known Satisfiability problem, wherein formula instances are limited to CNF formulae having exactly k literals in each clause at most one of which is un-negated. In addition, rather than seeking instantiations of the propositional variables that satisfy the formula, satisfying instantiations having at least $k-1$ variables set to 1 are required (such instantiations being referred to as *acceptable*). We show, using analytic methods, that this variant exhibits a phase transition whose behaviour is significantly different from that of the classical k-SAT problem concerning which various researchers have presented empirical evidence for the existence of a constant θ_k such that randomly chosen instances of k-*SAT* on n variables having $m(n)$ clauses are 'almost certainly' satisfiable if $m(n)/n < \theta_k$ and 'almost certainly' unsatisfiable if $m(n)/n > \theta_k$.

In this paper we prove a *sharp* threshold result for k-*HSAT for all* fixed $k \geq 2$. Specifically, for $\theta_k = (k-1)(k+1)/k!$ it is proved that if ϕ is chosen uniformly at random from instances of k-*HSAT* on n variables and $m(n)$ clauses then

$$\lim_{n\to\infty} Prob\,[\,\phi\ \textit{has an acceptable instantiation}\,] \;=\; 1 \;\text{ if }\; \frac{m(n)}{n^{k-1}\log_e n} < \theta_k$$
$$\lim_{n\to\infty} Prob\,[\,\phi\ \textit{has an acceptable instantiation}\,] \;=\; 0 \;\text{ if }\; \frac{m(n)}{n^{k-1}\log_e n} > \theta_k$$

Keywords: Phase transition, Satisfiability, Horn Clause formula.

1. Introduction

The Satisfiability problem, and in particular the restriction of this to propositional formulae in conjunctive normal form with exactly 3 literals in each clause (3-*SAT*) has been the subject of extensive investigation with respect to the phenomenon known as a *phase transition*. Informally, this is the behaviour whereby as the number of clauses, m, increases, a 'typical' m-clause instance of 3-*SAT* on n propositional variables progresses from being usually satisfiable to usually unsatisfiable. Furthermore, if one examines this transition in terms of the ratio of number-of-clauses to number-of-variables (m/n) there is an apparent sharp threshold, θ, such that typical instances with $m/n < \theta$ are almost always satisfiable, and typical instances with $m/n > \theta$ are almost always not satisfiable. More formally, if $S(m,n)$ is the proportion of possible instances of 3-*SAT* with n variables and m clauses, that are satisfiable, then experimental studies posit the existence of a threshold θ such that

$$\forall\ \alpha < \theta \ \lim_{n\to\infty} |\,S(\alpha n, n)\,| \;=\; 1$$
$$\forall\ \alpha > \theta \ \lim_{n\to\infty} |\,S(\alpha n, n)\,| \;=\; 0$$

The empirical studies presented by [5], [12], and [13] suggest a value for θ of about 4.24. Analytic methods confirm the existence of a sharp threshold although this has yet to be identified exactly [6]. Some progress, however, has been made in narrowing the region in which such a threshold would have to lie. Thus, initial upper bounds on the clause-to-variable ratio m/n were obtained in [4] where it was demonstrated that 'almost all' instances with $m/n > 5.19..$ are unsatisfiable. This has been improved in [10,11] where the bound is lowered to 4.598... A lower bound of 3.003 is announced in [7], improving the earlier bound of 2.99 given in [2]. Hence it is known that, for 3-*SAT*, $3.003 \leq \theta \leq 4.598$. Analytic techniques have proved more successful with 2-*SAT* (which can be decided in time linear in m), the work of [3] and [8] proving a sharp threshold at $m/n = 1$ for this problem. An interesting and more general problem is studied, by experimental techniques, in [14]. In this work the existence of prime implicates of length t in random 3-*SAT* instances is considered: 3-*SAT* being the special case of this problem corresponding to $t = 0$. It is shown that differing values of t induce differing thresholds for the phase transition.

Although the phase transition phenomenon in 3-*SAT* may be viewed as of purely combinatorial interest, there are, in fact, powerful pragmatic reasons for examining it. The correlation between instances lying

The work reported in this paper was carried out with support from EPSRC Grant GR/L/77089

close to the threshold region and instances which prove to place the greatest computational demands on heuristic search methods has been observed in empirical studies. A similar correlation is apparent between instances falling at ‘extreme’ distances from the threshold and instances for which search heuristics can deliver an answer very quickly. These, experimentally observed, connections provide two of the strong practical motivations for investigating phase transitions: instances lying on or close to the threshold provide a natural testbed of ‘hard’ cases with which to evaluate new heuristic methods; and the ease with which ‘extreme’ instances can often be solved can provide directions for designing efficient (on average) methods for solving some classes of hard, e.g. *NP*-complete, problems. A classic example of the latter application is the methods of [1] which give fast average-case methods for solving the Hamiltonian Cycle problem in graphs with ‘sufficiently many’ edges.

