Analysis of Pure Literal Elimination Rule for Non-uniform Random (MAX) k-SAT Problem with an Arbitrary Degree Distribution

Oleksii Omelchenko, Andrei A. Bulatov

School of Computing Science,
Simon Fraser University, Canada
{oomelche, abulatov}@cs.sfu.ca

Abstract

MAX k-SAT is one of the archetypal NP-hard problems. Its variation called random MAX k-SAT problem was introduced in order to understand how hard it is to solve instances of the problem on average. The most common model to sample random instances is the uniform model, which has received a large amount of attention. However, the uniform model often fails to capture important structural properties we observe in the real-world instances. To address these limitations, a more general (in a certain sense) model has been proposed, the configuration model, which is able to produce instances with an arbitrary distribution of variables' degrees, and so can simulate biases in instances appearing in various applications. Our overall goal is to expand the theory built around the uniform model to the more general configuration model for a wide range of degree distributions. This includes locating satisfiability thresholds and analysing the performance of the standard heuristics applied to instances sampled from the configuration model.

In this paper we analyse the performance of the pure literal elimination rule. We provide an equation that given an underlying degree distribution gives the number of clauses the pure literal elimination rule satisfies w.h.p. We also show how the distribution of variable degrees changes over time as the algorithm is being executed.

Introduction

MAX SAT and its variant MAX k-SAT are known NP-hard problems even for $k = 2$. Hence, it is unlikely there exists an efficient algorithm, unless P=NP, and so we must rely on heuristics and approximation algorithms. Therefore, it is natural to ask what heuristics are good at solving *typical* SAT instances.

There are many ways to define which instances are typical. One of the approaches is to construct a random distribution of SAT formulas, and call formulas sampled from the distribution as representative or typical. By constructing appropriate distributions we may imitate formulas coming from different domains.

The most well-studied random model of k-CNF formulas is the *uniform* model $F_k(n, rn)$, which samples equiprobably formulas having n variables and rn k-clauses, where quantity $r > 0$ is called *density* (Achlioptas 2009). It turns out that many key properties of the model depend on the density. One such property is the *satisfiability threshold*, a critical density $\rho_k = \rho_k(n)$, which depends on k (and maybe on n), such that with high probability (w.h.p.) ϕ is satisfiable whenever $r \leq (1 - \epsilon)\rho_k(n)$, and unsatisfiable if $r \geq (1 + \epsilon)\rho_k(n)$ for $\epsilon > 0$. Finding the satisfiability threshold for different values of k has been a very active and fruitful research direction, see, e.g., (Achlioptas 2009; Ding, Sly, and Sun 2014; Mitchell, Selman, and Levesque 1992; Larrabee and Tsuji 1992; Chvatal and Reed 1992; Goerdt 1996; Achlioptas 2001; Achlioptas and Sorkin 2000; Achlioptas and Peres 2003).

A similar quantity related to MAX k-SAT, the optimization version of k-SAT, is the (expected) number of clauses in a formula from $F_k(n, rn)$ that can be simultaneously satisfied. Although this value is not known, a straightforward argument gives a lower bound of $(1 - 2^{-k})rn$ and in the case of MAX 2-SAT (Coppersmith et al. 2003) places this value between $(3/4r + 0.34\sqrt{r})n$ and $(3/4r + 0.5\sqrt{r})n$.

Lower bounds on the satisfiability thresholds and on the number of satisfiable clauses of MAX k-SAT instances are often proved by analyzing relatively simple heuristics. For example, the $(1 - 2^{-k})rn$ bound is obtained as the expected number of satisfiable clauses, and this simple technique can be derandomized using the *method of conditional expectations* (Erdös and Selfridge 1973). A number of more sophisticated heuristics of different types have been analyzed as well, see, e.g., (Broder, Frieze, and Upfal 1993; Luby, Mitzenmacher, and Shokrollahi 1998; Molloy 2005; Achlioptas and Sorkin 2000; Achlioptas 2001; Kaporis, Kirousis, and Lalas 2002, 2006; Hajiaghayi and Sorkin 2003; Frieze and Suen 1996; Alekhnovich and Ben-Sasson 2007), for a survey see (Achlioptas 2009). In this paper we focus on the pure literal elimination heuristic. Given a CNF ϕ this algorithm simply selects a variable that appears in ϕ in only one polarity, that is, either only negated or only unnegated, assigns it the value that satisfies all its occurrences, and removes the clauses that get satisfied. (Broder, Frieze, and Upfal 1993) (see also (Kim 2008)) analysed the pure literal elimination rule for solving uniform random 3-SAT and MAX 3-SAT. They obtained an expression, which shows how many clauses are satisfied by each iteration of the algorithm, and proved that the maximal number of clauses

which can be satisfied by an assignment of pure literals is concentrated around the function $T_n(r, k)$. This function is however hard to calculate when r is constant. A somewhat easier to use expressions for estimating the efficacy of the pure literal elimination rule for any $k \geq 3$ were given in (Mitzenmacher 1997; Molloy 2005).

Most of the results we have mentioned so far concern the uniform random instances. However, there is a whole line of research suggesting that in many aspects random uniform formulas do not resemble natural instances coming from the real-world problems. For example, many industrial CNF instances exhibit *scale-free* property (Ansótegui, Bonet, and Levy 2009), they seem to consist of *clusters* of tightly connected communities (Ansótegui et al. 2019), have a rather smaller *fractal dimension* (Ansótegui et al. 2014), and other structural features not present in random instances (Ansótegui et al. 2008; Beyersdorff and Kullmann 2014). Moreover, it is unlikely that there exists a universal algorithm performing well on all instances coming from different domains.

One way to tune the Random k-SAT model so that random instances resemble instances coming from specific domains is to use more general random model. Several such models have been suggested, see, (Ansótegui, Bonet, and Levy 2009; Ansótegui, Bonet, and Levy 2019; Ansótegui et al. 2019; Ansótegui et al. 2014; Friedrich et al. 2017). In this paper, we use a random model called *configuration model*, parametrized by a sequence of random variables. In order to generate a k-CNF with n variables we fix a random variable ξ_i with values in $\mathbb{N}$ distributed accordingly to a distribution $\mathbb{D}_i$ for each variable v_i. The random variable ξ_i represents the degree or the number of occurrences of v_i in the formula. Then the degree ξ_i of v_i is sampled from $\mathbb{D}_i$ and we create ξ_i *clones* of v_i; each clone is negated with probability 1/2. Finally, the set of clones is randomly partitioned into k-sets to form k-clauses.

Depending on the distributions of ξ_i the configuration model allows one to generate a wide variety of random k-CNFs. For example, if every ξ_i is a constant we are dealing with CNFs with a fixed degree sequence. This case has been studied in (Cooper, Frieze, and Sorkin 2007) for 2-SAT. If every $\mathbb{D}_i$ is the same Poisson distribution, the configuration model is also known as the Poisson Cloning model. (Kim 2004) proves that in many aspects the Poisson Cloning model is equivalent to the uniform Random (MAX) k-SAT. (Omelchenko and Bulatov 2021b) and (Omelchenko and Bulatov 2021a) study several aspects of the configuration model when the distributions are heavy-tailed.

The overall research goal is therefore to expand the results known for the uniform model to the configuration model. The main difficulty in this enterprise is that it is not always possible to use the same well established and efficient tools in both theoretical and practical aspects of the problem. On the theoretical side, when the distributions $\mathbb{D}_i$ are not as well behaved as Poisson, especially if they are heavy-tailed or do not have higher moments, the standard probability theory tools such as concentration inequalities, do not apply, and have to be replaced with more complicated and less efficient ones. For example, (Borovkov and Borovkov 2008) and (Omelchenko and Bulatov 2019) obtained some (relatively weak) concentration bounds for heavy-tailed distributions. On the practical side, weak concentration properties mean that in order to achieve a visible trend in experimental results one needs to handle formulas with billions of variables.

In this work we study how the pure literal elimination rule performs in the configuration model, and how the distribution of degrees of variables affects the number of clauses which the rule can satisfy. Our main contribution is Theorem 1. The theorem gives an equation that depends on the "averaged" distribution of degrees and only assumes that the random variables ξ_i have finite first moments. Solving the equation one can find how many clauses the pure elimination rule satisfies. Note that a similar result for the Poisson Cloning model was obtained by (Kim 2008). Our secondary result is Lemma 6, which shows how the degree distribution evolve as the pure literal heuristics transforms the instance. This information is useful for further analysis of k-CNF formulas obtained *after* they have been processed with the pure literal elimination heuristic.

Our analysis uses two important tools: the Weak Law of Large Numbers and the Wormald's differential equations method. The former one gives us some concentration property, while the latter is a standard tool in analysis of SAT algorithms, see (Achlioptas 2001; Achlioptas and Sorkin 2000).

The paper is organized as follows. After reminding the basic definitions and notation, we introduce the configuration model. The main part of the paper is the "Analysis of Pure Literal Elimination Algorithm" section, where we give a number of lemmas needed to obtain the main result of this work, Theorem 1. We conclude the paper with a number of experimental results which show how the pure literal elimination rule performs on different random instances obtained from the configuration model in reality. Due to space restrictions many proofs will be given in Supplemental materials.

Notations & Definitions

MAX SAT, MAX k-SAT, and Random MAX k-SAT. Let $x_1, \ldots, x_n$ be boolean variables, and n will always denote their number. A *literal* is either a variable or its negation. A literal is *negative*, when it is a negated variable, and *positive* otherwise. A *clause* is a disjunction of literals, and a k-clause is a clause with exactly k not necessarily distinct literals. Then a *CNF formula* is a set of clauses, while a *k-CNF formula* is a set of k-clauses. So, we treat every CNF formula as a set of clauses, and each clause is a set of k literals.

The *MAX SAT* problem is an optimization problem where given a CNF formula ϕ we need to determine an assignment that satisfies the maximal number of clauses in ϕ. The *MAX k-SAT* is a special case of MAX SAT, where the given formula ϕ is k-CNF.

Random MAX k-SAT problem is a variant of MAX k-SAT problem, where instead of an arbitrary k-CNF formula we are given a random k-CNF formula, sampled from some probabilistic distribution over k-CNF formulas with n vari-

ables. The distribution we use in this paper is configuration model (see the "Configuration Model" section for details).

Probability Reminder. We say that a sequence of events A_n happens *with high probability* (w.h.p.), when $\Pr[A_n] \to 1$ with $n \to \infty$. We use acronyms *r.v.* for "random variable", *r.vs.* for "random variables", and *u.a.r.* for "uniformly at random". We use $[n]$ to denote the set $\{1, 2, \ldots, n\}$.

The next lemma shows that the sum of n independent r.vs. $\xi_1, \xi_2, \ldots, \xi_n$ with finite expectations $\mathbb{E}|\xi_i| < \infty$ is concentrated around its mean, i.e. $\sum_{i=1}^{n} \xi_i = \sum_{i=1}^{n} \mathbb{E}\xi_i + o(n)$ w.h.p. This result is known as the *Weak Law of Large Numbers* (WLLN), however, its classical versions require the r.vs. ξ_i's to be either identically distributed and/or to have moments beyond expectation. Variation of the WLLN we need and state below does not ask for such premises, while its proof is only a minor modification of the now classical proof of Khintchin's Law of Large Numbers (Borovkov 2013).

