



Evidence for Long-Tails in SLS Algorithms

Florian Wörz   

Institut für Theoretische Informatik, Universität Ulm, Germany

Jan-Hendrik Lorenz  

Institut für Theoretische Informatik, Universität Ulm, Germany

Abstract

Stochastic local search (SLS) is a successful paradigm for solving the satisfiability problem of propositional logic. A recent development in this area involves solving not the original instance, but a modified, yet logically equivalent one [23]. Empirically, this technique was found to be promising as it improves the performance of state-of-the-art SLS solvers.

Currently, there is only a shallow understanding of how this modification technique affects the runtimes of SLS solvers. Thus, we model this modification process and conduct an empirical analysis of the hardness of logically equivalent formulas. Our results are twofold. First, if the modification process is treated as a random process, a lognormal distribution perfectly characterizes the hardness; implying that the hardness is long-tailed. This means that the modification technique can be further improved by implementing an additional restart mechanism. Thus, as a second contribution, we theoretically prove that all algorithms exhibiting this long-tail property can be further improved by restarts. Consequently, all SAT solvers employing this modification technique can be enhanced.

2012 ACM Subject Classification Mathematics of computing → Probabilistic algorithms; Mathematics of computing → Distribution functions

Keywords and phrases Stochastic Local Search, Runtime Distribution, Statistical Analysis, Lognormal Distribution, Long-Tailed Distribution, SAT Solving

Digital Object Identifier 10.4230/LIPIcs.ESA.2021.82

Related Version *Full-length Version:* [arXiv:2107.00378](https://arxiv.org/abs/2107.00378) [41]

Supplementary Material See [42]:

Software (Base instances and modifications): <https://doi.org/10.5281/zenodo.4715893>

Software (Visual and statistical evaluations): <https://doi.org/10.5281/zenodo.5026180>

Funding *Florian Wörz:* Supported by the Deutsche Forschungsgemeinschaft (DFG) under project number 430150230, “Complexity measures for solving propositional formulas”.

Acknowledgements The authors acknowledge support by the state of Baden-Württemberg through bwHPC.

1 Introduction

Although algorithms for solving the NP-complete satisfiability problem, so-called SAT solvers, are nowadays remarkably successful in solving large instances, randomized versions of these solvers often show a high variation in the runtime required to solve a fixed instance over repeated runs [16]. In the past, research on randomized algorithms often focused on studying the unsteady behavior of statistical measures like the mean, variance, or higher moments of the runtime over repeated runs of the respective algorithm. In particular, these measures are unable to capture the long-tailed behavior of difficult instances. In a different line of work [13, 15, 32], the focus has shifted to studying the runtime distributions of search algorithms, which helps to understand these methods better and draw meaningful conclusions for the design of new algorithms.



© Florian Wörz and Jan-Hendrik Lorenz;
licensed under Creative Commons License CC-BY 4.0
29th Annual European Symposium on Algorithms (ESA 2021).

Editors: Petra Mutzel, Rasmus Pagh, and Grzegorz Herman; Article No. 82; pp. 82:1–82:16



Leibniz International Proceedings in Informatics

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Recently, the hybrid solver `GapSAT` [23] was introduced, combining a stochastic local search (SLS) solver with a conflict-driven clause learning (CDCL) solver. In the analysis conducted, it was empirically shown that adding new clauses is beneficial to the mean runtime (in flips) of the SLS solver `probsAT` [3] underlying the hybrid model. The authors also demonstrated that although adding new clauses can improve the mean runtime, there exist instances where adding clauses can harm the performance of SLS. This behavior is worth studying to help eliminate the risk of increasing the runtime of such procedures.

For this reason, we study the runtime (or more precisely, hardness) distribution of the procedure `Alfa`, introduced in this work, that models the addition of a set of logically equivalent clauses L to a formula F and the subsequent solving of this amended formula $F^{(1)} := F \cup L$ by an SLS solver. Our empirical evaluations show that this distribution is long-tailed. We want to stress the fact that studies on the runtime distribution of algorithms are quite sparse even though knowledge of the runtime distribution of an algorithm is extremely valuable: (1) Intuitively speaking, if the distribution is long-tailed, one knows there is a risk of ending in the tail and experiencing very long runs; simultaneously, the knowledge that the time the algorithm used thus far is in the tail of the distribution can be exploited to restart the procedure (and create a new logically equivalent instance $F^{(2)}$). We will prove this statement in a rigorous manner for all long-tailed algorithms. (2) Given the distribution of an algorithm’s sequential runtime, it was shown how to predict and quantify the algorithm’s expected speedup due to parallelization [1]. (3) If the distribution of hardness is known, experiments with few instances can lead to parameter estimations of the underlying distribution [13]. (4) Knowledge of the distribution can help compare competing algorithms: one can test if the difference in the means of algorithm runtimes is significant if the distributions are known [13].

1.1 Our Contributions

Our contributions consist of an empirical as well as theoretical part, specified below.

