The Probability of Pure Literals

John W. Rosenthal, *Department of Mathematics and Computer Science, Ithaca College, Ithaca, NY 14850*

E-Mail: rosentha@ithaca.edu

J. M. Plotkin, *Department of Mathematics, Michigan State University, East Lansing, MI 48824*

E-Mail: plotkin@math.msu.edu

John Franco, *Department of Computer Science, University of Cincinatti, Cincinatti, OH 45221*

E-Mail: franco@franco.csm.uc.edu

Address all correspondence to J.M. Plotkin at the above address/e-mail.

Abstract

We describe an error in earlier probabilistic analyses of the pure literal heuristic as a procedure for solving k-SAT. All probabilistic analyses are in the constant degree model in which a random instance C of k-SAT consists of m clauses selected independently and uniformly (with replacement) from the set of all k-clauses over n variables. We provide a new analysis for $k = 2$. Specifically, we show with probability approaching 1 as m goes to ∞ one can apply the pure literal rule repeatedly to a random instance of 2-SAT until the number of clauses is "small" provided $n/m \geq \lambda > 1$. But if $n/m \leq \lambda < 1$, with probability approaching 1 if the pure literal rule is applied as much as possible, then at least $m^{1/5}$ clauses will remain.

Keywords: 2-SAT, constant degree model, Davis-Putnam Procedure, pure literal (heuristic), probability of a pure literal

1 Introduction

The satisfiability problem for sets of propositional clauses is the primordial NP-complete problem (See[14, p.38]). Also as shown in [14], the satisfiability problem for sets of k-clauses (k-SAT) is NP-complete for $k > 2$. The lack of polynomial time algorithms for NP-complete problems has given rise to the search for fast heuristic schemes for their solution which will be successful "almost surely". Franco [10] and the rest of us [18] give probabilistic analyses of the pure literal heuristic as a procedure for solving k-SAT. Unfortunately a serious error vitiates the outcomes. We describe the error in §2. For 2-SAT we provide a different analysis in §3 showing the efficiency of the pure literal heuristic in the cases claimed by Franco. This analysis is based partly on the Chvátal and Reed [7] and Goerdt [15] analysis of the probability that a set of 2-clauses is satisfiable. In §4 we show the optimal nature of the result in §3.

All of these analyses are in the constant degree model for formulae in conjunctive normal form (CNF) having a sequence of m clauses each of which is a k-clause, that is, has a set of k literals selected from a set V of n Boolean variables and their negations. In this model clauses are selected independently from one another with the uniform distribution on the set of all possible k-clauses, denoted $Q_k(V)$ or $Q_k(n)$. To allow changes in parameters Franco and Paull [12] denote this distribution as $f(m, n, k)$. We omit k as it is clear. This model has been used in many studies of k-SAT including Brown and Purdom [3], Purdom and Brown [19], Franco and Paull[12], Franco [10], Franco, Plotkin, and Rosenthal [13], M-T Chao and Franco [5] and [6], Chvátal and Reed [7], and Goerdt [15]. It has also been used in

the study of *k-Exact SAT* in Rosenthal, Speckenmeyer, and Kemp [22] and Rosenthal [21]. The reader should be cautioned that among these papers there is confusing variation in the symbols used for the number of variables, the number of clauses, and the number of literals per clause.

The pure literal heuristic is based on the pure literal rule which is part of the Davis-Putnam Procedure (DPP) [8], an algorithm for satisfiability (SAT) and k-SAT. Franco and Paull [12] provide a recent description. [12], [10], [13], [5], [6], etc. study aspects of DPP using the constant degree model. Goldberg [16], Goldberg, Purdom, and Brown [16], Purdom and Brown [20], Bugrara, Pan, and Purdom [4], Franco [11], etc. study aspects of DPP for SAT using another distribution, the constant density model. A lot of the work in the constant density model involves the pure literal rule. Recently Broder, Freize, and Upfal [2] used another model to study the pure literal rule for 3-clauses. In this model there are also m k-clauses, but the km literals are chosen uniformly and independently from the set of $2n$ available literals.

A pure literal for a formula in CNF is a literal l which occurs in at least one clause, but whose negation does not occur in any clause. The pure literal rule declares l to be true, and deletes all clauses containing l. The pure literal heuristic PL keeps trying to apply the pure literal rule until the sequence of remaining clauses is so small that the splitting rule (another one of DPP's rules) will efficiently determine satisfiability. If the pure literal rule can not be applied this often, the pure literal heuristic gives up. Franco [10] provides a more formal description of PL.