Lemma 1 (WLLN)**.** *Let $\xi_1, \xi_2, \ldots, \xi_n$ be a collection of n independent r.vs. with finite expectation $\mathbb{E}|\xi_i| < \infty$. Then it holds w.h.p.*

$$\sum_{i=1}^{n} \xi_i = \sum_{i=1}^{n} \mathbb{E}\xi_i + o(n).$$

To analyse the evolution of algorithms, we rely on Wormald's *differential equations method* theorem (Wormald 1995, 1999). Some improvements to the method and its applications can be found in (Bohman 2009; Bohman and Keevash 2010; Warnke 2014, 2019). The main idea of the method is simple. Suppose you have a number of co-evolving in time scalar random processes $X_i(t)$, where $t \in \mathbb{N}$ and $i \in [\ell]$. Our goal is to describe how those processes evolve, and approximate their trajectory in time. The processes may affect each other in some intricate way during their evolution, which complicates their description. The theorem states that if the processes are "good", namely: (a) given state of all processes at any time t, the expected one time change of each process can be well-approximated (up to $o(1)$ factor) by some function $f_i\left(\frac{t}{n}, \frac{X_1(t)}{n}, \frac{X_2(t)}{n}, \ldots, \frac{X_\ell(t)}{n}\right)$; (b) it is highly unlikely that any process during its evolution will jump over $n^{1/2}$; and finally (c) functions f_i's satisfy Lipschitz condition, then we can be confident that the processes are concentrated around the deterministic trajectory, which can be obtained from a solution to the system of differential equations, for as long as all the above three conditions hold (see Supplemental Materials for a precise statement of the theorem).

Finally, we denote by $\xi \sim \mathbb{D}$ the fact that a r.v. ξ is distributed according to a probability distribution $\mathbb{D}$.

Configuration Model. *Configuration model* $\mathbb{C}_n^k((\xi_i)_{i=1}^n)$ is a probability distribution of k-CNF formulas with n variables. The model is parametrized by an ordered sequence of n independent random variables ξ_i with support on $\mathbb{N}^+$. We write $\phi \sim \mathbb{C}_n^k((\xi_i)_{i=1}^n)$, when formula ϕ is sampled from the configuration model. We do not require ξ_i's to be identically distributed as each $\xi_i \sim \mathbb{D}_i$ may come from its own distribution $\mathbb{D}_i$. But we do require them to be independent and with finite expectation $\mathbb{E}\xi_i < \infty$ for all $i \in [n]$.

Sampling formulas from $\mathbb{C}_n^k((\xi_i)_{i=1}^n)$ is usually done by constructing such formulas. By $\mathcal{V}(\phi)$ we denote the set of all variables in a sampled formula ϕ. Then $|\mathcal{V}(\phi)| = n$ always. The *degree* of a variable $v \in \mathcal{V}(\phi)$ is the number of times it appears in ϕ as an atom; each occurrence of v in ϕ we call a *clone* of v. Let $\mathcal{S}(\phi)$ be the set of all clones of all variables in ϕ. Each clone in $\mathcal{S}(\phi)$ can be positive or negative thus corresponding to a positive or negative literal. A literal in ϕ is said to be *pure literals*, if ϕ does not contain its negation.

A formula in the configuration model is generated in two steps. The first step, called CREATECLONES$((\xi_i)_{i=1}^n, k)$, starts with a sequence $(\xi_i)_{i=1}^n$ of random variables and creates clones of variables and assigns a sign to each of them. In order to do that for each $i \in [n]$ we sample degree d_i of each variable $v_i \in \mathcal{V}(\phi)$ from ξ_i and create d_i clones of v_i. If the total number of clones is not a multiple of k we discard these clones and start all over. Otherwise we assign a sign to each clone equiprobably. Let $\mathcal{S}$ denote the resulting set of (signed) clones. During the process we say that a clone $c \in \mathcal{S}(\phi)$ is *paired*, when there is a clause in ϕ, which contains this clone c; otherwise, the clone is said to be *unpaired*.

In the second step, called CONFIGURATIONMODEL-CNF$(\mathcal{S}, k)$, we randomly partition the set $\mathcal{S}$ into k-element subset that form clauses of the formula.

Note that as long as the second step results in a random partition of $\mathcal{S}$ into k-element subsets, it makes no difference how exactly partitioning is done. The most basic way is to pick k-element subsets from $\mathcal{S}$ until all the clones are put into some clause. However, often it is convenient to choose a more elaborate way, which may ease the analysis of the process. For example, we can use any rule (deterministic or probabilistic) to pick the very first clone for every clause, since in the end all clones must end up in some clauses. This observation was one of the key features used in (Cooper, Frieze, and Sorkin 2007; Omelchenko and Bulatov 2021b) to produce an alternative algorithm called TSPAN for constructing 2-CNF formulas. On the other hand, this freedom of picking the very first clone is coupled with the restriction that the other $k-1$ clones of every clause must be picked u.a.r. without replacement from the set of unpaired clones. When these two conditions are satisfied, we can be sure that the formulas produced by the alternative algorithms are equivalent to the formulas from $\mathbb{C}_n^k((\xi_i)_{i=1}^n)$.

To end this section, we introduce several quantities, which we frequently use in the subsequent analysis. First, let $S_n = |\mathcal{S}(\phi)|$ be the total number of clones in ϕ. Then, clearly, the number of clauses in ϕ is $|\phi| = \frac{S_n}{k}$. Let $\gamma := \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\xi_i$ denote the average degree of variables, and since all $\mathbb{E}\xi_i < \infty$, it follows that $\gamma < \infty$. Moreover, as the expectations of ξ_i's are finite, we have the following simple result:

Lemma 2. *It holds that $S_n = (1 + o(1))\,\gamma n$ w.h.p.*

Let ξ be a r.v. over $\mathbb{N}$ with probability distribution function

$$\Pr[\xi = d] = \frac{1}{n}\sum_{i=1}^{n} \Pr[\xi_i = d] \qquad \text{for all } d \in \mathbb{N}. \quad (1)$$

It turns out that many quantities in the configuration model can be expressed via this "averaged" r.v. ξ. For example, $\gamma = \mathbb{E}\xi$, and so $S_n = (1 + o(1))\, n\mathbb{E}\xi$ w.h.p. Next, we will abuse notation by using the same letter p with different subscripts to denote two different probabilities distinguishing them by the number of indices. Let $p_d := \Pr[\xi = d]$, while we use $p_{i,j}$ to denote the probability that a u.a.r. chosen variable produces i positive clones and j negative clones. The next lemma expresses $p_{i,j}$ in terms of p_{i+j}.

Lemma 3. *The probability a randomly chosen variable produces exactly i positive and j negative clones is*

$$p_{i,j} = 2^{-(i+j)} \binom{i+j}{i} \Pr[\xi = i + j] = 2^{-(i+j)} \binom{i+j}{i} p_{i+j}.$$

Finally, let $N_{i,j}$ be the number of *variables*, which produced exactly i positive and j negative clones. Then by Lemma 1 and Lemma 3, it readily follows that w.h.p.

$$N_{i,j} = (1 + o(1))\, np_{i,j}. \tag{2}$$

Analysis of Pure Literal Elimination

As was mentioned in the Introduction, our goal here is to analyse the efficiency of the pure literal elimination heuristic on random k-CNF instances sampled from the configuration model.

The main entities of interest in pure literal elimination rule are pure literals. As was mentioned in previous sections, we call a literal ℓ pure in a $k-$CNF formula ϕ if its complement literal $\bar{\ell}$ does not appear in ϕ. Hence, we can safely satisfy ℓ, and eliminate all clauses in which ℓ is present, thus, simplifying ϕ. We continue this elimination process as long as we have pure literals to satisfy. Observe, that during clause elimination it may happen that a literal, which was not initially pure, becomes pure if all occurrences of its complement literal happen to be in the eliminated clauses, which fuels the algorithm with new pure literals.

For the sake of analysis we slightly modify the algorithm (keeping the same outcome) by employing deferred decision. That is, instead of feeding the algorithm a complete k-CNF formula, we instead "reveal" only those clauses that contain pure literals, satisfy them, update the set of pure literals, and repeat this process as long as we have unpaired clones of pure literals. This way the unrevealed part of the formula remains random, and so probabilistic analysis can be applied.

The deferred decision version of the pure literal elimination algorithm is given in Algorithm 1. We start with creation of clones for all n variables by calling the CREATECLONES$((\xi_i)_{i=1}^n, k)$ procedure with $(\xi_i)_{i=1}^n$ being a sequence of r.vs. from which degrees are sampled. It may happen that some variables possess only clones of one sign (it happens w.h.p. when we have $p_d > 0$ for some constant d). Such variables form an initial pool of pure literals.

We generate formula ϕ iteratively, one clause at a time. We start with an empty formula. Next, as long as we have pure literals, we pick one of them u.a.r. without replacement. Observe that if we satisfy the chosen literal l, then every clause in which a clone of l is present is satisfied as well.

Algorithm 1: Pure literal elimination algorithm with deferred decision.

```
1:  function PURELITERAL((ξ_i)_{i=1}^n, k)
2:      φ_1 ← {};
3:      S ← CREATECLONES((ξ_i)_{i=1}^n, k);
4:      while there are clones in S with no complementary
        counterparts do
5:          c ← u.a.r. picked clone from the set of clones
            with no complementary counterparts in S;
6:          Satisfy the literal ℓ associated with c;
7:          C ← the set of all unpaired clones of ℓ;
8:          while C ≠ ∅ do
9:              c ← pick an arbitrary clone from C;
10:             cl ← sample k − 1 clones u.a.r. w/o replace-
                ment from S − {c};
11:             cl ← {c} ∪ cl;
12:             Mark cl as satisfied;
13:             φ_1 ← φ_1 ∪ {cl};
14:             C ← C − cl;
15:             S ← S − cl;
16:         end while
17:     end while
18:     return CONFIGURATIONMODELCNF(S, k);
19: end function
```

Hence, to know how many clauses we satisfy by satisfying l, we calculate how many clauses contain clones of l. To do so take all clones of the chosen pure literal l and distribute them in the following manner. Pick an arbitrary unpaired clone of l and form a new 1-clause out of it. Next, to finish the formation of the new clause, sample other $k - 1$ unpaired clones from $\mathcal{S}$ (line 1.10), and append them to the newly created 1-clause (line 1.11). We obtain a complete k-clause, which we add to the formula ϕ (line 1.13). Since we have paired clones in the new clause, they cannot appear in any other clauses, and so we remove them from the sets $\mathcal{C}$ and $\mathcal{S}$ (lines 1.14 and 1.15 respectively). We continue formation of new clauses containing clones of l as long as there remain unpaired clones of it. This generation process of clauses containing clones of l continues until we deplete all its unpaired clones. After that we pick next pure literal, if there exists any, and continue producing new clauses. Observe that the algorithm not just creates new clauses but actually satisfies them as well.

Just like in the original algorithm, it may happen that during elimination process some literals, which were not pure initially, become pure, when all clones of their complements get paired. Or, in other words, all their complement counterparts appear only in clauses, which we have already satisfied. In that case we add the newly-formed pure literals to the pool of all not-yet-paired pure literals, and continue the pairing process until we exhaust all the unpaired clones of pure literals. When all pure literals are used up we generate a k-CNF from the remaining clones in the usual way.