The fact that phase transition phenomena arise in the particular form of Satisfiability represented by k-*SAT* and in the more general problem examined in [14] raises the question of how the transition functions may be affected by considering further restrictions on instances of k-*SAT*. The present paper is concerned with a variant of k-*SAT* but one which can be decided in polynomial time. Furthermore the specific problem we study is strongly motivated by issues arising in the domain of Logic Programming. A formal definition of the problem, k-*HSAT*, is given in the next section, however, in order to provide a preliminary motivation for it we observe that the set of admissible clauses in a given instance of k-*HSAT* is such that at most one literal in each clause may occur in positive form, and it is required that satisfying instantiations set at least $k-1$ variables to the Boolean value 1.

The reason for choosing this restriction on the clauses is that it corresponds to the knowledge representation paradigms popular in Artificial Intelligence. Thus the Logic Program paradigm typically restricts itself to logic programs which use the Horn Clause. A Horn Clause is of the form: $a_1 \leftarrow b_1 \wedge b_2 \wedge \cdots \wedge b_n$ where a_1 is either a positive literal or the empty clause and $b_1, \ldots, b_n$ are positive literals. As is evident such clauses, when transformed to CNF, limit us to at most one positive literal in each elementary disjunction. Thus k-*HSAT* is sufficient to express all typical logic programs and production rule bases.

The main result of this paper is an analytic proof of a sharp threshold for k-*HSAT*, wherein it is established that the function $r(n,k) = n^{k-1}\log_e n$, has the following property: $\exists\, \theta_k$, (a constant> 0) such that ‘almost all’ $m(n)$-clause instances, ϕ, of k-*HSAT*(n) with $m(n)/r(n,k) < \theta_k$ (resp. $m(n)/r(n,k) > \theta_k$) are such that k-*HSAT*$(\phi) = 1$ (resp. k-*HSAT*$(\phi) = 0$). The proof of this result, furthermore, identifies the exact value of θ_k as $(k-1)(k+1)/k!$.

We note here a number of points of interest about these results. Firstly, with the exception of 2-*SAT* in [3,8], no explicit (as opposed to existential) *sharp* threshold results had been proved analytically for variants of Satisfiability: our result not only improves this situation, but also gives the first example of a sequence of sharp threshold results parameterised in terms of the clause length k. A second point is that the analysis uses only established combinatorial mechanisms, such as the ‘first-moment’ method, and thus raises the possibility that these may be sufficiently powerful to sharpen the threshold for more general problems: a similar indication to this effect was already apparent in [10,11]. We note here, that Istrate and Ogihara have, independently, considered the phase-transition problem for satisfiability of general Horn-clause formulae [9], i.e. without the restriction that each clause contain exactly k literals, proving that a sharp threshold does not exist for this class..

In the next section we present some basic definitions and the notations used subsequently. In Section 3 the main analytic result on the threshold function for k-*HSAT*(n) is proved. Conclusions are presented in Section 4.

2. Definitions and Notation

$\mathbf{X_n} = < x_1, x_2, \ldots, x_n >$ denotes a set of n propositional variables. A *literal* is either a variable, x, or its negation $\bar{x}$. In the former case, x is said to be a positive *literal*; in the latter case it is said to be a *negative* literal. A *clause*, s, is a disjunction of literals; s is said to be *trivial* if it contains both the literal x and its negation, and is *non-trivial* otherwise. Similarly, a *product*, p, is a conjunction of literals. A *CNF formula*, ϕ, over $\mathbf{X_n}$ is a conjunction of (non-trivial) clauses $\{ s_1, s_2, \ldots, s_m \}$. For any CNF formula, ϕ, f_ϕ denotes the n-variable propositional logic function represented by ϕ. The decision problem *satisfiability* (*SAT*) asks whether a given CNF formula, ϕ, is such that there exists any instantiation, α, of the propositional variables $\mathbf{X_n}$ of ϕ, for which $f_\phi(\alpha) = 1$. A k-CNF formula is a CNF formula in which every clause has exactly k

literals. The decision problem k-SAT is the satisfiability problem restricted to k-CNF formulae.