In [10] Franco examines the performance of PL applied to random (according to the constant degree model) formulae with m clauses from $Q_k(n)$ where $n = \lambda m$ and $\lambda > 1$. His intention is to show that asymptotically (in m) almost always PL can be applied until only $\log_2 m$ clauses remain. He does not analyze PL directly, but rather analyzes a procedure PL' whose probability of giving up is at least the probability PL gives up. In PL' if a pure literal occurs in $r > 1$ clauses, then one reintroduces $r - 1$ random k-clauses built using the variables not yet assigned a truth value. For the reader's convenience we paraphrase Franco's description of PL'.

In the following description A denotes the set of variables not yet assigned a truth value, D denotes the sequence of remaining k-clauses, and E denotes the set of k-clauses that are reintroduced. Initially $A = V$, D is the sequence of k-clauses from the original random formulae, and $E = \emptyset$

Procedure $PL'(D, A, h)$

(1) while $|D| \geq h$ do

(2) if there is no pure literal in D, then "give up"

(3) else begin

(4) choose a pure literal l in D;

(5) delete the variable of l from A;

(6) delete all clauses containing l from D;

(7) if r clauses were removed in line (6), select $r - 1$ clauses independently and uniformly from $Q_k(A)$, and adjoin these to both D and E.

end; (8) return "satisfiable" or "unsatisfiable" as determined by the splitting rule applied to $D - E$;

end PL'

2 The Error

Lemma 1 of [10] is the calculation of the probability that PL' does not give up on an $f(m,n)$-random instance (where $n > m$). It is claimed that after the i^{th} iteration of the while loop in PL' the sequence D is an $f(m-i,n-i)$-random instance, i.e. that the clauses of D are independently and uniformly selected from $Q_k(A)$. We show this is false. A similar error occurs in [17] and is essentially a sophisticated version of Bertrand's paradox.

To show that the claim is false we must first clarify line (4) of PL'. The choice of a pure literal in D may be made algorithmically or randomly (viewing PL' as a randomized algorithm). In the random case for every occurrence of line (4) one assigns probabilities (or "weights") to the possible choices of pure literal. We call any such assignment a *weighting scheme.* For example, we could assign each pure literal equal weight. The algorithmic case can be viewed as a special instance of the random case in which we use a weighting scheme that always assigns one pure literal weight 1 and the others weight 0.

We show that for no weighting scheme does one retain independence after the completion of one pass through the while loop of PL'. For ease of exposition we do this for $m = 4, n = 6, k = 3$.

We represent a clause by listing its variables and beneath each variable placing $+$ or $-$ to indicate a positive or negative occurrence of that variable.

Let $A = \{v_i \mid 1 \leq i \leq 6\}$ and $e_j \in \{+,-\}$ for $1 \leq j \leq 3$. For each sequence D consisting of four clauses from $Q_3(A)$, let $W_D(e_1,e_2,e_3)$ be the probability that in the first execution of the while loop on line (1) of $PL'(D,A,h)$ the pure literal chosen on line (4) was v_4, v_5, or v_6 and that the new D arising

at the end of line (7) is

$$\left\{\begin{array}{cccccccccccc} v_1 & v_2 & v_3 & ; & v_1 & v_2 & v_3 & ; & * & * & * \\ + & + & + & & e_1 & e_2 & e_3 & & * & * & * \end{array}\right\}$$

where $*$'s indicate arbitrary entries.

Let $\sigma(e_1, e_2, e_3) = \sum_{|D|=4} W_D(e_1, e_2, e_3)$. The assumption that the result of one pass through the while loop is an $f(3,5)$-random instance implies $\sigma(e_1, e_2, e_3)$ is independent of the values of e_1, e_2, e_3. Let us examine what would cause an imbalance between $\sigma_+ = \sigma(+,+,+)$ and $\sigma_- = \sigma(-,-,-)$. We need only consider cases where v_4, v_5, or v_6 is pure. If either or both of the v_1, v_2, v_3-clauses listed are obtained via E (on line (7)) then these two v_1, v_2, v_3-clauses are obtained independently and so no imbalance could occur. So we can restrict our attention to the remaining case where the two v_1, v_2, v_3-clauses listed came from the original D and v_4, v_5, or v_6 is pure. We call such a D a potential threatening contributor to σ_+ or σ_-, respectively. σ_+ and σ_- have corresponding potential threatening contributors. The only difference is that in such contributors to σ_- v_1, v_2, and v_3 are not pure and hence