To analyse Algorithm 1, we need to introduce a few quantities that play major role in the analysis. The first one is the number of *variables* with i unpaired positive clones and j unpaired negative clones at time $t \in \mathbb{N}$, which we denote

by $N_{i,j}(t)$. Time t here measures how many clauses Algorithm 1 has formed, or, equivalently, the total number of times the inner loop (lines 8-16) has been executed. Hence, initially we have $t = 0$. As we form new clauses and t increases, $N_{i,j}(t)$ changes as well. If it happens that Algorithm 1 runs long enough for t to reach $|\phi| = \frac{S_n}{k}$, we conclude that the pure literal elimination rule was able to find a satisfying assignment. Also note that due to the random nature of the pairing process, $N_{i,j}(t)$ for all $i, j \geq 0$ are in fact random processes.

Note that $\sum_{i\geq 1}(N_{i,0}(t) + N_{0,i}(t))$ measures the number of pure literals at time t, since $N_{i,0}(t)$ and $N_{0,i}(t)$ count the number of variables with i unpaired clones all of which belong to a single literal. Hence, the algorithm runs as long as $\sum_{i\geq 1}(N_{i,0}(t) + N_{0,i}(t)) > 0$. Let t_0 denote the moment of time, when the algorithm stops, i.e. when the quantity $\sum_{i\geq 1}(N_{i,0}(t_0) + N_{0,i}(t_0))$ hits zero. Note that t_0 is a random variable, due to the random nature of dynamics of the processes $N_{i,j}(t)$. Clearly, to study how well the pure literal elimination heuristic performs on random instances from $\mathbb{C}_n^k((\xi_i)_{i=1}^n)$, we need to learn the distribution of t_0.

In order to apply the differential equations method to approximate dynamics of the processes $N_{i,j}(t)$, we must verify that the processes are "good". The next lemma proves that the processes $N_{i,j}(t)$ are indeed good candidates for approximation using the differential equations method.

Lemma 4. *Let $H(t) := \bigcup_{0\leq t'\leq t}\bigcup_{i,j\geq 1}\{N_{i,j}(t')\}$ be the complete history of the evolution of the processes $N_{i,j}(t')$ up to and including time t. Then for all $i, j \geq 1$ and all $0 < t < t_0$, it holds:*

1. $\mathbb{E}\left[N_{i,j}(t+1) - N_{i,j}(t) \mid H(t)\right]$
$$= f_{i,j}\left(\frac{t}{n}, \frac{N_{i,j}(t)}{n}, \frac{N_{i+1,j}(t)}{n}, \frac{N_{i,j+1}(t)}{n}\right) + o(1),$$
where
$$f_{i,j}\left(\tau, n_{i,j}(\tau), n_{i+1,j}(\tau), n_{i,j+1}(\tau)\right) = -\frac{k-1}{\gamma - k\tau}(i+j)n_{i,j}(\tau) + \frac{k-1}{\gamma - k\tau}\Big[(i+1)n_{i+1,j}(\tau) + (j+1)n_{i,j+1}(\tau)\Big];$$
2. $\Pr\left[\left|N_{i,j}(t+1) - N_{i,j}(t)\right| > k \mid H(t)\right] = 0$;
3. $f_{i,j}\left(\tau, n_{i,j}(\tau), n_{i+1,j}(\tau), n_{i,j+1}(\tau)\right)$ *is continuous and satisfies Lipschitz condition for* $0 \leq \tau \leq \frac{t_0}{n} < \frac{\gamma}{k}$.

Remark. *As it follows from the lemma, we use $n_{i,j}(\tau) = \frac{N_{i,j}(\tau n)}{n}$ with $\tau = \frac{t}{n}$, to denote the scaled number of variables with i positive and j negative unpaired clones at the scaled time τ.*

Hence, now we construct the system of differential equations describing the dynamics of the scaled number of variables $n_{i,j}(\tau)$ for all $i, j \geq 1$

$$\begin{aligned}\frac{d\, n_{i,j}}{d\tau} &= f_{i,j}\left(\tau, n_{i,j}(\tau), n_{i+1,j}(\tau), n_{i,j+1}(\tau)\right)\\ &= -\frac{k-1}{\gamma - k\tau}(i+j)n_{i,j}(\tau)\\ &+ \frac{k-1}{\gamma - k\tau}\Big[(i+1)n_{i+1,j}(\tau) + (j+1)n_{i,j+1}(\tau)\Big], \qquad (3)\end{aligned}$$

with initial values $n_{i,j}(0) = \frac{N_{i,j}(0)}{n} = 2^{-(i+j)}\binom{i+j}{i}p_{i+j}$. which follows from (2) after plugging in the initial number of variables $N_{i,j}$. The system has a single solution.

Lemma 5. *The system* (3) *defined for all $i, j \geq 1$ together with initial values $n_{i,j}(0) = 2^{-(i+j)}\binom{i+j}{i}p_{i+j}$ has a unique solution*

$$n_{i,j}(\tau) = 2^{-(i+j)}\binom{i+j}{i} \times \sum_{\ell\geq 1}\binom{\ell}{i+j} z(\tau)^{i+j}\left(1 - z(\tau)\right)^{\ell-(i+j)},$$

where

$$z(\tau) = \left(1 - \frac{k\tau}{\gamma}\right)^{1-\frac{1}{k}}.$$

Now, we apply differential equations method to approximate dynamics of the processes $N_{i,j}(t)$.

Lemma 6. *Let $N_{i,j}(t)$ be the number of variables with i positive and j negative unpaired clones at time t, where $i, j \geq 1$. Then it holds w.h.p. that $N_{i,j}(t) = n \cdot n_{i,j}\left(\frac{t}{n}\right) + o(n)$, where function $n_{i,j}\left(\frac{t}{n}\right)$ is the solution from Lemma 5 for all $i, j \geq 1$.*

Now after learning how $N_{i,j}(t)$'s evolve over time, we determine when the pure literal elimination algorithm stops and how many clauses it satisfies. First let $C(t)$ denote the number of unpaired *clones* of pure literals at time t, and let $c(\tau) := \frac{C(\tau n)}{n}$ be its scaled version at scaled time τ. Then we have the following result:

Lemma 7. *The number of unpaired pure clones at time $0 \leq t \leq t_0$ is*

$$C(t) = n \cdot c\left(\frac{t}{n}\right) + o(n),$$

where

$$\begin{aligned}c(\tau) &= \gamma - k\tau - z(\tau)\gamma + z(\tau)\sum_{\ell\geq 1}\ell p_\ell\left(1 - \frac{z(\tau)}{2}\right)^{\ell-1}\\ &= \gamma - k\tau - z(\tau)\gamma - \frac{z(\tau)}{2}\frac{d}{dx}\left[G\left(1 - \frac{x}{2}\right)\right]\Big|_{x=z(\tau)}.\end{aligned}$$

Here $G(x) = \mathbb{E}\left[x^\xi\right]$ is the probability-generating function of the r.v. ξ given by (1)*.*

Now, given Lemma 7, the stopping time of Algorithm 1 becomes almost self-evident.

Theorem 1. *Let $\phi \sim \mathbb{C}_n^k(\{\xi_i\}_{i=1}^n)$, where $\mathbb{E}\xi_i < \infty$. Then the pure literal elimination heuristic satisfies w.h.p. $(1 + o(1))\,\tau_0 n$ clauses with $\tau = \tau_0 > 0$ being the smallest solution of the equation*

$$\gamma - k\tau - z(\tau)\gamma - \frac{z(\tau)}{2}\frac{d}{dx}\left[G\left(1 - \frac{x}{2}\right)\right]\Big|_{x=z(\tau)} = 0, \quad (4)$$

where $z(\tau)$ is the function defined in Lemma 5.

Proof. Recall that the Algorithm 1 stops as soon as all pure clones get paired. The number of unpaired pure clones at time t is $C(t) = n \cdot c\left(\frac{t}{n}\right) + f(n)$ w.h.p., where $f(n) = o(n)$,

as it follows from Lemma 7. Introduce an increasing function $\lambda(n)$, such that $f(n) \ll \lambda(n) \ll n$. For the sake of analysis instead of stopping the algorithm when all pure clones get exhausted, we stop its execution when $C(t)$ becomes less than $\lambda(n)$. Although, the algorithm could still continue working and satisfy more clauses, but as it will become apparent from the proof, difference in the number of clauses that Algorithm 1 satisfies, which are not satisfied by stopping the algorithm at the $\lambda(n)$ mark is of order at most $\lambda(n) = o(n)$, which is negligible comparing to the number of clauses we do satisfy.

Hence, it follows that the algorithm's stopping time t_0 is the time, when $C(t_0) \leq \lambda(n)$ for the first time. When we consider the same condition in terms of the scaled version of the number of unpaired pure clones, we have that $\tau_0 = \frac{t_0}{n}$ is the time, when $c(\tau_0) = \frac{C(t_0)}{n} \leq \frac{\lambda(n)}{n} = o(1)$. In other words, $\tau = \tau_0$ is the first time $c(\tau)$ crosses 0 (up to $o(1)$ additive term, since $c(\tau)$ is a smooth continuous function).

Therefore, by finding the closest to zero $\tau_0 > 0$, which satisfies equation (4) gives us the stopping time of Algorithm 1, from which we obtain the number of satisfied clauses $t_0 = (1 + o(1))\,\tau_0 n$ (the $(1 + o(1))$ multiplicative factor here is caused by us stopping the algorithm at $\lambda(n) = o(n)$ level instead of when $C(t)$ turns to 0). □

Note that if $\xi \sim D(\theta)$ for some parametric probability distribution over $\mathbb{N}^+$, then the evolution of the scaled number of pure clones can be itself viewed as not only a function of scaled time τ, but also of parameter θ, i.e. $c(\tau) = c_\theta(\tau)$. In that case the most interesting values of θ are the ones, when $c_\theta(\tau)$ only touches zero at $\tau = \tau_0$ and bounces back as τ increases. In that case we should see sharp and sudden changes in the efficacy of the pure literal algorithm in the neighbourhood of such critical θ's. We demonstrate examples of this phenomenon in the next section, where we discuss our experiment.

Experiments

To verify our results experimentally, we picked 4 probability distributions over $\mathbb{N}^+$, coming from different classes:

1. *Subgaussian distribution* $S(\mu, \sigma)$ with probability distribution function $\Pr[S(\mu,\sigma) = x] = C_{\mu,\sigma} e^{-\left(\frac{x-\mu}{\sigma}\right)^2}$, where $C_{\mu,\sigma}$ is its normalizing constant. $S(\mu, \sigma)$ resembles normal distribution but with support for whole numbers only. As the name suggests, subgaussian distribution comes from the class of subgaussian distributions (Rivasplata 2012).
2. *Poisson distribution* $P(\lambda)$ with distribution $\Pr[P(\lambda) = x] = \frac{\lambda^{x-1} e^{-\lambda}}{(x-1)!}$. The distribution is identical to classical Poisson distribution but with support defined only for positive natural numbers. The Poisson distribution is an example of exponentially decaying distributions.
3. *Log-normal distribution* $L(\eta, \sigma)$ distributed as $\Pr[L(\eta,\sigma) = x] = \frac{M_{\eta,\sigma}}{x} e^{-\left(\frac{\log x - \mu}{\sigma}\right)^2}$, where $M_{\eta,\sigma}$ is the normalizing constant. $L(\eta, \sigma)$ was inspired by the classical continuous log-normal distribution, and it belongs to the class of distributions with tail's decay rate being in-between exponentially decaying distributions and the ones with polynomially heavy tails.
4. *Zeta distribution* $Z(\alpha)$ with distribution $\Pr[Z(\alpha) = x] = \frac{x^{-\alpha}}{\zeta(\alpha)}$, where $\zeta(\alpha) = \sum_{x \geq 1} x^{-\alpha}$ is the Riemann zeta function. $Z(\alpha)$ is a canonical example of a discrete heavy-tailed distribution.