Statistical Runtime/Hardness Distribution Analysis. By conducting a plethora of experiments (total CPU time 80 years) and using several statistical tools for the analysis of empirical distributions, we conjecture that `Alfa` equipped with SLS solvers based on Schönig’s Random Walk Algorithm [36], `SRWA` for short, follows a long-tailed distribution (Conjecture 8). The evidence obtained further suggests that this distribution is, in fact, lognormal (Conjecture 6). We measure the goodness-of-fit of our results over the whole domain using the χ^2 -statistic.

Restarts Are Useful For Long-Tailed Algorithms. Lorenz [22] has analyzed the lognormal and the generalized Pareto distribution for the usefulness of restarts and their optimal restart times. Given that our Strong Conjecture 6 holds, this result implies that restarts are useful for `Alfa`. We will also show that this is the case if only the Weak Conjecture 8 holds: We theoretically prove that restarts are useful for the class of algorithms exhibiting a long-tailed distribution.

1.2 Related Work and Differentiation

In [13], the authors presented empirical evidence for the fact that the distribution of the effort (more precisely, the number of consistency checks) required for backtracking algorithms to solve constraint satisfaction problems randomly generated at the 50% satisfiable point

can be approximated by the Weibull distribution (in the satisfiable case) and the lognormal distribution (in the unsatisfiable case). These results were later extended to a wider region around the 50% satisfiable point [32]. It should be emphasized that this study created all instances using the same generation model. This resulted in the creation of similar yet logically non-equivalent formulas. We, however, will firstly use different models to rule out any influence of the generation model and secondly generate logically equivalent modifications of a base instance (see Algorithm 1). This approach lends itself to the analysis of existing SLS solvers [23]. The major advantage is that the conducted work is not lost in the case of a restart: only the logically equivalent instance could be changed while keeping the current assignment.

In [16], the cost profiles of combinatorial search procedures were studied. The authors showed that they are often characterized by the Pareto-Lévy distribution and empirically demonstrated how rapid randomized restarts can eliminate this tail behavior. We will theoretically prove the effectiveness of restarts for the larger class of long-tailed distributions.

The paper [1] studied the solvers **Sparrow** and **CCASAT** and found that for randomly generated instances the lognormal distribution is a good fit for the runtime distributions. For this, the Kolmogorov-Smirnov statistic $\sup_{t \in \mathbb{R}} |\hat{F}_n(t) - F(t)|$ was used. Although the KS-test is very versatile, this comes with the disadvantage that its statistical power is rather low. Clearly, the KS statistic is also nearly useless in the tails of a distribution: A high relative deviation of the empirical from the theoretical cumulative distribution function in either tail results in a very small absolute deviation. It should also be remarked that the paper studies only few formulas in just two domains, 10 randomly generated and 9 crafted. Our work will address both shortcomings in this paper: The χ^2 -test gives equal importance to the goodness-of-fit over the entire support; and various instance domain models (both theoretical and applied) are considered in this paper.

► **Remark.** Unfortunately, the term heavy- or long-tailed distribution is not used consistently in the literature. We will follow [12] and use the notion given in Definition 7.

2 Preliminaries

We assume familiarity with terminologies such as Boolean variable, literal, clause, CNF formula, the SAT problem, and assignment flips in SLS solvers and refer the reader to e.g. [37]. We furthermore trust that the reader has basic knowledge of the proof system Resolution [8, 33]. *Stochastic local search* (SLS) solvers operate on complete assignments for a formula F . These solvers are started with a randomly generated complete initial assignment α_0 . If α_0 satisfies F , a solution is found. Otherwise, the SLS solver tries to find a solution by performing a random walk over the set of complete assignments for the underlying formula. A formula F *logically implies* a clause C if every complete truth assignment which satisfies F also satisfies C , for which we write $F \models C$. If L is a set of clauses we write $F \models L$ if $F \models C$ for all $C \in L$.

► **Definition 1.** Let X be a random variable for the runtime of an SLS algorithm \mathcal{A} on some input. For $t > 0$, the algorithm \mathcal{A}_t is obtained by restarting \mathcal{A} after time t if no solution was found. Restarts are useful if there is a $t > 0$ such that

$$E[X_t] < E[X],$$

where X_t models the runtime of \mathcal{A}_t .

► **Definition 2** ([19]). Let X be a real-valued random variable.

■ Its cumulative distribution function (cdf) is the function $F: \mathbb{R} \rightarrow [0, 1]$ with

$$F(t) := \Pr[X \leq t].$$

82:4 Evidence for Long-Tails in SLS Algorithms

- Its quantile function $Q: (0, 1) \rightarrow \mathbb{R}$ is given by $Q(p) := \inf\{t \in \mathbb{R} \mid F(t) \geq p\}$.
- A non-negative, integrable function f such that $F(t) = \int_{-\infty}^t f(u) du$ is called probability density function (pdf) of X .

► **Definition 3** ([40]). An absolutely continuous, positive random variable X is (three-parameter) lognormally distributed with parameters $\sigma^2 > 0$, $\gamma > 0$, and $\mu \in \mathbb{R}$, if $\log(X - \gamma)$ is normally distributed with mean μ and variance σ^2 . In the following, we refer to σ as the shape, μ as the scale, and γ as the location parameter.