$$\text{Prob(pure chosen is in } \{v_4, v_5, v_6\}) = 1$$

whereas in such contributors to σ_+, v_1, v_2, and/or v_3 may be pure. Thus to guarantee that σ_+ is as large as σ_- the weighting scheme when handling potential threatening contributors to σ_+ must give no weight to any pure literal in $\{v_1, v_2, v_3\}$. In particular if the original D is

$$\left\{\begin{array}{cccccccccccccccc} v_1 & v_2 & v_3 & ; & v_1 & v_2 & v_3 & ; & v_4 & v_5 & v_6 & ; & v_4 & v_5 & v_6 \\ + & + & + & & + & + & + & & + & + & + & & + & + & + \end{array}\right\}$$

then v_1, v_2, and v_3 must not be selected. Mutatis mutandi v_4, v_5, and v_6 must not be selected. We have arrived at a contradiction.

It is worth noting that a similar argument shows that PL' never decreases the probability of impurity. More precisely, let V be a set of n Boolean variables, V' a set of $n - s$ Boolean variables (where $s > 0$), C a sequence of m clauses selected independently and uniformly from $Q_k(V)$, C' a sequence of $m - s$ clauses selected independently and uniformly from $Q_k(V')$. Let $t \geq 0$. Let E' be the event that after t complete passes through the while loop of line (1) of $PL'(C', V', h)$ there is a pure literal and let E be the event that after $s + t$ complete passes through the while loop of line (1) of $PL'(C, V, h)$ there is a pure literal. Then Prob(E)$\leq$ Prob(E').

3 Pure Literal Rule for 2-SAT

Using a different analysis we now show that the pure literal heuristic succeeds when $k = 2$ provided the ratio of the number of variables over the number of clauses is asymptotically greater than 1.

THEOREM 3.1

Let $\lambda > 1$. Assume $n \geq \lambda m$. Let C be a random sequence of m clauses from $Q_2(n)$. Then

Prob(the pure literal rule may be applied to C until at most $\log_2(n)$ clauses remain)$= 1 - o(1)$.

NOTES: n is a function of m.

Throughout §3 and §4 all asymptotic notations are asymptotic in m.

REMARK 3.2

The proof, especially the Configuration Lemma below, is based on the Chvátal-Reed [7] proof that under the same hypotheses

$$\text{Prob}(C \text{ is satisfiable}) = 1 - o(1).$$

More generally we show:

THEOREM 3.3

Same hypotheses as Theorem 3.1. If $1 = o(t)$, then

$$\text{Prob(the pure literal rule may be applied to C until at most } t \text{ clauses remain)} = 1 - o(1).$$

NOTE: t is a function of m.

PROOF. For notational simplicity we assume λm is an integer. Adding clauses to C only makes it harder to reduce to t clauses. So we may as well assume $n = \lambda m$.

DEFINITION 3.4 (Pure literal block)

A *pure literal block* (PL block) is a sequence of clauses with no pure literals, that is, every variable which occurs does so both positively and negatively.

To prove Theorem 3.3 it suffices to prove

$$\text{Prob(there is a PL block of size at least } t) = o(1).$$

It is well known that 2-SAT has fast algorithms (see e.g. Aspvall, Plass, and Tarjan[1]). Underlying this result is the observation that a 2-clause $l_1 \vee l_2$ may be viewed as the implications $\neg l_1 \rightarrow l_2$ and/or $\neg l_2 \rightarrow l_1$. Hence, a sequence of 2-clauses may be viewed as a directed graph on the set of literals. From this viewpoint a PL block B is a sequence of directed edges so that

every vertex of B is both the initial vertex of a directed edge and the terminal vertex of a directed edge.

DEFINITION 3.5 (Cycle)

A *cycle* is a set $l_1 \to l_2, l_2 \to l_3, \cdots, l_s \to l_1$ of directed edges. Henceforth, we write $l_1 \to l_2 \to \cdots \to l_s \to l_1$.

As it has only finitely many vertices any PL block includes a cycle. More generally any vertex of a PL block is either on a cycle or lies between two cycles.

PROPOSITION 3.6 (CONFIGURATION LEMMA)

Same hypotheses as Theorem 3.1.

a) Prob(C has two cycles connected by a path)= $o(1)$.

b) Prob(C has two directly connected cycles)= $o(1)$.

c) Prob(C has two overlapping cycles)= $o(1)$.