Remark. *For $S(\mu, \sigma)$ and $L(\eta, \sigma)$ we fix their second parameter $\sigma = 1$, and now all 4 distributions parametrized by a single parameter.*

The goal of the experiment is to measure performance of the pure literal elimination rule on random k-CNFs formulas having different degree distributions. The distributions we picked cover a wide spectrum of tail dynamics at infinity, hence, this should help us to develop an intuition on how distributions of degrees affect performance of the pure literal elimination.

Let us denote by $D(\theta)$ an arbitrary one-parameter probability distribution with θ being the distribution's parameter. Let $\mathcal{D} = \{S(\mu, 1), P(\lambda), L(\eta, 1), Z(\alpha)\}$ be the set of the 4 chosen distributions. For every fixed $D(\theta) \in \mathcal{D}$ let θ_γ be the value of its parameter, so that the expected value $\mathbb{E}[D(\theta_\gamma)] = \gamma$.

Next we describe setup of the experiment. First, we fix a list of 13 average degrees $\mathcal{E} = \{3, 4, 5, \ldots, 14, 15\}$. Then for every $D(\theta) \in \mathcal{D}$ and every $\gamma \in \mathcal{E}$, we estimate θ_γ. Next, we produce random[1] 3-CNF formulas $\phi \sim \mathbb{C}^3_n((\xi_i)_{i=1}^n)$, where $\xi_i \sim D(\theta_\gamma)$ are i.i.d. r.vs. with $\mathbb{E}\xi_i = \gamma$, and n, the number of variables of sampled formulas, we fix to be 100,000. We solve the generated formulas with pure literal elimination algorithm, and record how many clauses the algorithm satisfies. Let the number of satisfied clauses be T, and its scaled version $\tau_0 = \frac{T}{n} = 10^{-5} T$. Then with every sampled formula ϕ we associate a tuple $(\phi, D(\theta_\gamma), \gamma, \tau_0)$, where ϕ is the formula itself, $D(\theta_\gamma)$ is the distribution, which was used for degree sequence generation, γ is the average degree, and finally τ_0 is the scaled number of clauses of ϕ that the pure literal algorithm was able to satisfy. By sampling many formulas for each $D(\theta) \in \mathcal{D}$ and $\gamma \in \mathcal{E}$, we should be able to get a good estimate of the "true" τ_0 value. In our experiment we produced 7,777 instances for each distribution and each average degree γ.

After collecting enough data points, we averaged τ_0 for each γ and every $D(\theta)$, and plotted this estimate of τ_0 on Figure 1. We also calculated the 0.99 confidence interval of the obtained estimates, which we also added to the figure (surprisingly, whiskers, marking the bounds of the confidence intervals, are almost indiscernible, which means that the estimates seem to converge rapidly to their true values. We say it is rather surprising, since, based on our experience, quantities derived from heavy-tailed distributions quite often do not exhibit nice convergence rates. One case of this phenomenon can be seen, for example, in (Omelchenko and Bulatov 2021b)).

[1] We use std::random_device() from the standard library of C++ to generate seeds for PRG.

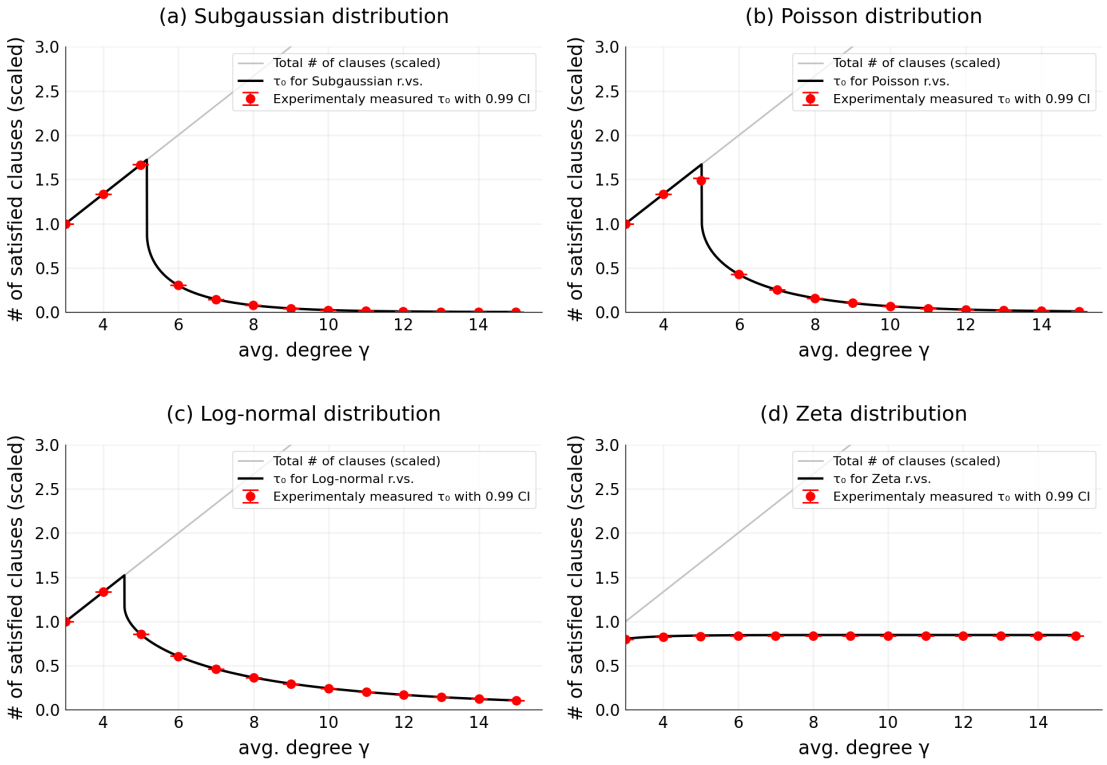

Figure 1: Experimentally obtained efficacy of the pure literal heuristic on formulas from $\mathbb{C}_n^3((\xi_i)_{i=1}^n)$ vs. theoretically predicted efficacy, where r.vs. ξ_i's follow (a) subgaussian, (b) Poisson, (c) log-normal, and (d) zeta distribution. Black curves represent the scaled number of satisfied clauses $\tau_0(\gamma)$ as a function of γ. Grey slanted lines show the typical number of total clauses, and round markers are experimentally obtained estimates of $\tau_0(\gamma)$ with whiskers showing bounds of the 0.99 confidence intervals.

Also for each $D(\theta) \in \mathcal{D}$ we calculate numerically $\tau_0 = \tau_0(\theta)$ by solving the equation (4). Note that now $\tau(\theta)$ can be viewed as a function of the distribution's parameter, which, in turn, controls the expected value $\gamma(\theta)$. Hence, we can plot $(\gamma(\theta), \tau_0(\theta))$ by varying θ. We pick θ's so that $\gamma(\theta) \in [3, 15]$, and calculate corresponding $\tau_0(\theta)$. Obtained functions of the "true" values of τ_0 are represented as black curves on Figure 1. Additionally, we plot as grey slanted lines the scaled average number of clauses that formulas with average degree γ have, i.e. for each γ this quantity is equal to $\frac{|\phi|}{n} = \frac{\gamma}{k} = \frac{\gamma}{3}$.

As it follows from the experiment, we have obtained good evidence supporting Theorem 1. The "true" values of the scaled number of satisfied clauses τ_0, given by equation 4, predict well how the pure literal elimination performs in reality. There are some more observations we can make from the experimental data. It seems the heuristic performs really well, when the distribution of degrees has a rapidly decaying tail (like Poisson and subgaussian), and γ is at most around 5. However, for even slightly larger γ's performance of the pure literal heuristic drops significantly and it becomes almost useless. Hence, this average degree $\gamma = 5$ and the corresponding values of θ_γ for Poisson and subgaussian distributions seem to be critical.

However, the more heavy-tailed distributions exhibit a somewhat reversed dynamic. The pure literal algorithm does not seem to perform well on them for lower values of γ. For example, the algorithm was able to satisfy only around 1/2 of all clauses of formulas with log-normally distributed degrees when $\gamma = 5$, while on subgaussian r.vs. it was able to satisfy all or almost all clauses at the same average degree. But as we increase γ, the rule starts performing better on the more heavy-tailed distributions comparing to their light-tailed counterparts. Consider, for example, Poisson and zeta distributions. When $\gamma > 10$, the algorithm is unable to satisfy any reasonable number of clauses in case of Poisson distributed degrees, but for zeta distribution efficacy of the pure literal elimination rule drops much and much slower as we increase γ.

References

Achlioptas, D. 2001. Lower bounds for random 3-SAT via differential equations. *Theoretical Computer Science*, 265.

Achlioptas, D. 2009. Random Satisfiability. *Frontiers in Artificial Intelligence and Applications*, 185.

Achlioptas, D.; and Peres, Y. 2003. The threshold for random k-SAT is $2^k \ln 2 - O(k)$. 223–231.

Achlioptas, D.; and Sorkin, G. 2000. Optimal myopic algorithms for random 3-SAT. *Annual Symposium on Foundations of Computer Science - Proceedings*.

Alekhnovich, M.; and Ben-Sasson, E. 2007. Linear Upper Bounds for Random Walk on Small Density Random 3-CNFs. *SIAM J. Comput.*, 36(5): 1248–1263.

Ansótegui, C.; Bonet, M. L.; Giráldez-Cru, J.; and Levy, J. 2014. The Fractal Dimension of SAT Formulas. In Demri, S.; Kapur, D.; and Weidenbach, C., eds., *Automated Reasoning*, 107–121. Cham: Springer International Publishing. ISBN 978-3-319-08587-6.

Ansótegui, C.; Bonet, M. L.; Giráldez-Cru, J.; Levy, J.; and Simon, L. 2019. Community Structure in Industrial SAT Instances. *J. Artif. Intell. Res.*, 66: 443–472.

Ansótegui, C.; Bonet, M. L.; and Levy, J. 2019. Phase Transition in Realistic Random SAT Models. In Sabater-Mir, J.; Torra, V.; Aguiló, I.; and Hidalgo, M. G., eds., *Artificial Intelligence Research and Development - Proceedings of the 22nd International Conference of the Catalan Association for Artificial Intelligence, CCIA 2019, Mallorca, Spain, 23-25 October 2019*, volume 319 of *Frontiers in Artificial Intelligence and Applications*, 213–222. IOS Press.

Ansótegui, C.; Bonet, M.; and Levy, J. 2009. On the Structure of Industrial SAT Instances. volume 5732, 127–141. ISBN 978-3-642-04243-0.

Ansótegui, C.; Bonet, M. L.; Levy, J.; and Manyà, F. 2008. Measuring the Hardness of SAT Instances. volume 1, 222–228.

Beyersdorff, O.; and Kullmann, O. 2014. Unified Characterisations of Resolution Hardness Measures. In *SAT*.

Bohman, T. 2009. The triangle-free process. *Advances in mathematics (New York. 1965)*, 221(5): 1653–1677.

Bohman, T.; and Keevash, P. 2010. The early evolution of the H-free process. *Inventiones mathematicae*, 181(2): 291–336.

Borovkov, A.; and Borovkov, K., eds. 2008. *Asymptotic analysis of random walks: heavy-tailed distributions*, volume 118 of *Encyclopedia of mathematics and its applications*. Cambridge University Press. ISBN 978-0-51172-139-7.