A *rule* is an expression of the form $z \leftarrow y_1 \wedge y_2 \cdots \wedge y_k$ where $\{ y_1, \ldots, y_k \}$ are distinct variables from the set $\mathbf{X_n}$ and z is either the Boolean constant 0 or a variable in $\mathbf{X_n}$. All variables in a rule must be positive and it is assumed that $z \notin \{ y_1, \ldots, y_k \}$. If $p = y_1 \wedge \cdots \wedge y_k$, then we may express a rule, $z \leftarrow p$ by the logically equivalent clause, $(z \vee \bar{y}_1 \vee \bar{y}_2 \vee \cdots \vee \bar{y}_k)$.

From the definition of rule it follows that an equivalent clause contains at most one positive literal. We can thus consider the following class of clauses and CNF formulae.

The set of clauses over $\mathbf{X_n}$ containing exactly k literals at most one of which is positive is denoted $HC(n, k)$. The set of k-CNF formulae all of whose clauses are in $HC(n, k)$ is denoted $HCNF(n, k)$ with the subset of these that have exactly m clauses denoted $H(n, k, m)$. The decision problem with which this paper is concerned is a variant of k-SAT restricted to formulae in $HCNF(n, k)$.

Definition 2.1: k-$HSAT$ is the decision problem which takes as input a formula, $\phi \in HCNF(n, k)$ and asks if there is an instantiation, $\alpha \in \{ 0, 1 \}^n$ of $\mathbf{X_n}$ such that $f_\phi(\alpha) = 1$ and *at least* $k-1$ propositional variables in $\mathbf{X_n}$ are instantiated to 1 under α. If α is an instantiation that certifies an instance ϕ of k-$HSAT$, we say that α is *acceptable* for ϕ. An instance ϕ for which k-$HSAT(\phi)$ holds is said to be *useful*. •

A *property*, Π, is a subset of $\bigcup_{n=0}^{\infty} HCNF(n, k)$. A formula $\phi \in HCNF(n, k)$ is said to have property Π if $\phi \in \Pi$. A property is *monotone increasing* if $\phi \in \Pi \Rightarrow \psi \in \Pi$ for all ψ whose clauses are a superset of those in ϕ. Similarly, a property is *monotone decreasing* if $\phi \in \Pi \Rightarrow \psi \in \Pi$ for all ψ whose clauses form of *subset* of those of ϕ. It is easy to see that the property defined by useful instances of k-$HSAT$ is monotone decreasing.

3. The sharp threshold for random instances of k-HSAT.

It may be noted that any instantiation of $\mathbf{X_n}$ in which at most $k-2$ variables take the value 1 will, trivially, satisfy any $\phi \in HCNF(n, k)$: after setting any subset of at most $k-2$ variables to 1, any clause will either be satisfied or will have at least one (non-instantiated) negative literal remaining, thus ϕ can be satisfied by assigning all the remaining variables the value 0.

There are, however, good reasons why we should consider such satisfying instantiations as being uninteresting, hence the second condition imposed in Definition 2.1. In particular, the requirement that we have at least $k-1$ variables assigned the value **true** does not exclude any meaningful knowledge bases. In a knowledge based system, the rules are applicable only if the conditions in the antecedents are satisfied. Even though the falsity of the antecedent guarantees the truth of the rule, it does not justify its inclusion in a knowledge base, since such a rule could never be applied. If fewer than $k-1$ variables are **true**, no antecedent can be satisfied and so no rule can be applicable and hence the knowledge base as a whole is inapplicable in that situation. Our restriction thus does no more than admit only cases where the corresponding knowledge base has a potential application.

We consider the following two probability distributions on n-variable instances of k-$HSAT$.

D1. The distribution in which a random formula denoted subsequently by $F_{n,k,p}$ is chosen by considering each $s \in HC(n, k)$ in turn and including it with probability p, excluding it with probability $1-p$. All such decisions being made independently.

D2. The distribution in which a random formula denoted subsequently by $G_{n,k,m}$ is chosen uniformly at random from $HCNF(n, k, m)$.

Suppose that $r : (\mathbf{N} \times \mathbf{N}) \rightarrow \mathbf{R}$ is some (positive) real-valued function. In order to present results regarding phase transitions in k-$HSAT(n)$ we wish to determine: an appropriate function $r(n, k)$ and a constant, θ_k such that,

$$\forall\ \alpha > \theta_k \quad \lim_{n \rightarrow \infty} P\,[\, G_{n,k,\alpha r(n,k)} \text{ is useful} \,] \;=\; 0 \tag{3.1}$$

$$\forall\ \alpha < \theta_k \quad \lim_{n \rightarrow \infty} P\,[\, G_{n,k,\alpha r(n,k)} \text{ is useful} \,] \;=\; 1 \tag{3.2}$$

Distributions similar to $G_{n,k,m}$ are frequently employed in empirical studies. Analysis of this, however, can become rather cumbersome. For this reason our analytic results are derived in terms of the distribution $F_{n,k,p}$. It is well-known from the analogous random graph distributions (in which edges are included with probability p; resp. an m-edge graph is chosen uniformly at random from the set of n-vertex, m-edge graphs) that results obtained in one distribution may be translated into the other. In particular, if Π is a monotone (decreasing) property of formulae in the class $HCNF(\, n, k\,)$, the following fact holds.