PROOF.

a) Two cycles connected by a path consist of

$l_1 \to l_2 \to \cdots \to l_{\rho_1} \to l_1$, the first cycle

$m_1 \to m_2 \to \cdots \to m_{\rho_2} \to m_1$, the second cycle

and

$l_1 \to n_2 \to \cdots \to n_{\rho_3} \to m_1$, the connecting path.

b) Two directly connected cycles consist of

$l_1 \to l_2 \to \cdots \to l_{\rho_1} \to l_1$, one cycle

and

$l_1 \to m_2 \to \cdots \to m_{\rho_2} \to l_1$, the other cycle.

c) Two overlapping cycles consist of

$l_1 \to l_2 \to \cdots \to l_{\rho_1} \to \cdots \to l_{\rho_1+\rho_2} \to l_1$, one cycle

and

$l_1 \to m_2 \to \cdots \to m_{\rho_3} \to l_{\rho_1}$, the nonoverlapping part of the other cycle.

The proofs for cases b) and c) are omitted as they are similar to the proof for case a). The crucial point in each case is that the number of literals used is one less than the number of clauses used.

For the moment fix ρ_1, ρ_2, and ρ_3. In case a) one must choose $\rho = \rho_1 + \rho_2 + \rho_3$ clauses and $\rho - 1$ literals for these cycles and their connecting path. The literals may be chosen in at most $(2n)^{\rho-1}$ ways and the clauses may be chosen in at most m^ρ ways. Thus, the probability of this configuration for fixed ρ_1, ρ_2, and ρ_3 is at most

$$\begin{aligned} \frac{(2n)^{\rho-1}m^\rho}{\left(2^2\binom{n}{2}\right)^\rho} &= \frac{1}{2n}\frac{1}{\lambda^\rho}\left(\frac{n}{n-1}\right)^\rho \\ &\leq \frac{4}{2n}\frac{1}{\lambda^\rho} \quad (\text{as } \rho < n) \end{aligned}$$

So the probability of this configuration for some ρ_1, ρ_2, and ρ_3 with $\rho = \rho_1 + \rho_2 + \rho_3$ is at most $\frac{4}{2n}\frac{\rho^2}{\lambda^\rho}$. Hence, the probability of this configuration for some ρ_1, ρ_2, and ρ_3 is at most

$$\frac{4}{2n}\sum_{\rho \geq 1}\frac{\rho^2}{\lambda^\rho}$$

By the ratio test $\sum_{\rho \geq 1}\frac{\rho^2}{\lambda^\rho}$ is finite and so the desired probability is

$$O\left(\frac{1}{2n}\right) = o(1).$$

The Configuration Lemma tells that by ignoring a $o(1)$ piece of the sample space of formulae, we may assume that no cycles in a formula are connected to one another. Thus, the proof of Theorem 3.3 is completed by showing :

PROPOSITION 3.7

Same hypotheses as Theorem 3.1.

$$\text{Prob(the sum of the lengths of all cycles in } C \text{ is at least } t \text{ and}$$
$$\text{no cycles of } C \text{ are connected to one another)} = o(1).$$

PROOF. Say the sum of the lengths is $t' \geq t$.

Say there are n_i cycles of length t_i for $i = 1, \ldots, s$ with total length $t' = \sum_{i=1}^{s} n_i t_i$.

As $l \rightarrow l$ is not a clause, each $t_i \geq 2$.

One must choose t' clauses and t' literals for these cycles. The literals may be chosen in at most $(2n)^{t'}$ ways and the clauses may be chosen in at most $m^{t'}$ ways. But as any cycle of length t_i may be rotated in t_i ways, and the n_i cycles of length t_i may be permuted in $n_i!$ ways, the number of ways to choose literals was overcounted by a factor of $\prod_{i=1}^{s} {t_i}^{n_i} n_i!$.

So the probability of n_i cycles of length t_i for $i = 1, \ldots, s$ is at most

$$\frac{1}{\prod_{i=1}^{s} {t_i}^{n_i} n_i!} \frac{(2n)^{t'} m^{t'}}{\left(2^2 \binom{n}{2}\right)^{t'}} \leq \frac{4}{\prod_{i=1}^{s} {t_i}^{n_i} n_i! \lambda^{t'}}$$

It is well known that in the uniform distribution on the set of permutations on t' letters

$$\frac{1}{\prod_{i=1}^{s} {t_i}^{n_i} n_i!}$$

is the probability that a permutation on t' letters has precisely n_i cycles of length t_i for $i = 1, \ldots, s$.