Borovkov, A. A. 2013. *Probability Theory by Alexandr A. Borovkov.* Universitext. Springer London : Imprint: Springer, 1st ed. 2013. edition. ISBN 1-4471-5200-X.

Broder, A. Z.; Frieze, A. M.; and Upfal, E. 1993. On the Satisfiability and Maximum Satisfiability of Random 3-CNF Formulas. In *Proceedings of the Fourth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '93, 322–330. USA: Society for Industrial and Applied Mathematics. ISBN 0898713137.

Chvatal, V.; and Reed, B. 1992. Mick gets some (the odds are on his side) (satisfiability). In *Proceedings., 33rd Annual Symposium on Foundations of Computer Science*, 620–627.

Cooper, C.; Frieze, A.; and Sorkin, G. B. 2007. Random 2-SAT with Prescribed Literal Degrees. *Algorithmica*, 48(3): 249–265.

Coppersmith, D.; Gamarnik, D.; Hajiaghayi, M. T.; and Sorkin, G. B. 2003. Random MAX SAT, random MAX CUT, and their phase transitions. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms, January 12-14, 2003, Baltimore, Maryland, USA*, 364–373. ACM/SIAM.

Ding, J.; Sly, A.; and Sun, N. 2014. Proof of the Satisfiability Conjecture for Large k. *To Appear in STOC*.

Erdös, P.; and Selfridge, J. 1973. On a combinatorial game. *Journal of Combinatorial Theory, Series A*, 14(3): 298–301.

Friedrich, T.; Krohmer, A.; Rothenberger, R.; Sauerwald, T.; and Sutton, A. M. 2017. Bounds on the Satisfiability Threshold for Power Law Distributed Random SAT. In Pruhs, K.; and Sohler, C., eds., *25th Annual European Symposium on Algorithms, ESA 2017, September 4-6, 2017, Vienna, Austria*, volume 87 of *LIPIcs*, 37:1–37:15. Schloss Dagstuhl - Leibniz-Zentrum für Informatik.

Frieze, A. M.; and Suen, S. 1996. Analysis of Two Simple Heuristics on a Random Instance of k-SAT. *J. Algorithms*, 20(2): 312–355.

Goerdt, A. 1996. A Threshold for Unsatisfiability. *J. Comput. Syst. Sci.*, 53(3): 469–486.

Hajiaghayi, M. T.; and Sorkin, G. B. 2003. The satisfiability threshold of random 3-SAT is at least 3.52. Technical report, IBM.

Kaporis, A. C.; Kirousis, L. M.; and Lalas, E. G. 2002. The Probabilistic Analysis of a Greedy Satisfiability Algorithm. In Möhring, R. H.; and Raman, R., eds., *Algorithms - ESA 2002, 10th Annual European Symposium, Rome, Italy, September 17-21, 2002, Proceedings*, volume 2461 of *Lecture Notes in Computer Science*, 574–585. Springer.

Kaporis, A. C.; Kirousis, L. M.; and Lalas, E. G. 2006. The probabilistic analysis of a greedy satisfiability algorithm. *Random Struct. Algorithms*, 28(4): 444–480.

Kim, J. H. 2004. The Poisson Cloning Model for Random Graphs, Random Directed Graphs and Random k-SAT Problems. In Chwa, K.; and Munro, J. I., eds., *Computing and Combinatorics, 10th Annual International Conference, COCOON 2004, Jeju Island, Korea, August 17-20, 2004, Proceedings*, volume 3106 of *Lecture Notes in Computer Science*, 2. Springer.

Kim, J. H. 2008. Finding cores of random 2-SAT formulae via Poisson cloning. *CoRR*, abs/0808.1599.

Larrabee, T.; and Tsuji, Y. 1992. Evidence for a Satisfiability Threshold for Random 3CNF Formulas.

Luby, M.; Mitzenmacher, M.; and Shokrollahi, M. A. 1998. Analysis of Random Processes via And-Or Tree Evaluation. In Karloff, H. J., ed., *Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, 25-27 January 1998, San Francisco, California, USA*, 364–373. ACM/SIAM.

Mitchell, D.; Selman, B.; and Levesque, H. 1992. Hard and Easy Distributions of SAT Problems.

Mitzenmacher, M. 1997. Tight thresholds for the pure literal rule. Technical report.

Molloy, M. 2005. Cores in random hypergraphs and Boolean formulas. *Random Struct. Algorithms*, 27(1): 124–135.

Omelchenko, O.; and Bulatov, A. 2021a. Satisfiability and Algorithms for Non-uniform Random k-SAT. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(5): 3886–3894.

Omelchenko, O.; and Bulatov, A. A. 2019. Concentration inequalities for sums of random variables, each having power bounded tails.

Omelchenko, O.; and Bulatov, A. A. 2021b. Satisfiability Threshold for Power Law Random 2-SAT in Configuration Model. *Theoretical Computer Science*.

Rivasplata, O. 2012. Subgaussian random variables: An expository note.

Warnke, L. 2014. When does the K4-free process stop? *Random structures & algorithms*, 44(3): 355–397.

Warnke, L. 2019. On Wormald's differential equation method. *ArXiv*, abs/1905.08928.

Wormald, N. 1999. *The differential equation method for random graph processes and greedy algorithms*, 73–155. Wydawnictwo Naukowe Pwn.

Wormald, N. C. 1995. Differential Equations for Random Processes and Random Graphs. *The Annals of applied probability*, 5(4): 1217–1235.

Supplemental Materials

Here we provide proofs of the results from the paper, as well as state Wormald's differential equations method theorem. For the reader's convenience we give our lemmas and the theorem together with their proofs.

Wormald's Differential Equations Method Theorem

In what follows, we say $X = O(f(n))$ for a r.v. X when we mean that

$$\max\{\, x \mid \Pr[X = x] \neq 0\} = O(f(n)).$$

Also function $f : \mathbb{R}^\ell \to \mathbb{R}$ is said to satisfy *Lipschitz condition* on $\mathcal{D} \subseteq \mathbb{R}^\ell$ if there exists a constant $L > 0$, such that $|f(x) - f(y)| \leq L||x - y||_1$ for all $x, y \in \mathcal{D}$.

Theorem A ((Wormald 1995)). *Let $X_i(t)$ be a sequence of ℓ real-valued random processes with $t \in \mathbb{N}$, such that for all $i \in [\ell]$ and every t, it holds that $|X_i(t)| = O(n)$. Let $H(t) = \bigcup_{0 \leq j \leq t} \{X_1(j), X_2(j), \ldots, X_\ell(j)\}$ denote the complete history of evolution of processes $X_i(j)$ up to and including time $t \in \mathbb{N}$.*

Denote by $\vec{X}(t) = (X_1(t), X_2(t), \ldots, X_\ell(t))$, and let

$$\mathcal{I} = \Big\{(x_1, x_2, \ldots, x_N) \in \mathbb{R}^N \;\Big|\; \Pr\Big[\vec{X}(0) = (x_1 n, x_2 n, \ldots, x_N n)\Big] > 0\Big\}$$

be the set of possible initial values for the process $\vec{X}(t)$. Let $\mathcal{D} \subset \mathbb{R}^{\ell+1}$ be some open bounded connected set containing the closure of $\{(\tau, x_1, x_2, \ldots, x_\ell) \in \mathbb{R}^{\ell+1} \mid \tau \geq 0\}$ with $\{(0, x_1, x_2, \ldots, x_\ell) \mid (x_1, x_2, \ldots, x_\ell) \in \mathcal{I}\}$.

Let $f_i : \mathbb{R}^{\ell+1} \to \mathbb{R}$ be a collection of ℓ functions, and suppose the following three conditions hold for all $0 \leq t < T_0(n)$ for some function $T_0(n) = O(n)$, and any $i \in [\ell]$:

1. $\mathbb{E}[X_i(t+1) - X_i(t) \mid H(t)] = f_i\left(\frac{t}{n}, \frac{X_1(t)}{n}, \frac{X_2(t)}{n}, \ldots, \frac{X_\ell(t)}{n}\right) + o(1)$;
2. $\Pr\Big[\Big|X_i(t+1) - X_i(t)\Big| \geq n^{1/2} | H(t)\Big] = o\left((nl)^{-1}\right)$;
3. *function f_i is continuous and satisfies Lipschitz condition on $\mathcal{D}$.*

Then the following are true:

(a) *For $(0, \hat{z}_1, \hat{z}_2, \ldots, \hat{z}_\ell) \in \mathcal{D}$ the system of differential equations*

$$\frac{dz_i}{d\tau} = f_i(\tau, z_1, z_2, \ldots, z_\ell), \qquad i \in [\ell]$$

has a unique solution in $\mathcal{D}$ for $z_i : \mathbb{R} \to \mathbb{R}$ passing through $z_i(0) = \hat{z}_i$ for all $i \in [n]$, and extending arbitrarily close to the boundary of domain $\mathcal{D}$;

(b) *it holds almost surely that for any $i \in [\ell]$*

$$X_i(t) = n z_i\left(\frac{t}{n}\right) + o(n),$$

uniformly for $0 \leq t < min\{T(n), \sigma(n)\}$, where z_i is the solution from part (a), and $\sigma(n)$ is the supremum of those t to which z_i's can be extended without violating the Lipschitz condition.

Proof of Lemma 1

Lemma 1 (WLLN). *Let $\xi_1, \xi_2, \ldots, \xi_n$ be a collection of n independent r.vs. with finite expectation $\mathbb{E}|\xi_i| < \infty$. Then it holds w.h.p.*

$$\sum_{i=1}^n \xi_i = \sum_{i=1}^n \mathbb{E}\xi_i + o(n).$$

Proof. Let $\zeta_i = \xi_i - \mathbb{E}\xi_i$. Clearly, $\mathbb{E}\zeta_i = 0$. Then

$$\sum_{i=1}^n \xi_i = \sum_{i=1}^n (\zeta_i + \mathbb{E}\xi_i) = \sum_{i=1}^n \mathbb{E}\xi_i + \sum_{i=1}^n \zeta_i =: \sum_{i=1}^n \mathbb{E}\xi_i + Z,$$

where $Z = \sum_{i=1}^n \zeta_i$.