Fact 3.1: If $p(n)$ is such that $\forall\ p \geq p(n)\ \lim_{n \to \infty} P\,[\, F_{n,k,p} \in \Pi\,] \ = \ 0$, then

$$m \geq m(n) = p(n)|HC(n, k)| \ \Rightarrow \ \lim_{n \to \infty} P\,[\, G_{n,k,m} \in \Pi\,] \ = \ 0$$

Similarly, if $p(n)$ is such that $\forall\ p \leq p(n)\ \lim_{n \to \infty} P\,[\, F_{n,k,p} \in \Pi\,] \ = \ 1$, then

$$m \leq m(n) = p(n)|HC(n, k)| \ \Rightarrow \ \lim_{n \to \infty} P\,[\, G_{n,k,m} \in \Pi\,] \ = \ 1$$

Proof: Follows directly from the analogous translations between random graph models, see, e.g. [1]. □

It follows from this fact, that we can identify a suitable function $r(n, k)$ to consider in (3.1, 3.2) by identifying a suitable probability p governing the presence or absence of clauses in $F_{n,k,p}$.

A central result in the derivation of our analytic bounds is the following characterisation of necessary and sufficient conditions for ϕ to be useful.

Lemma 3.1: ϕ is useful for k-$HSAT(\, n\,)$ if and only if there is a subset B of $\mathbf{X_n}$ such that $|B| \geq k-1$ and for which

Q1. $\forall\ \{\, y_1, \ldots, y_k\,\} \subseteq B,\ (\,\bar{y}_1 \vee \bar{y}_2 \cdots \vee \bar{y}_k\,) \notin \phi$.

Q2. $\forall\ \{\, y_1, \ldots, y_{k-1}\,\} \subseteq B,\ \forall\ z \in \mathbf{X_n} - B,\ (\,\bar{y}_1 \vee \bar{y}_2 \cdots \vee \bar{y}_{k-1} \vee z\,) \notin \phi$.

For brevity we shall subsequently refer to the relationship whereby a subset B has size at least $k-1$ and satisfies Q1 and Q2 with respect to a formula ϕ using the notation $Q(\, B, \phi\,)$.

Proof: $\Leftarrow$: Let $B \subseteq \mathbf{X_n}$ such that $Q(\, B, \phi\,)$ and consider the partial instantiation $x := 1, \forall\ x \in B$. Let ψ be the formula the results after removing, from ϕ, clauses which take the value 1 and literals which take the value 0 under this instantiation. Since Q1 holds, ψ cannot have any clauses that evaluate to 0; since Q2 holds, no clause in ψ can consist of a *single* positive literal and, therefore, every remaining clause must contain a negated literal. It follows that setting all uninstantiated variables to 0 will result in an instantiation that satisfies ϕ and has at least $k-1$ variables set to 1, i.e. is acceptable.

$\Rightarrow$: Suppose that ϕ is useful and let α be an acceptable instantiation. Choose B to be $\{\, x : x = 1\ \mathit{under}\ \alpha\,\}$. Since α is acceptable, $|\,B\,| \geq k-1$. Furthermore condition Q1 must hold (otherwise $f_\phi(\,\alpha\,) = 0$). Similarly, Q2 must hold, for otherwise there is a subset Y of B of size $k-1$ and a variable $z \notin B$ with $(\, z \vee \bigvee_{y \in Y} \bar{y}\,) \in \phi$, but any such clause takes the value 0 under α, contradicting the condition that α is an acceptable instantiation. □

Lemma 3.1 leads to the main result of this paper.