So the probability that the sum of lengths of all cycles is t' is at most

$$\text{Prob(a permutation on } t' \text{ letters has no fixed point)}\frac{4}{\lambda^{t'}} \leq \frac{4}{\lambda^{t'}}.$$

So the probability the sum of the lengths of all cycles of C is at least t and no cycles of C are connected is at most

$$4\sum_{t' \geq t} \frac{1}{\lambda^{t'}} = O\left(\frac{1}{\lambda^t}\right).$$

So, as $1 = o(t)$, this probability is $o(1)$.

Given the phenomenal success of the pure literal heuristic under the hypotheses of Theorem 3.3 it is reasonable to ask if under the same hypotheses the pure literal rule asymptotically almost always eliminates all clauses. This is false as

PROPOSITION 3.8

If $n \sim \lambda m$ where $\lambda > 1$, then the probability there is a cycle of length 2 is asymptotically > 0.

PROOF. Let p_2 be the probability there is no cycle of length 2. We show p_2 is asymptotic to $\exp\left(-\frac{1}{4\lambda^2}\right)$.

The main source of difficulty in computing this probability is that clauses may occur more than once. So we write p_2 as:

$$\sum_{t \geq 0} \text{Prob(there are } t \text{ repetitions of clauses and no cycle of length 2)}$$

.

Pick t' so that $1 = o(t')$ and $t' = o(m^{1/2})$.

CLAIM 3.9

$$\text{Prob(there are at least } t' \text{ repetitions)} = o(1).$$

PROOF OF CLAIM 3.9. Say we get exactly t repetitions as we list the m clauses. Say these occur as the $i_1^{\text{st}}, i_2^{\text{nd}}, \ldots, i_t^{\text{th}}$ clauses where $i_1 < i_2 < \ldots < i_t$. Then there are $i_1 - 1$ choices for the first repetition, $i_2 - 2$ choices for the second repetition,..., and $i_t - t$ choices for the t^{th} repetition. And there are at most $(2N)^{m-t}$ choices for the other $m - t$ clauses (where $N = n(n-1)$). Thus, the probability of exactly t repetitions is at most $\frac{s_{t,m}}{(2N)^t}$ where

$$s_{t,m} = \sum_{1<i_1<i_2<\cdots<i_t\leq m} (i_1 - 1)(i_2 - 2)\cdots(i_t - t).$$

By induction we have:

LEMMA 3.10

$$\frac{(m-t)^{2t}}{2^t t!} \leq s_{t,m} \leq \frac{m^{2t}}{2^t t!}.$$

So the probability of exactly t repetitions is at most $\frac{1}{t!}\left(\frac{m^2}{4N}\right)^t$. So the probability of at least t' repetitions is at most $\sum_{t\geq t'} \frac{1}{t!}\left(\frac{m^2}{4N}\right)^t$, which is $o(1)$ as $1 = o(t')$.

Thus, p_2 is asymptotically

$$\sum_{0\leq t<t'} \text{Prob(there are } t \text{ repetitions of clauses and no cycle of length 2).}$$

As no cycle of length 2 is produced as we list the m clauses if and only if each clause not repeating a previous clause avoids the converse of each previous clause, a similar argument shows

$$\text{Prob(there are } t \text{ repetitions of clauses and no 2 cycle)} = s_{t,m}\,\frac{N!}{2^t N^m (N-(m-t))!}.$$

By Lemma 3.10 as $t' = o(t^{1/2})$, $s_{t,m}$ is uniformly asymptotic to $\frac{m^{2t}}{2^t t!}$ for $t \leq t'$.

Thus, p_2 is asymptotic to

$$\sum_{0 \le t < t'} \frac{m^{2t} N!}{2^{2t} N^m t! (N - (m-t))!}.$$

By Stirling's formula and $\ln(1-x) = -x - \frac{x^2}{2} + O(x^3)$, this is asymptotic to

$$\sum_{0 \le t < t'} \frac{1}{t!} \left(\frac{m^2}{4N}\right)^t \exp\left(-\frac{m^2}{2N}\right).$$

By the power series expansion of $\exp(x)$, this is asymptotic to

$$\exp\left(\frac{m^2}{4N}\right) \exp\left(-\frac{m^2}{2N}\right),$$

which is asymptotic to $\exp\left(-\frac{1}{4\lambda^2}\right)$.