Let $\phi_{\frac{Z}{n}}(t) = \mathbb{E}e^{it\frac{Z}{n}}$ be the characteristic function of $\frac{Z}{n}$. Our goal is to show that for any *fixed* $t \in \mathbb{R}$ function $\phi_{\frac{Z}{n}}(t) \to 1 \equiv \phi_0(t)$, when $n \to \infty$, which implies that $\frac{Z}{n} \to 0$ in probability by Levy's continuity theorem. Since Z is the sum of independent r.vs. $\zeta_1, \zeta_2, \ldots, \zeta_n$, it follows that

$$\begin{aligned}
\phi_{\frac{Z}{n}}(t) &= \mathbb{E}\left[e^{it\frac{\sum \zeta_i}{n}}\right] \\
&= \prod_{i=1}^n \mathbb{E}\left[e^{i\frac{t}{n}\zeta_i}\right] \\
&= \prod_{i=1}^n \phi_{\zeta_i}\left(\frac{t}{n}\right) \\
&= \prod_{i=1}^n \phi_{\zeta_i}(x), \qquad (5)
\end{aligned}$$

where $x = \frac{t}{n} \to 0$ when $n \to \infty$. Since $\mathbb{E}\zeta_i = 0$, and so it is finite, it holds that for any $i \in [n]$ and any fixed $t \in \mathbb{R}$ function $\phi_{\zeta_i}(x)$ can be expanded via Taylor series around $x_0 = 0$ such that

$$\begin{aligned}
\phi_{\zeta_i}(x) &= \mathbb{E}\left[e^{ix\zeta_i}\right] \\
&= \mathbb{E}[1 + ix\zeta_i + o(|x|)] \\
&= 1 + ix\mathbb{E}\zeta_i + o(|x|) \\
&= 1 + o(|x|).
\end{aligned}$$

Hence, $\phi_{\zeta_i}(x) = 1 + o(|x|) \to e^{o(|x|)}$, when $x \to 0$. Recall that $x = \frac{t}{n}$ and t is fixed, so $\phi_{\zeta_i}(x) \to e^{o\left(\frac{1}{n}\right)}$. Therefore, we conclude that (5) approaches

$$\phi_{\frac{Z}{n}}(t) = \prod_{i=1}^n \phi_{\zeta_i}(x) \to \prod_{i=1}^n e^{o(n^{-1})} = e^{o(1)} \to 1,$$

when $n \to \infty$, and we conclude that $\frac{Z}{n} \to 0$ in probability, or, equivalently, that for any $\epsilon > 0$

$$\Pr\left[\left|\frac{Z}{n}\right| > \epsilon\right] = \Pr\Big[\,|Z| > \epsilon n\Big] \to 0,$$

when $n \to 0$. In other words, $Z = \sum_{i=1}^{n} \zeta_i = o(n)$ w.h.p., and so it follows that w.h.p.

$$\sum_{i=1}^{n} \xi_i = \sum_{i=1}^{n} \mathbb{E}\xi_i + Z = \sum_{i=1}^{n} \mathbb{E}\xi_i + o(n),$$

and the lemma follows.

□

Proof of Lemma 2

Lemma 2. *It holds that $S_n = (1 + o(1))\,\gamma n$ w.h.p.*

Proof. As it follows from the CREATECLONES$((\xi_i)_{i=1}^{n}, k)$ procedure, S_n is a sum of r.vs. ξ_i's. Hence, $S_n = \sum_{i=1}^{n} \xi_i$, and since each $\mathbb{E}\xi_i < \infty$, it follows that w.h.p. $S_n = \sum_{i=1}^{n} \xi_i = (1 + o(1)) \sum_{i=1}^{n} \mathbb{E}\xi_i$ by Lemma 1. Finally, by definition, $\gamma = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\xi_i$, thus, it holds that w.h.p. $S_n = (1 + o(1))\,\gamma n$. The lemma is proved. □

Proof of Lemma 3

Lemma 3. *The probability a randomly chosen variable produces exactly i positive and j negative clones is*

$$p_{i,j} = 2^{-(i+j)} \binom{i+j}{i} \Pr[\xi = i + j] = 2^{-(i+j)} \binom{i+j}{i} p_{i+j}.$$

Proof. There are n variables, and the probability that a randomly selected variable produces $i+j$ clones in total is thus $\frac{1}{n}\sum_{m=1}^{n} \Pr[\xi_m = i + j] = \Pr[\xi = d] = p_d$. Now, given that the chosen variable has produced $i + j$ clones, the probability that i of them will obtain positive sign and j will get negative sign is $\binom{i+j}{i} 2^{-(i+j)}$, since we have $i + j$ clones in total and the positive sign is assigned to each clone independently with probability $1/2$. Therefore, after combining probabilities of the two above events together, we prove the lemma. □

Proof of Lemma 4

Lemma 4. *Let $H(t) := \bigcup_{0 \le t' \le t} \bigcup_{i,j \ge 1} \{N_{i,j}(t')\}$ be the complete history of the evolution of the processes $N_{i,j}(t')$ up to and including time t. Then for all $i, j \ge 1$ and all $0 < t < t_0$, it holds:*

1. $\mathbb{E}[N_{i,j}(t+1) - N_{i,j}(t) \mid H(t)]$
$= f_{i,j}\left(\frac{t}{n}, \frac{N_{i,j}(t)}{n}, \frac{N_{i+1,j}(t)}{n}, \frac{N_{i,j+1}(t)}{n}\right) + o(1),$
where
$$f_{i,j}(\tau, n_{i,j}(\tau), n_{i+1,j}(\tau), n_{i,j+1}(\tau)) = -\frac{k-1}{\gamma - k\tau}(i+j)n_{i,j}(\tau) + \frac{k-1}{\gamma - k\tau}\Big[(i+1)n_{i+1,j}(\tau) + (j+1)n_{i,j+1}(\tau)\Big];$$
2. $\Pr\Big[\Big|N_{i,j}(t+1) - N_{i,j}(t)\Big| > k \mid H(t)\Big] = 0$;
3. $f_{i,j}(\tau, n_{i,j}(\tau), n_{i+1,j}(\tau), n_{i,j+1}(\tau))$ *is continuous and satisfies Lipschitz condition for* $0 \le \tau \le \frac{t_0}{n} < \frac{\gamma}{k}$.

Proof. 1. Recall how the pure literal elimination algorithm (Algorithm 1) works. We pick a clone of a pure literal, and form a new clause out of it. Then we are left with $k-1$ empty placeholders in the newly created clause, which we fill by sampling u.a.r. without replacement $k-1$ unpaired clones. Since at each iteration we pair exactly k clones, we have $S_n - kt$ unpaired clones at time t. Hence, the probability that we pick a clone of a variable from $N_{i,j}(t)$ is $\frac{1}{S_n - kt}(i+j)N_{i,j}(t)$, and since we have $k-1$ "tries" to pick their clones, it follows that on average $N_{i,j}(t)$ loses $\frac{k-1}{S_n - kt}(i+j)N_{i,j}(t)$ variables.

Likewise, it may happen that an empty placeholder will be filled by a positive clone of a variable from $N_{i+1,j}(t)$ or by a negative clone of a variable from $N_{i,j+1}(t)$, which happens with probability $\frac{1}{S_n - kt}(i+1)N_{i+1,j}(t)$ and $\frac{1}{S_n - kt}(j+1)N_{i,j+1}(t)$ respectively. In that case $N_{i,j}(t)$ gains an extra variable. On average it happens $\frac{k-1}{S_n - kt}\left[(i+1)N_{i+1,j}(t) + (j+1)N_{i,j+1}(t)\right]$ times, since we have $k-1$ placeholders.

Note that there exist other cases when $N_{i,j}(t)$ gains and loses variables. For example, it may happen that we sample two positive clones of a variable from $N_{i+2,j}(t)$. In that case $N_{i+2,j}(t)$ would lose this one variable, while $N_{i,j}(t)$ would gain it. However, such "bad" events are rare and the number of such clauses is dominated by the number of clauses, where each variable appears only once.

Hence, after combining the averages of incoming and outgoing variables to and from $N_{i,j}(t)$, we obtain the expected one step change in $N_{i,j}(t)$

$$\begin{aligned}
&\mathbb{E}[N_{i,j}(t+1) - N_{i,j}(t) \mid H(t)] \\
&= -\frac{k-1}{S_n - kt}(i+j)N_{i,j}(t) \\
&+ \frac{k-1}{S_n - kt}\Big[(i+1)N_{i+1,j}(t) + (j+1)N_{i,j+1}(t)\Big] + o(1) \\
&= -\frac{k-1}{\gamma - k\tau}(i+j)n_{i,j}(\tau) \\
&+ \frac{k-1}{\gamma - k\tau}\Big[(i+1)n_{i+1,j}(\tau) + (j+1)n_{i,j+1}(\tau)\Big] + o(1)
\end{aligned}$$

after substituting $n_{i,j}(\tau) = \frac{N_{i,j}(\tau n)}{n}$, $\tau = \frac{t}{n}$, and $S_n = (1 + o(1))\,\gamma n$, which holds w.h.p.

2. The proof is almost self-evident. The pure elimination algorithm pairs at each step exactly k clones, which means that we affect at most k variables. Therefore, probability of the difference $|N_{i,j}(t+1) - N_{i,j}(t)|$ jumping over k is 0, and the result follows.

3. Note that the functions

$$\begin{aligned}
&f_{i,j}(\tau, n_{i,j}(\tau), n_{i+1,j}(\tau), n_{i,j+1}(\tau)) \\
&= -\frac{k-1}{\gamma - k\tau}(i+j)n_{i,j}(\tau) \\
&+ \frac{k-1}{\gamma - k\tau}\Big[(i+1)n_{i+1,j}(\tau) + (j+1)n_{i,j+1}(\tau)\Big]
\end{aligned}$$

are in fact rational functions of variables $\tau, n_{i,j}(\tau), n_{i+1,j}(\tau)$, and $n_{i,j+1}(\tau)$, and so they satisfies Lipschitz condition when $\tau < \frac{\gamma}{k}$. Moreover, recall that functions $n_{i,j}(\tau) = \frac{N_{i,j}(\tau n)}{n}$ denote the scaled number of variables with *unpaired* i positive and j negative clones, therefore,

$$\begin{aligned}(i+j)n_{i,j} &\le \gamma - k\tau,\\ (i+1)n_{i+1,j} &\le \gamma - k\tau,\\ (j+1)n_{i,j+1} &\le \gamma - k\tau,\end{aligned}$$

since the RHS in all of these inequalities is the scaled *total* number of unpaired clones at time t. Hence, we conclude that for any $\tau < \frac{\gamma}{k}$

$$\begin{aligned}&|f_{i,j}(\tau, n_{i,j}(\tau), n_{i+1,j}(\tau), n_{i,j+1}(\tau))|\\ &\quad\le (k-1)\left|\frac{(i+j)n_{i,j}(\tau)}{\gamma - k\tau}\right|\\ &\qquad + (k-1)\left|\frac{(i+1)n_{i+1,j}(\tau)}{\gamma - k\tau}\right|\\ &\qquad\quad + (k-1)\left|\frac{(j+1)n_{i,j+1}(\tau)}{\gamma - k\tau}\right|\\ &\quad\le 3(k-1).\end{aligned}$$

Thus, verification of all 3 claims completes proof of the lemma. □

Proof of Lemma 5

Lemma 5. *The system* (3) *defined for all $i, j \ge 1$ together with initial values $n_{i,j}(0) = 2^{-(i+j)}\binom{i+j}{i}p_{i+j}$ has a unique solution*

$$\begin{aligned}n_{i,j}(\tau) &= 2^{-(i+j)}\binom{i+j}{i}\\ &\quad\times \sum_{\ell\ge 1}\binom{\ell}{i+j} z(\tau)^{i+j}\left(1 - z(\tau)\right)^{\ell-(i+j)},\end{aligned}$$

where

$$z(\tau) = \left(1 - \frac{k\tau}{\gamma}\right)^{1-\frac{1}{k}}.$$

Proof. The proof consists simply in verifying that taking derivative of the above stated functions $n_{i,j}(\tau)$ results in correct derivatives, and also that they satisfy the initial conditions.