Theorem 3.1: Let $p = p(n) = \left(\dfrac{\log_e n}{n}\right)$.

$$\forall\ k \geq 2 \qquad \forall\ \alpha > k-1\ \lim_{n \to \infty} P\,[\, F_{n,k,\alpha\, p}\ \mathit{is\ useful}\,] = 0 \tag{a}$$

$$\forall\ \alpha < k-1\ \lim_{n \to \infty} P\,[\, F_{n,k,\alpha\, p}\ \mathit{is\ useful}\,] = 1 \tag{b}$$

Proof: For part (a) we employ a 'first-moment' argument. Let

$$Acc(\,\phi\,) \ =_{def} \ \{\ B \subseteq \mathbf{X_n} : Q(\, B, \phi\,)\ \}$$

That is, for a given ϕ, $Acc(\,\phi\,)$ is the set of distinct subsets of $\mathbf{X_n}$ that would guarantee the useful nature of ϕ from Lemma 3.1. For $\phi \in F_{n,k,p}$, $|\,Acc(\,\phi\,)\,|$ is a random variable taking values between 0 and $2^n - O(\, n^{k-2}\,)$. Letting $E\,[\,|\,Acc\,|\,]$ denote the expected value of this random variable, $\phi \in F_{n,k,p}$ be a random instance of k-$HSAT(\, n\,)$, and I_ϕ denote the indicator function which is 1 if ϕ is useful and 0 otherwise, we have,

$$P[\,\phi \textit{ is useful}\,] \;=\; \sum_{\phi} P[\,\phi\,]\, I_{\phi} \;\le\; \sum_{\phi} P[\,\phi\,]\,|Acc(\phi)| \;=\; E[\,|Acc(\phi)|\,] \tag{3.3}$$

To prove part (a) it suffices to identify p such that $\lim_{n\to\infty} E[|Acc(F_{n,k,p})|] = 0$. The Markov inequality in (3.3) then shows that $\phi \in F_{n,k,p}$ is almost certainly not useful.

$$E[\,|Acc(\phi)|\,] \;=\; \sum_{\phi} P[\phi]\,|Acc(\phi)| \;=\; \sum_{\phi} \sum_{B \subseteq \mathbf{X_n}} P[\phi]\, I_{Q(B,\phi)}$$

(where $I_{Q(B,\phi)}$ is the indicator function for the relation Q)

$$= \sum_{B}\sum_{\phi} P[\phi]\, I_{Q(B,\phi)} \;\le\; \sum_{B} P[\,Q(B,\phi)\,]$$

$$= \sum_{B} P[\,B \in Acc(\phi)\,] \tag{3.4}$$

We thus have,

$$P[\,\phi \textit{ is useful}\,] \;\le\; E[\,|Acc(\phi)|\,] \;\le\; \sum_{B \subseteq \mathbf{X_n}} P[\,B \in Acc(\phi)\,]$$

Hence,

$$E[\,|Acc(\phi)|\,] \;\le\; \sum_{r=k-1}^{n} \;\sum_{B:|B|=r} P[\,B \in Acc(\phi)\,] \;\le\; \sum_{r=k-1}^{n} \binom{n}{r} (1-p)^{\binom{r}{k}+\binom{r}{k-1}(n-r)} \tag{3.5}$$

Since, from Lemma 3.1, in order for $Q(B,\phi)$ to hold, the condition Q1 forbids $\binom{r}{k}$ clauses in ϕ and the condition Q2 forbids a further, $\binom{r}{k-1}(n-r)$ clauses. Simplifying (3.5) using the relationships $1+x \le \exp(x)$ and $\binom{n}{r} \le (n\,e/r)^r$, we obtain

$$E[\,|Acc(\phi)|\,] \;\le\; \sum_{r=k-1}^{n} \exp\left(-p\left\{ \binom{r}{k} + n\binom{r}{k-1} - r\binom{r}{k-1} \right\} + r\log_e(n/r) + r \right)$$

$$=_{def} \;\sum_{r=k-1}^{n} \exp(\,\beta(r)\,)$$

Substituting $p = \alpha \log_e n/n$, we get,

$$\beta(r) \;=\; -r\log_e n \left\{ \frac{\alpha}{r}\binom{r}{k-1} + \frac{\alpha}{nr}\binom{r}{k} - \frac{\alpha}{n}\binom{r}{k-1} - 1 + \frac{\log_e r}{\log_e n} - \frac{1}{\log_e n} \right\}$$

$$= -r\log_e n \left\{ \alpha\binom{r}{k-1}\left\{ \frac{1}{r} - \frac{1}{n} \right\} - 1 + \frac{\alpha}{nr}\binom{r}{k} + \frac{\log_e r}{\log_e n} - \frac{1}{\log_e n} \right\}$$

$$=_{def} \; -r\log_e n\, \gamma(r) \tag{3.6}$$

The remainder of the proof of part (a) falls into 4 parts. Recall that we have established so far, that $E[\,|Acc(\phi)|\,] \le \sum_{r=k-1}^{n} \exp(-r\log_e n\,\gamma(r))$, where the function $\gamma(r)$ depends also on α and n. If we can show for an appropriate choice of α and all r in the range of the summation that $\gamma(r) > \delta > 0$ (for some fixed δ) then this summation is bounded above by the geometric progression $\sum_{r=k-1}^{n} n^{-\delta r}$ which approaches 0 as n increases. We, thus, complete the proof as follows.