Proposition 3.8 is in striking contrast to Broder, Frieze, and Upfal's [2] results for the pure literal rule for 3-clauses. They show that in their model with asymptotically 1 probability there is no pure literal block provided $\frac{m}{n}$ is sufficiently small (< 1.63).

Proposition 3.8 and Erdös and Renyi's [9] results on the occurrences of cycles in random graphs suggest the following conjectures, assuming $\lambda_1 m < n < \lambda_2 m$ where the λ_i's are constants > 1.

CONJECTURE 3.11

For any fixed t the asymptotic probability there is a PL block of size t is strictly between 0 and 1.

CONJECTURE 3.12

The asymptotic probability there is a PL block of some size is strictly between 0 and 1. (By the Configuration Lemma, Conjecture 3.12 is equivalent to: The asymptotic probability there is a cycle of some length is strictly between 0 and 1.)

4 Optimality of Theorem 3.1

We show Theorem 3.1 is near optimal by showing if $n = \lambda m$ where $0 < \lambda < 1$, then with asymptotic probability 1 there is a PL block of size m^ε for some $\varepsilon > 0$. This follows immediately from the proof in Chvátal and Reed [7] that for such λ, Prob(C is satisfiable)$= o(1)$. They use the second moment method to show Prob(C has a "snake" of size m^ε)$= 1 - o(1)$ for any $\varepsilon < \frac{1}{8}$. It is trivial to observe a snake is a PL block.

A larger ε with a simpler proof may be obtained by using cycles instead of snakes.

THEOREM 4.1

Let $0 < \lambda < 1$. Assume $n \leq \lambda m$. Let $s = o(n^{1/4})$. Let C be a random sequence of m clauses from $Q_2(n)$. Then

Prob(the pure literal rule may be applied to C until fewer than s clauses remain)$= o(1)$.

NOTE: s is a function of m. It suffices to prove this theorem for $1 = o(s)$.

PROOF. As in Theorem 3.3 we may as well assume that λm is an integer and $n = \lambda m$.

DEFINITION 4.2 (s-cycle)

An *s-cycle* is a cycle $l_1 \to l_2 \to \cdots \to l_s \to l_1$ in which $l_1, \ldots, l_s$ have distinct variables.

We show Prob(C has an s-cycle)$= 1 - o(1)$.

For any s-cycle A let

$$\chi_A = \begin{cases} 1 & \text{if each clause of } A \text{ occurs exactly once in each } C \\ 0 & \text{otherwise} \end{cases}$$

Let $\chi = \sum\{\chi_A \mid A \text{ is an } s\text{-cycle}\}$. We show with probability $1 - o(1)$ some s-cycle occurs by showing $\text{Prob}(\chi > 0) = 1 - o(1)$. This is accomplished by the second moment method. That is, we show

$$E(\chi^2) = E(\chi)^2(1 + o(1)).$$

Then by Chebyshev's inequality,

$$\text{Prob}(|\chi - E(\chi)| \geq E(\chi)) \leq \frac{E(\chi^2) - E(\chi)^2}{E(\chi)^2} = o(1)$$

and so $\text{Prob}(\chi = 0) = o(1)$.

As in [7],

$$E(\chi_A) = \left(\frac{m}{2n^2}\right)^s (1 + o(1)) \text{ uniformly in } A.$$

There are $\binom{n}{s} s!2^s = (2n)^s(1 + o(1))$ ways to choose the vertices of an s-cycle. So

$$E(\chi) = \left(\frac{m}{n}\right)^s (1 + o(1)).$$

Also if A and B are s-cycles sharing exactly i edges, then

$$E(\chi_A\chi_B) = \left(\frac{m}{2n^2}\right)^{2s-i} (1 + o(1)) \text{ uniformly in } A, B, \text{ and } i.$$

Thus

$$E(\chi_A\chi_B) = \left(\frac{2n^2}{m}\right)^i E(\chi_A)E(\chi_B)(1 + o(1)) \text{ uniformly in } A, B, \text{ and } i.$$

Let $p_i(n)$ (usually written p_i) be the probability that a random s-cycle B shares i edges with a fixed s-cycle A.

So

$$E(\chi^2) = \sum_{i=0}^{s} p_i \left(\frac{2n^2}{m}\right)^i E(\chi^2)(1 + o(1)).$$

The main work involves obtaining estimates of the p_i. As there are $(2n)^s(1+o(1))$ s-cycles,

$$p_i = \frac{\text{number of } s\text{-cycles } B \text{ such that } A \text{ and } B \text{ share } i \text{ edges}}{(2n)^s(1+o(1))}.$$

Let $A \cap B$ denote the edges in both A and B. Let k be the number of connected components of $A \cap B$. Let $N(i,k)$ be the number of s-cycles B such that $A \cap B$ has i edges and k components.