First, we check that these functions truly satisfy the condition that $n_{i,j}(0) = 2^{-(i+j)}\binom{i+j}{i}p_{i+j}$.

$$\begin{aligned}n_{i,j}(0) &= 2^{-(i+j)}\binom{i+j}{i}\\ &\quad\times \sum_{\ell\ge 1}\binom{\ell}{i+j} z(0)^{i+j}\left(1 - z(0)\right)^{\ell-(i+j)} p_\ell\\ &= 2^{-(i+j)}\binom{i+j}{i}p_{i+j}, \text{ since } z(0) = 1.\end{aligned}$$

Next, introduce function $B_{\alpha,\beta}(x) = \binom{\beta}{\alpha}x^\alpha(1-x)^\beta$. Then functions $n_{i,j}(\tau)$ can be expressed as

$$\begin{aligned}n_{i,j}(\tau) &= 2^{-(i+j)}\binom{i+j}{i}\\ &\quad\times \sum_{\ell\ge 1}\binom{\ell}{i+j} z(\tau)^{i+j}\left(1 - z(\tau)\right)^{\ell-(i+j)} p_\ell\\ &= 2^{-(i+j)}\binom{i+j}{i}\sum_{\ell\ge 1} B_{i+j,\ell}(z(\tau))p_\ell. \qquad (6)\end{aligned}$$

Note that

$$\begin{aligned}\frac{d}{d\tau}z(\tau) &= \frac{d}{d\tau}\left[\left(1 - \frac{k\tau}{\gamma}\right)^{1-\frac{1}{k}}\right]\\ &= -\frac{k-1}{\gamma}\left(1 - \frac{k\tau}{\gamma}\right)^{-\frac{1}{k}}\\ &= -\frac{k-1}{\gamma}\left[\frac{1-\frac{k\tau}{\gamma}}{1-\frac{k\tau}{\gamma}}\right]\left(1 - \frac{k\tau}{\gamma}\right)^{-\frac{1}{k}}\\ &= -\frac{k-1}{\gamma - k\tau}\left(1 - \frac{k\tau}{\gamma}\right)^{1-\frac{1}{k}}\\ &= -\frac{k-1}{\gamma - k\tau}z(\tau). \qquad (7)\end{aligned}$$

Next, we take the derivative of the function $B_{i+j,\ell}(z(\tau))$

$$\begin{aligned}&\frac{d}{d\tau}B_{i+j,\ell}(z(\tau))\\ &\quad= \frac{dB_{i+j,\ell}(z(\tau))}{dz(\tau)}\cdot\frac{dz(\tau)}{d\tau}\\ &\quad= -\frac{k-1}{\gamma - k\tau}z(\tau)\left(\frac{d}{dx}B_{i+j,\ell}(x)\right)\Big|_{x=z(\tau)}\end{aligned}$$

after substituting the derivative of $z(\tau)$ with (7). Therefore,

$$\begin{aligned}&\frac{d}{d\tau}B_{i+j,\ell}(z(\tau))\\ &\quad= -\frac{k-1}{\gamma - k\tau}z(\tau)\left(\frac{d}{dx}B_{i+j,\ell}(x)\right)\Big|_{x=z(\tau)}\\ &\quad= -\frac{k-1}{\gamma - k\tau}z(\tau)\binom{\ell}{i+j}\\ &\qquad\times\left(\frac{d}{dx}\left[x^{i+j}(1-x^{\ell-(i+j)})\right]\right)\Big|_{x=z(\tau)}\\ &\quad= -\frac{k-1}{\gamma - k\tau}z(\tau)\binom{\ell}{i+j}\\ &\qquad\times\Big((i+j)x^{i+j-1}(1-x^{\ell-(i+j)})\\ &\qquad\quad -(\ell-(i+j))x^{i+j}(1-x^{\ell-(i+j)-1})\Big)\Big|_{x=z(\tau)}\\ &\quad= -\frac{k-1}{\gamma - k\tau}(i+j)\binom{\ell}{i+j}z(\tau)^{i+j}(1-z(\tau))^{\ell-(i+j)}\\ &\qquad + \frac{k-1}{\gamma - k\tau}(\ell-(i+j))\binom{\ell}{i+j}\end{aligned}$$

$$\times z(\tau)^{i+j+1}(1-z(\tau))^{\ell-(i+j+1)}$$
$$= -\frac{k-1}{\gamma-k\tau}(i+j)B_{i+j,\ell}(z(\tau)) + \frac{k-1}{\gamma-k\tau}(i+j+1)\binom{\ell}{i+j+1} \times z(\tau)^{i+j+1}(1-z(\tau))^{\ell-(i+j+1)}$$
$$= -\frac{k-1}{\gamma-k\tau}(i+j)B_{i+j,\ell}(z(\tau)) + \frac{k-1}{\gamma-k\tau}(i+j+1)B_{i+j+1,\ell}(z(\tau)).$$

Therefore, after taking derivative of $n_{i,j}(\tau)$ expressed as (6), we obtain

$$\frac{d}{d\tau}n_{i,j}(\tau) = 2^{-(i+j)}\binom{i+j}{i}\sum_{\ell\geq 1}\frac{d}{d\tau}B_{i+j,\ell}(z(\tau))p_\ell$$
$$= -\frac{k-1}{\gamma-k\tau}(i+j)2^{-(i+j)}\binom{i+j}{i}\sum_{\ell\geq 1}B_{i+j,\ell}(z(\tau))p_\ell$$
$$+ \frac{k-1}{\gamma-k\tau}(i+j+1)2^{-(i+j)}\binom{i+j}{i}\sum_{\ell\geq 1}B_{i+j+1,\ell}(z(\tau))p_\ell$$
$$= -\frac{k-1}{\gamma-k\tau}(i+j)n_{i,j}(\tau), \text{ due to equality (6)}$$
$$+ \frac{k-1}{\gamma-k\tau}(i+j+1)2^{-(i+j)}\binom{i+j}{i}\sum_{\ell\geq 1}B_{i+j+1,\ell}(z(\tau))p_\ell$$
$$= -\frac{k-1}{\gamma-k\tau}(i+j)n_{i,j}(\tau)$$
$$+ \frac{k-1}{\gamma-k\tau}((i+1)+j)2^{-((i+1)+j)}\binom{i+j}{i} \times \sum_{\ell\geq 1}B_{(i+1)+j,\ell}(z(\tau))p_\ell$$
$$+ \frac{k-1}{\gamma-k\tau}(i+(j+1))2^{-(i+(j+1))}\binom{i+j}{i} \times \sum_{\ell\geq 1}B_{i+(j+1),\ell}(z(\tau))p_\ell$$
$$= -\frac{k-1}{\gamma-k\tau}(i+j)n_{i,j}(\tau)$$
$$+ \frac{k-1}{\gamma-k\tau}(i+1)2^{-((i+1)+j)}\binom{i+j+1}{i} \times \sum_{\ell\geq 1}B_{(i+1)+j,\ell}(z(\tau))p_\ell$$
$$+ \frac{k-1}{\gamma-k\tau}(j+1)2^{-(i+(j+1))}\binom{i+j+1}{i} \times \sum_{\ell\geq 1}B_{i+(j+1),\ell}(z(\tau))p_\ell$$
$$= -\frac{k-1}{\gamma-k\tau}(i+j)n_{i,j}(\tau) + \frac{k-1}{\gamma-k\tau}(i+1)n_{i+1,j}(\tau) + \frac{k-1}{\gamma-k\tau}(j+1)n_{i,j+1}(\tau),$$

where substitutions on the last two lines are done again due to the equality (6). Hence, functions $n_{i,j}(\tau)$ from Lemma 5 are truly a solution to the system of differential equations (3) with initial condition $n_{i,j}(\tau) = 2^{-(i+j)}\binom{i+j}{i}p_{i+j}$. And since functions $f_{i,j}(\tau, n_{i,j}(\tau), n_{i+1,j}(\tau), n_{i,j+1}(\tau))$ satisfy Lipschitz condition for $\tau < \frac{\gamma}{k}$ (Lemma 4), it follows that the above solution is unique. □

Proof of Lemma 6

Lemma 6. *Let $N_{i,j}(t)$ be the number of variables with i positive and j negative unpaired clones at time t, where $i, j \geq 1$. Then it holds w.h.p. that $N_{i,j}(t) = n \cdot n_{i,j}\left(\frac{t}{n}\right) + o(n)$, where function $n_{i,j}\left(\frac{t}{n}\right)$ is the solution from Lemma 5 for all $i, j \geq 1$.*

Proof. Proof readily follows from Lemmas 4, 5, and Theorem A. □

Proof of Lemma 7

Lemma 7. *The number of unpaired pure clones at time $0 \leq t \leq t_0$ is*

$$C(t) = n \cdot c\left(\frac{t}{n}\right) + o(n),$$

where

$$c(\tau) = \gamma - k\tau - z(\tau)\gamma + z(\tau)\sum_{\ell\geq 1}\ell p_\ell\left(1-\frac{z(\tau)}{2}\right)^{\ell-1}$$
$$= \gamma - k\tau - z(\tau)\gamma - \frac{z(\tau)}{2}\frac{d}{dx}\left[G\left(1-\frac{x}{2}\right)\right]\Big|_{x=z(\tau)}.$$

Here $G(x) = \mathbb{E}\left[x^\xi\right]$ is the probability-generating function of the r.v. ξ given by (1)*.*

Proof. Consider approximations of functions $N_{i,j}(t)$ from Lemma 6

$$N_{i,j}(t) = n \cdot n_{i,j}\left(\frac{t}{n}\right) + o(n) \quad := n \cdot n_{i,j}(\tau) + g_{i,j}(n), \tag{8}$$

where $g_{i,j}(n) = o(n)$ is an approximation error. Hence, there must exist some $\Delta_n \to \infty$ with $n \to \infty$, such that $\Delta_n^2 \cdot \max_{i,j\geq 1}\{g_{i,j}(n)\} = o(n)$.

Also observe that the initial sum of degrees of non-pure variables with degree at least Δ_n, which we denote as $C_{\Delta_n} := \sum_{\substack{i,j\geq 1\\ i+j\geq\Delta_n}}(i+j)N_{i,j}(0)$, decreases as we increase Δ_n, and becomes $C_{\Delta_n} = o(n)$ for any $\Delta_n \to \infty$ when $n \to \infty$, since the total number of all unpaired clones at time $t = 0$ is $\sum_{i,j\geq 1}N_{i,j}(t) \leq S_n$, where $S_n = (1+o(1))\gamma n$ w.h.p. by Lemma 2, and $\gamma = \mathbb{E}\xi < \infty$.

Therefore, we can fix some $\Delta_n \to \infty$ such that $\max\left\{C_{\Delta_n}, \Delta_n^2 \cdot \max_{i,j\geq 1}\{g_{i,j}(n)\}\right\} = o(n)$.

Note also that during each iteration of the algorithm, variables only lose their unpaired clones, therefore, $\sum_{\substack{i,j\geq 1\\ i+j\geq\Delta_n}}(i+j)N_{i,j}(t) \leq C_{\Delta_n} = o(n)$ at any time $0 \leq t \leq t_0$.

Next, recall that initially we have S_n unpaired clones, and during each algorithm's iteration we pair exactly k clones.