a1) We show that $\forall\, k \geq 2$, $\forall\, \varepsilon > 0$, $\exists\, N_\varepsilon$ such that with $\alpha = k - 1 + \varepsilon$, $\gamma(\,k - 1\,) > \varepsilon/2k$ whenever $n \geq N_\varepsilon$.

a2) We show that $\forall\, k \geq 2$, $\gamma(r) < \gamma(r + 1)$ for $k - 1 \leq r < s(n, k)$ where $s(n, k) = (k - 2)n/(k - 1)$ for $k \geq 3$, and $2n/(\,(1 + \varepsilon) \log_e n\,)$ for $k = 2$.[1)]

a3) We prove that $\forall\, k \geq 2$, $\gamma(\,r\,) > \varepsilon/2k$ for $s(n, k) \leq r \leq n$.

In combination, (a1)-(a3), allow us to deduce that:

$\forall\, k \geq 2,\ \forall\, \varepsilon > 0,\ \forall\, r : k - 1 \leq r \leq n,\ \forall\, n \geq N_\varepsilon \quad \gamma(\,r\,) > \varepsilon/2k$

a4) Finally, we prove that the geometric progression bounding $E\,[\,|Acc(\,\phi\,)|\,]$ approaches 0 as n approaches infinity.

Recall that the function $\gamma(\,r\,)$ whose behaviour we analyse is:

$$\gamma(\,r\,) \;=\; \alpha \binom{r}{k-1} \left\{ \frac{1}{r} - \frac{1}{n} \right\} - 1 + \frac{\alpha}{nr} \binom{r}{k} + \frac{\log_e r}{\log_e n} - \frac{1}{\log_e n} \tag{3.7}$$

Proof of (a1): We wish to show that, $\forall\, k \geq 2$, $\forall\, \varepsilon > 0$, $\exists\, N_\varepsilon$ such that with $\alpha = k - 1 + \varepsilon$, $\gamma(\,k - 1\,) > \varepsilon/2k$ whenever $n \geq N_\varepsilon$. We have,

$$\gamma(k-1) = \frac{\alpha}{k-1} - \left(1 + \frac{\alpha}{n} \right) + \frac{\log_e(k-1) - 1}{\log_e n}$$

$$\geq \frac{\varepsilon}{k-1} - \frac{k - 1 + \varepsilon}{n} - \frac{1}{\log_e n} \;>\; \frac{\varepsilon}{2k} \qquad \text{when } n \text{ is large enough}$$

Proof of (a2): We first deal with the case, $k \geq 3$. Consider the value of $\gamma(\,r + 1\,) - \gamma(\,r\,)$. From (3.7) this reduces, after basic algebraic manipulations, to

$$\frac{\alpha}{r(r + 2 - k)} \binom{r}{k-1} \left(k - 2 - \frac{r(k-1)}{n} \right) + \frac{\alpha}{nr(r + 1 - k)} \binom{r}{k} (k - 1) + \frac{\log_e(1 + 1/r)}{\log_e n} \tag{3.8}$$

For $k \geq 3$ and $k - 1 \leq r \leq (k - 2)n/(k - 1)$, the term

$$\frac{\alpha}{r(r + 2 - k)} \binom{r}{k-1} \left(k - 2 - \frac{r(k-1)}{n} \right)$$

is always non-negative. Since the other terms in (3.8) are positive this establishes (a2) for $k \geq 3$.

For the case $k = 2$, we have

$$\gamma(\,r\,) \;=\; \alpha - \left(1 + \frac{\alpha r}{n} \right\} + \frac{\alpha(r-1)}{2n} + \frac{\log_e r}{\log_e n} - \frac{1}{\log_e n}$$

Substituting $\alpha = 1 + \varepsilon$, this becomes

$$\gamma(\,r\,) \;=\; \varepsilon - \frac{(1+\varepsilon)r}{2n} - \frac{1+\varepsilon}{2n} + \frac{\log_e r}{\log_e n} - \frac{1}{\log_e n}$$

Differentiating $\gamma(\,r\,)$ with respect to r gives

$$\frac{d\gamma}{dr} = -\frac{1+\varepsilon}{2n} + \frac{1}{r \log_e n}$$

From which we can deduce that $\gamma(\,r\,)$ attains its maximal value at $r = \dfrac{2n}{(1+\varepsilon)\log_e n}$, as claimed.