So for $i > 0$

$$p_i = \frac{\sum_{k \geq 1} N(i,k)}{(2n)^s(1+o(1))}.$$

For $0 < i < s$, $N(i,k)$ may be overestimated by the product of

i) the number of ways of placing in A the k components of $A \cap B$

ii) the number of ways of placing in B the k components of $A \cap B$

iii) the number of ways of placing in B the edges for each component of $A \cap B$

and

iv) the number of ways of assigning literals to the vertices in B, but not in $A \cap B$.

i) and ii) may be overestimated as follows: Mark the beginning of each component with a $+$ and the end with a $-$. So we must choose positions for the $2k$ markers giving at most $\binom{s}{2k} \leq \frac{s^{2k}}{(2k)!}$ choices.

iii) is the number of ways of permuting same size components of $A \cap B$ in B. This is largest when all the components are the same size and, hence, is at most $k!$.

As $i < s$, $A \cap B$ has $i + k$ vertices and hence B has $s - (i + k)$ vertices not in $A \cap B$. So iv) is at most $(2n)^{s-(i+k)}$. Thus,

$$N(i,k) \leq \left(\frac{s^{2k}}{(2k)!}\right)^2 k!(2n)^{s-(i+k)} \leq \frac{s^{4k}(2n)^s}{(2n)^{i+k}}.$$

So

$$p_i \leq \frac{1}{(2n)^i} \sum_{k \geq 1} \frac{s^{4k}}{(2n)^k}(1 + o(1)) = \frac{1}{(2n)^i} \frac{s^4}{2n} \frac{1}{1 - \frac{s^4}{2n}}(1 + o(1)).$$

As $s = o(n^{1/4})$,

$$p_i \leq \frac{1}{(2n)^i} \frac{s^4}{2n}(1 + o(1)) \text{ uniformly in } i \text{ for } 0 < i < s.$$

If $i = s$, the only choice for B is which vertex of A is l_1^B.

So

$$p_s = \frac{s}{(2n)^s}(1 + o(1).$$

So

$$\sum_{i=1}^{s} p_i \leq \left(\frac{s}{(2n)^s} + \frac{s^4}{2n} \sum_{i \geq 1} \frac{1}{(2n)^i}\right)(1 + o(1)) = o(1).$$

So

$$p_0 = 1 - o(1).$$

So

$$E(\chi^2) - E(\chi)^2 = \sum_{i=1}^{s} p_i(n) \left(\frac{2n^2}{m}\right)^i E(\chi)^2(1 + o(1)).$$

By the above estimates of p_i,

$$E(\chi^2) - E(\chi)^2 \leq \frac{s^4}{2n} \sum_{i=1}^{s-1} \left(\frac{n}{m}\right)^i E(\chi)^2(1 + o(1)) + s\left(\frac{n}{m}\right)^s E(\chi)^2(1 + o(1)).$$

But $\frac{s^4}{2n} \sum_{i=1}^{s-1} \left(\frac{n}{m}\right)^i = o(1)$ as $\frac{n}{m} = \lambda < 1$ and $s = o(n^{1/4})$; and $s\left(\frac{n}{m}\right)^s = o(1)$ as $\frac{n}{m} = \lambda < 1$ and $1 = o(s)$.

QUESTION 4.3

How much larger of a pure literal block occurs asymptotically almost always for $n = \lambda m$ where $0 < \lambda < 1$?

5 Topics for further study

For 2-clauses we have seen there is an abrupt transition in the performance of the pure literal heuristic. It occurs for $n = \lambda m$ as λ switches from < 1 to > 1. It is striking that this is the same λ at which occurs the abrupt transition in 2-satisfiability shown by Chvátal and Reed in [7] and Goerdt in [15]. Is there a comparable transition in the performance of the pure literal heuristic for k-clauses for $k \geq 3$? First one should find an α (depending on k) such that for $n \geq \alpha m$, asymptotically there are almost never pure literal blocks of size at least $\log(m)$. And one should find a β (depending on k) such that for $n \leq \beta m$, asymptotically there are almost always pure literal blocks of size m^{ε} for some $\varepsilon > 0$. Next one should determine if as for $k = 2$, the inf of the possible α's = the sup of the possible β's. We are confident that unlike for 2-clauses such a transition would be larger than the conjectured transition for satisfiability.