Hence, at time t we are left with $S_n - kt$ unpaired clones, and $\sum_{i,j\geq 1}(i+j)N_{i,j}(t)$ of them belong to non-pure variables. Therefore, at time t we have

$$\begin{aligned} C(t) &= (S_n - kt) - \sum_{i,j\geq 1}(i+j)N_{i,j}(t) \\ &= (1+o(1))\,\gamma n - kt - \sum_{i,j\geq 1}(i+j)N_{i,j}(t), \end{aligned}$$

since $S_n = (1+o(1))\,\gamma n$ w.h.p. (Lemma 2). Then

$$\begin{aligned} C(t) &= (1+o(1))\,\gamma n - kt - \sum_{i,j\geq 1}(i+j)N_{i,j}(t) \\ &= \gamma n - kt - \sum_{i,j\geq 1}(i+j)N_{i,j}(t) + o(n) \\ &= \gamma n - kt - \sum_{\substack{i,j\geq 1 \\ i+j\leq \Delta_n}}(i+j)N_{i,j}(t) + o(n) \\ &= \gamma n - kt - \sum_{\substack{i,j\geq 1 \\ i+j\leq \Delta_n}}(i+j)N_{i,j}(t) + o(n) \\ &\qquad - \sum_{\substack{i,j\geq 1 \\ i+j> \Delta_n}}(i+j)N_{i,j}(t) \\ &= \gamma n - kt - \sum_{\substack{i,j\geq 1 \\ i+j\leq \Delta_n}}(i+j)N_{i,j}(t) + o(n), \end{aligned}$$

since $\sum_{\substack{i,j\geq 1 \\ i+j\geq \Delta_n}}(i+j)N_{i,j}(t) \leq C_{\Delta_n} = o(n)$ by the choice of Δ_n. Next, we use the approximations (8) for $N_{i,j}(t)$

$$\begin{aligned} C(t) &= \gamma n - kt - \sum_{\substack{i,j\geq 1 \\ i+j\leq \Delta_n}}(i+j)N_{i,j}(t) + o(n) \\ &= \gamma n - kt + o(n) \\ &\qquad - \sum_{\substack{i,j\geq 1 \\ i+j\leq \Delta_n}}(i+j)\left[n\cdot n_{i,j}\left(\frac{t}{n}\right) + g_{i,j}(n)\right] \\ &= \gamma n - kt + o(n) - n\sum_{\substack{i,j\geq 1 \\ i+j\leq \Delta_n}}(i+j)n_{i,j}\left(\frac{t}{n}\right), \end{aligned}$$

since again by the choice of Δ_n

$$\begin{aligned} \sum_{\substack{i,j\geq 1 \\ i+j\leq \Delta_n}}(i+j)g_{i,j}(n) &\leq \sum_{\substack{i,j\geq 1 \\ i+j\leq \Delta_n}}(i+j)\max_{i,j\geq 1}\{g_{i,j}(n)\} \\ &\leq \Delta_n^2 \max_{i,j\geq 1}\{g_{i,j}(n)\} \\ &= o(n). \end{aligned}$$

Therefore, we obtain

$$C(t) = \gamma n - kt + o(n) - n\sum_{\substack{i,j\geq 1 \\ i+j\leq \Delta_n}}(i+j)n_{i,j}\left(\frac{t}{n}\right). \quad (9)$$

We could stop derivation of $C(t)$ here, but to make it much easier to work with, we are going to "extract" dynamics of non-pure variables with degrees at least Δ_n from the $o(n)$ error term. Recall, as was noted at the beginning of the proof, that during any time t the number of unpaired clones of those "heavy" variables is $\sum_{\substack{i,j\geq 1 \\ i+j\geq \Delta_n}}(i+j)N_{i,j}(t) \leq C_{\Delta_n} = o(n)$. Hence, we can add them to (9) without violating the equality

$$C(t) = \gamma n - kt + o(n) - n\sum_{i,j\geq 1}(i+j)n_{i,j}\left(\frac{t}{n}\right),$$

and then

$$\begin{aligned} C(t) &= \gamma n - kt + o(n) - n\sum_{i,j\geq 1}(i+j)n_{i,j}\left(\frac{t}{n}\right) \\ &= n\cdot c\left(\frac{t}{n}\right) + o(n), \end{aligned}$$

where

$$c\left(\frac{t}{n}\right) = \gamma - k\frac{t}{n} - \sum_{i,j\geq 1}(i+j)n_{i,j}\left(\frac{t}{n}\right).$$

Next, recall our convention to call $\tau = \frac{t}{n}$ as scaled time, and so $c\left(\frac{t}{n}\right) = c(\tau)$ can be expressed as

$$\begin{aligned} c(\tau) &= \gamma - k\tau - \sum_{i,j\geq 1}(i+j)n_{i,j}(\tau) \\ &= \gamma - k\tau - \sum_{i,j\geq 1}(in_{i,j}(\tau) + jn_{i,j}(\tau)) \\ &= \gamma - k\tau - \sum_{i,j\geq 1} in_{i,j}(\tau) - \sum_{i,j\geq 1} jn_{i,j}(\tau) \\ &= \gamma - k\tau - \sum_{i,j\geq 1} i\,2^{-(i+j)}\binom{i+j}{i} \\ &\qquad \times \sum_{\ell\geq 0}\binom{\ell}{i+j}z(\tau)^{i+j}(1-z(\tau))^{\ell-(i+j)}p_\ell \\ &\qquad - \sum_{i,j\geq 1} j\,2^{-(i+j)}\binom{i+j}{i} \\ &\qquad \times \sum_{\ell\geq 0}\binom{\ell}{i+j}z(\tau)^{i+j}(1-z(\tau))^{\ell-(i+j)}p_\ell \\ &= \gamma - k\tau - 2\sum_{i,j\geq 1} i\,2^{-(i+j)}\binom{i+j}{i} \\ &\qquad \times \sum_{\ell\geq 0}\binom{\ell}{i+j}z(\tau)^{i+j}(1-z(\tau))^{\ell-(i+j)}p_\ell \\ &= \gamma - k\tau - 2\sum_{j\geq 1}\sum_{i\geq 1} i\binom{i+j}{i}\left(\frac{z(\tau)}{2(1-z(\tau))}\right)^{i+j} \\ &\qquad \sum_{\ell\geq 0}\binom{\ell}{i+j}(1-z(\tau))^{\ell}p_\ell. \end{aligned}$$

Now let $x := \frac{z(\tau)}{2(1-z(\tau))}$. Then

$$
\begin{aligned}
c(\tau) &= \gamma - k\tau - 2\sum_{j\geq 1}\sum_{i\geq 1} i\binom{i+j}{i}\left(\frac{z(\tau)}{2(1-z(\tau))}\right)^{i+j} \\
&\qquad \times \sum_{\ell\geq 0}\binom{\ell}{i+j}(1-z(\tau))^{\ell}p_{\ell} \\
&= \gamma - k\tau - 2\sum_{j\geq 1}\sum_{i\geq 1} i\binom{i+j}{i}x^{i+j} \\
&\qquad \times \sum_{\ell\geq 0}\binom{\ell}{i+j}(1-z(\tau))^{\ell}p_{\ell} \\
&= \gamma - k\tau - 2\sum_{\ell\geq 0}(1-z(\tau))^{\ell}p_{\ell} \\
&\qquad \times \sum_{j\geq 1}\sum_{i\geq 1} i\binom{\ell}{i+j}\binom{i+j}{i}x^{i+j} \\
&= \gamma - k\tau - 2\sum_{\ell\geq 0}(1-z(\tau))^{\ell}p_{\ell} \\
&\qquad \times \sum_{j\geq 1}\sum_{i\geq 1} i\binom{\ell}{j}\binom{d-j}{i}x^{i+j} \\
&= \gamma - k\tau - 2\sum_{\ell\geq 0}(1-z(\tau))^{\ell}p_{\ell} \\
&\qquad \times x\sum_{j\geq 1}\binom{\ell}{j}x^{j}\sum_{i\geq 0} i\binom{\ell-j}{i}x^{i-1} \\
&= \gamma - k\tau - 2\sum_{\ell\geq 0}(1-z(\tau))^{\ell}p_{\ell} \\
&\qquad \times x\sum_{j\geq 1}\binom{\ell}{j}x^{j}\frac{d}{dx}\left[\sum_{i\geq 0}\binom{\ell-j}{i}x^{i}\right] \\
&= \gamma - k\tau - 2\sum_{\ell\geq 0}(1-z(\tau))^{\ell}p_{\ell} \\
&\qquad \times x\sum_{j\geq 1}\binom{\ell}{j}x^{j}\frac{d}{dx}\left[(1+x)^{\ell-j}\right] \\
&= \gamma - k\tau - 2\sum_{\ell\geq 0}(1-z(\tau))^{\ell}p_{\ell} \\
&\qquad \times x\sum_{j\geq 1}(\ell-j)\binom{\ell}{j}x^{j}(1+x)^{(\ell-1)-j} \\
&= \gamma - k\tau - 2\sum_{\ell\geq 0}(1-z(\tau))^{\ell}p_{\ell} \\
&\qquad \times x\sum_{j\geq 1}\ell\binom{\ell-1}{j}x^{j}(1+x)^{(\ell-1)-j} \\
&= \gamma - k\tau - 2\sum_{\ell\geq 0}\ell(1-z(\tau))^{\ell}p_{\ell} \\
&\qquad \times x\left((1+2x)^{\ell-1} - (1+x)^{\ell-1}\right).
\end{aligned}
$$

Since $x = \frac{z(\tau)}{2(1-z(\tau))}$,

$$
\begin{aligned}
c(\tau) &= \gamma - k\tau - 2\sum_{\ell\geq 0}\ell(1-z(\tau))^{\ell}p_{\ell} \\
&\qquad \times x\left((1+2x)^{\ell-1} - (1+x)^{\ell-1}\right) \\
&= \gamma - k\tau - 2\sum_{\ell\geq 0}\ell(1-z(\tau))^{\ell}p_{\ell} \\
&\qquad \times \frac{z(\tau)}{2(1-z(\tau))}\left(1+\frac{z(\tau)}{1-z(\tau)}\right)^{\ell-1} \\
&\qquad + 2\sum_{\ell\geq 0}\ell(1-z(\tau))^{\ell}p_{\ell} \\
&\qquad \times \frac{z(\tau)}{2(1-z(\tau))}\left(1+\frac{z(\tau)}{2(1-z(\tau))}\right)^{\ell-1} \\
&= \gamma - k\tau - z(\tau)\sum_{\ell\geq 0}\ell(1-z(\tau))^{\ell}p_{\ell}\,(1-z(\tau))^{-\ell} \\
&\qquad + z(\tau)\sum_{\ell\geq 0}\ell(1-z(\tau))^{\ell}p_{\ell}\,\frac{\left(1-\frac{z(\tau)}{2}\right)^{\ell-1}}{(1-z(\tau))^{\ell}} \\
&= \gamma - k\tau - z(\tau)\gamma + z(\tau)\sum_{\ell\geq 0}\ell p_{\ell}\left(1-\frac{z(\tau)}{2}\right)^{\ell-1}.
\end{aligned}
$$

What is left is to show that the function $c(\tau)$ can be expressed in terms of probability-generating function $G(x) = \mathbb{E}\left[x^{\xi}\right]$ of the r.v. ξ:

$$
\begin{aligned}
c(\tau) &= \gamma - k\tau - z(\tau)\gamma + z(\tau)\sum_{\ell\geq 1}\ell p_{\ell}\left(1-\frac{z(\tau)}{2}\right)^{\ell-1} \\
&= \gamma - k\tau - z(\tau)\gamma \\
&\qquad - \frac{z(\tau)}{2}\frac{d}{dx}\left[\mathbb{E}\left[\left(1-\frac{x}{2}\right)^{\xi}\right]\right]\Big|_{x=z(\tau)} \\
&= \gamma - k\tau - z(\tau)\gamma - \frac{z(\tau)}{2}\frac{d}{dx}\left[G\left(1-\frac{x}{2}\right)\right]\Big|_{x=z(\tau)},
\end{aligned}
$$

and this completes the proof.

□