1) Strictly speaking, there is no need to consider the cases $k = 2$ and $k \geq 3$ separately, since, in principle, the method used for the former could be adapted to deal with all k. Such an approach, however, introduces further complications to the analysis due to the form of the binomial coefficients in the definition of $\gamma(r)$ when $k \geq 3$. For this reason a more direct approach is used to deal with $k \geq 3$, even though this has the disadvantage of dividing the proof into two distinct sections.

Proof of (a3): For the case $k = 2$, (a2) has established that $\gamma(r) < \gamma(r+1)$ when $1 \le r < 2n/((1+\varepsilon)\log_e n)$, and the value of $\gamma(r)$ decreases subsequently. Given that (a1) establishes $\gamma(1) > \varepsilon/4$ when n is sufficiently large and that the maximum value of r is n, it therefore suffices to show $\gamma(n) > \varepsilon/4$. From (3.7)

$$\gamma(n) = \varepsilon - \frac{1+\varepsilon}{2} - \frac{1+\varepsilon}{2n} + 1 - \frac{1}{\log_e n}$$

$$= \frac{1+\varepsilon}{2} - \frac{1+\varepsilon}{2n} - \frac{1}{\log_e n} > \frac{\varepsilon}{4} \qquad \text{for } n \text{ large enough}$$

For $k \ge 3$, and $\frac{(k-2)n}{(k-1)} \le r \le n$, From (3.7) we have,

$$\gamma(r) = \alpha \binom{r}{k-1} \left\{ \frac{1}{r} - \frac{1}{n} \right\} - 1 + \frac{\alpha}{nr} \binom{r}{k} + \frac{\log_e r}{\log_e n} - \frac{1}{\log_e n}$$

This, under the conditions stated is,

$$\ge \frac{\alpha}{n^2} \binom{(k-2)n/(k-1)}{3} + \frac{\log_e((k-2)/(k-1)) + \log_e n - 1}{\log_e n} - 1$$

$$= \Omega(n)$$

which can be made arbitrarily large.

In summary it has just been proved that for all $\varepsilon > 0$ and n sufficiently large, if we fix $\alpha = k - 1 + \varepsilon$, then $\forall\, k \ge 2$, $\forall\, r : k-1 \le r \le n$, it holds that

$$\gamma(r) > \frac{\varepsilon}{2k} \tag{3.9}$$

Hence, from (3.6),

$$E[\,|Acc(\phi)|\,] \le \sum_{r=k-1}^{n} \exp(-r \log_e n\, \gamma(r))$$

$$< \sum_{r=k-1}^{n} \exp(-(r\varepsilon/2k)\log_e n) = \sum_{r=k-1}^{n} n^{-\delta r}$$

(where $\delta = \varepsilon/2k$)

$$= \frac{1 - n^{-\delta(n+1)}}{1 - n^{-\delta}} - \frac{1 - n^{-\delta(k-1)}}{1 - n^{-\delta}}$$

$$= \frac{n^{-\delta(k+1)} - n^{-\delta(n+1)}}{1 - n^{-\delta}}$$

and since,

$$\lim_{n \to \infty} \frac{n^{-\delta(k+1)} - n^{-\delta(n+1)}}{1 - n^{-\delta}} = 0$$

we have completed the proof that $\forall\, k \ge 2$, $\forall\, \varepsilon > 0$, if $\alpha = k - 1 + \varepsilon$ then

$$\lim_{n \to \infty} P[\, F_{n,k,\alpha p} \textit{ is useful} \,] \le \lim_{n \to \infty} E[\,|Acc(F_{n,k,\alpha p})|\,] = 0$$

For part (b), we have a simpler analysis by considering only condition Q2 with subsets of size $k-1$

Hence,

$$P[\,F_{n,k,\alpha p}\ \textit{is useful}\,] \;\geq\; P[\,\exists\, B : |B| = k-1 \ \textit{and}\ \forall\, z \in \mathbf{X_n} - B\ (\,z \vee \bigvee_{x \in B} \bar{x}\,) \notin F\,]$$

$$= \; 1 - P[\,\forall\, B : |B| = k-1,\ \exists\, z\,(\,z \vee \bigvee_{x \in B} \bar{x}\,) \in F\,]$$

$$= \; 1 - (\,P[\,\exists\, z\,(\,z \vee \bar{x}_1 \vee \cdots \vee \bar{x}_{k-1}\,) \in F\,]\,)^{\binom{n}{k-1}}$$

(since the probabilities of individual clauses being chosen are independent).