References

[1] B. Aspvall, M.F. Plass, and R.E. Tarjan. A Linear-time algorithm for Testing the Truth of Certain Quantified Boolean Formulas. *Information Processing Letters*, **8**, 121-123, 1979.

[2] A.Z. Broder, A.M. Frieze, and E. Upfal. On the satisfiability and maximum satisfiability of random 3-CNF formulas. In *Proc. 4th Annual ACM-SIAM Symposium on Discrete Algorithms*, pp.322-330, ACM, New York 1993.

[3] C.A. Brown and P.W. Purdom. An Average Time Analysis of Backtracking. *SIAM J. Computing*, **10**, 583-593, 1981.

[4] K.M. Bugrara, Y. Pan, and P.W. Purdom. Exponential Average Time for the Pure Literal Rule. *SIAM J. Computing*, **18**, 409-418, 1989

[5] M-T Chao and J. Franco, Probabilistic Analysis of Two Heuristics for the 3-Satisfiability Problem. *SIAM J. Computing*, **15**, 1106-1118, 1986.

[6] M-T Chao and J. Franco. Probabilistic Analysis of a Generalization of the Unit Clause Literal Selection Heuristic for the k-Satisfiability Problem. *Information Sciences*, **51**, 289-314, 1990.

[7] V. Chvátal and B. Reed. Mick gets Some (The Odds are on his Side). In *Proc. 33rd Annual Symposium on the Foundation of Computer Science*, IEEE, pp. 620-627, 1992.

[8] M. Davis and H. Putnam. A Computing Procedure for Quantification Theory. *Journal of the Association for Computing Machinery*, **7**, 201-215, 1960.

[9] P. Erdös and A. Renyi. On the Evolution of Random Graphs. *Magyar Tud. Akad. Mat. Kutató Int. Közl.*, **5**, 17-61, 1960.

[10] J. Franco. Probabilistic Analysis of the Pure Literal Heuristic for the Satisfiability Problem. In *Annals of Operations Research I*, pp. 273-289, 1984.

[11] J. Franco. Elimination of Infrequent Variables Improves Average Case Performance of Satisfiability Algorithms. *SIAM J. Computing*, **20**, 1119-1127, 1991.

[12] J. Franco and M. Paull. Probabilistic Analysis of the Davis Putnam Procedure for Solving the Satisfiability Problem. *Discrete Applied Mathematics*, **5**, 77-87, 1983.

[13] J. Franco, J.M. Plotkin, and J.W. Rosenthal. Correction to Probabilistic Analysis of the Davis Putnam Procedure for Solving the Satisfiability Problem. *Discrete Applied Mathematics*, **17**, 295-299, 1987.

[14] M. Garey and D. Johnson. *Computers and Intractability, A Guide to the Theory of NP-Completeness.* W.H. Freeman, San Francisco, 1979.

[15] A. Goerdt. A threshold for unsatisfiability. In *Mathematical foundations of computer science 1992.* Vol. 692 *Lecture Notes In Computer Science*, pp. 264-274. Springer, Berlin, 1992.

[16] A. Goldberg. Average Case Complexity of the Satisfiability Problem. In *Proc. 4th Workshop on Automated Deduction*, pp. 1-6, Austin Texas, 1979.

[17] A. Goldberg, P. Purdom, and C. Brown. Average Time Analysis of Simplified Davis- Putnam Procedure. *Information Processing Letters*, **15**, 72-75, 1982.

[18] J.M. Plotkin and J.W. Rosenthal. Probabilistic Analysis of the Pure Literal for the Satisfiability Problem. *Abstracts of the AMS*, **6**, Number 3, p. 267, 1985.

[19] P.W. Purdom and C.A. Brown. An Analysis of Backtracking with Search Rearrangement. *SIAM J. Computing*, **12**, 717-733, 1983.

[20] P.W. Purdom and C.A. Brown. The Pure Literal Rule and Polynomial Average Time. *SIAM J. Computing*, **14**, 943-953, 1985.

[21] J.W. Rosenthal. Fine Transitions in the Size of a Search Tree for Exact Satisfiers, in preparation.

[22] J.W. Rosenthal, E. Speckenmeyer, and R. Kemp. Exact Satisfiability: A Natural Extension of Set Partition and its Average Case Behavior. In *Annals of Math. and Art. Intell.*, **6**, 185-200, 1992.