$$\geq \; 1 - (\,1 - (\,1 - \alpha p\,)^{n-k+1}\,)^{\binom{n}{k-1}}$$

$$\geq \; 1 - \exp\left(-\binom{n}{k-1}(\,1 - \alpha p\,)^{n-k+1} \right)$$

Substituting $p = (\,\log_e n\,)/n$ and noting that

$$\lim_{n \to \infty} \left(1 - \frac{\alpha \log_e n}{n} \right)^{n-k+1} = n^{-\alpha} \quad ; \quad \binom{n}{k-1} \geq \frac{n^{k-1}}{(k-1)!} - o(\,n^{k-1}\,)$$

we see that the probability of being useful is bounded below by

$$1 - \frac{1}{\exp(\,n^{k-1-\alpha}/(k-1)!\,)} - o(1)$$

hence if $\alpha < k-1$ this probability approaches 1 as n approaches infinity. □

Corollary 3.1: Let $\theta_k = \dfrac{(k-1)(k+1)}{k!}$.

$$\forall\, \alpha > \theta_k \quad \lim_{n \to \infty} P[\,G_{n,k,\alpha n^{k-1}\log_e n}\ \textit{is useful}\,] = 0$$
$$\forall\, \alpha < \theta_k \quad \lim_{n \to \infty} P[\,G_{n,k,\alpha n^{k-1}\log_e n}\ \textit{is useful}\,] = 1$$

Proof: Recall that $|HC(n,k)| = (k+1)\binom{n}{k} \approx \dfrac{k+1}{k!}\, n^k$, and that Theorem 3.1. establishes for $\alpha > k-1$, and $p = \dfrac{\alpha \log_e n}{n}$ $P[\,F_{n,k,p}\ \textit{useful}\,] \to 0$; for $\alpha < k-1$ and $p = \dfrac{\alpha \log_e n}{n}$, $P[\,F_{n,k,p}\ \textit{useful}\,] \to 1$. Combining these with Fact 3.1., proves the result. □

4. Conclusions

In this paper a variation of *k-SAT* — *k-HSAT* — which is motivated by a natural application in Logic Programming has been considered with respect to its phase transition behaviour. Through the use of purely analytic methods a complete characterisation of the phase transition threshold, for all constant $k \geq 2$ has been obtained.

5. References

[1] D.Angluin and L.G. Valiant. Fast probabilistic algorithms for Hamiltonian Circuits and Matchings. *Journal of Computer and System Sciences*, 18, (1979), 82-93

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

[3] V. Chvátal and B. Reed. Mick gets some (the odds are on his side); *Proc. 33rd IEEE FOCS*, (1992), 620-627

[4] V. Chvátal and E. Szemerédi. Many hard examples for resolution; *Jnl. of the ACM*, 35, (1988), 759-768

[5] J.M. Crawford and L.D. Auton. Experimental results on the crossover point in random 3-SAT, *Artificial Intelligence* 81(1-2), (1996), 31-57.

[6] E. Friedgut and G. Kalai. Every monotone graph property has a sharp threshold. *Proc. Amer. Mathl. Soc.*, 124, (1996), 2993-3002

[7] A.M. Frieze and S. Suen. Analysis of simple heuristics for random instances of 3-SAT; Manuscript, 1993

[8] A. Goerdt. A threshold for unsatisfiability, *Proc. 17th Symposium on Mathematical Foundations of Computer Science*, (Eds: I. Havel, V. Krbec), (1992), Springer-Verlag LNCS Vol. 629, 264-274

[9] G. Istrate and M. Ogihara. The phase transition in random Horn satisfiability. (to appear: *Proc. 4th Conf. on A.I. and Mathematics*, 1998)

[10] L.M. Kirousis, E.Kranakis, and D. Krizanc. A better upper bound for the unsatisfiability threshold. *Proc. DIMACS Workshop Satisfiability Problem: Theory and Applications*, (March 1996)

[11] L.M. Kirousis, E.Kranakis, and D. Krizanc. Approximating the unsatisfiability threshold of random formulas. *Proc. 4th Annual European Symposium on Algorithms, ESA '96*, (September 1996)

[12] T. Larrabee and Y. Tsuji. Evidence of a satisfiability threshold for random 3CNF formulas. Technical Report UCSC-CRL-92-42, Univ. of California at Santa Cruz, (1992)

[13] B. Selman, D.G. Mitchell, H.J. Levesque, Generating hard satisfiability problems, *Artificial Intelligence*, 81(1-2), (1996) 17-29.

[14] R. Schrag and J.M. Crawford. Implicates and prime implicates in random 3-SAT, *Artificial Intelligence*, 81 (1996), 199-222