► **Definition 4.** Let X_1, \dots, X_n be independent, identically distributed real-valued random variables with realizations x_i of X_i . Then the empirical cumulative distribution function (ecdf) of the sample (x_1, \dots, x_n) is defined as

$$\hat{F}_n(t) := \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{x_i \leq t\}}, \quad t \in \mathbb{R},$$

where $\mathbb{1}_A$ is the indicator of event A .

3 Design of the Adjusted Logical Formula Algorithm Alfa

Our SLS solver **Alfa** (Adjusted logical formula algorithm) receives a satisfiable formula F as input. The algorithm then proceeds by adding to F a set L of logically generated clauses. It finally calls an SLS solver to solve the clause set $F \cup L$.

■ **Algorithm 1** Alfa acts as a base algorithm that can use different SLS algorithms.

Input: Boolean formula F , **Promise:** $F \in \text{SAT}$
 Generate **randomly** a set L of clauses such that $F \models L$
 Call $\text{SLS}(F \cup L)$ for some SLS solver SLS

Definition 5 is used in Algorithm 2 as a natural way to sample a set L of logically equivalent clauses with respect to a base instance F .

■ **Algorithm 2** Generation of the random set L with resolution.

Input: Boolean formula F , integer w , probability $p \in (0, 1]$, Boolean *shuffle*
foreach $R \in \text{Res}_w^*(F) \setminus F$ **do**
 | **with probability** p **do** $L := L \cup \{R\}$
if *shuffle* **then return** $\text{Shuffle}(L)$ **else return** L ;

► **Definition 5.** Let F be a clause set, and w be a positive integer. We define the operator

$$\text{Res}_w(F) := F \cup \{R \mid R \text{ is a resolvent of two clauses in } F \text{ and } |R| \leq w\}.$$

Also, we inductively define $\text{Res}_w^0(F) := F$ and

$$\text{Res}_w^{n+1}(F) := \text{Res}_w(\text{Res}_w^n(F)), \text{ for } n \geq 0.$$

Finally, we set

$$\text{Res}_w^*(F) := \bigcup_{n \geq 0} \text{Res}_w^n(F).$$

4 Empirical Evaluation

4.1 Experimental Setup, Instance Types, and Solvers Used

Hoos and Stützle [18] introduced the concept of *runtime distribution* to characterize the cdf of Las Vegas algorithms, where the runtime can vary from one execution to another, even with the same input. To obtain enough data for a fitting of such a distribution, for each base instance F we created 5000 modified instances $F^{(1)}, \dots, F^{(5000)}$ by generating resolvent sets $L^{(1)}, \dots, L^{(5000)}$ each by using Algorithm 2 with $w = 4$ and a value of p such that the expected number of resolvents being added was $\frac{1}{10}|F|$. Note that we also conducted a series of experiments to rule out the influence of p on our results. Each of these modified instances was solved 100 times, each time using a different seed. For $i = 1, \dots, 5000$ and $j = 1, \dots, 100$ we thus obtained the values $\text{flips}_S(F^{(i)}, s_j)$ indicating how many flips were used to solve the modified instance $F^{(i)}$ with solver S when using the seed s_j . Next, we calculated the mean number of flips $\text{mean}_S(F^{(i)}) := \frac{1}{100} \sum_{j=1}^{100} \text{flips}_S(F^{(i)}, s_j)$ required to solve $F^{(i)}$ with solver S whose hardness distribution we are going to analyze.

All experiments were performed on bwUniCluster 2.0 and three local servers. Sputnik [39] was used to distribute the computation and to parallelize the trials. Due to the heterogeneity of the computer setup, measured runtimes are not directly comparable to each other. Consequently, we instead measured the number of variable flips performed by the SLS solver. This is a hardware-independent performance measure with the benefit that it can also be analyzed theoretically. To give an indication of how flips relate to wall-clock time, one million flips take about one second of computing time on one of our servers. To give an idea of the computational effort involved, obtaining the data for the ecdf of a 100 variable base instance with SRWA took an average of 17,193,517 seconds (≈ 199 days) when unparallelized. This clearly prohibited examining instances having a number of variables currently being routinely solved by the state-of-the-art SLS algorithms. For the experiments the following instance types were used:

1. **Hidden Solution:** We implemented the CDC algorithm [5, 6] in [24] to generate instances with a hidden solution. For this, at the beginning, a complete assignment α is specified to ensure the generated formula's satisfiability. Then, repeatedly a randomly generated clause C is added to the formula with a weighted probability p_i depending on the number i of correct literals in C with respect to α . We included this type of instances because SLS solvers struggle to solve such instances. Experiments like these might be beneficial to find theoretical reasons for this behavior.
2. **Hidden Solution With Different Chances:** We also created sets of formulas with different underlying p_i values to rule out the influence of these.
3. **Uniform Random:** To generate uniform, random k -SAT instances with n variables and m clauses, each clause is generated by sampling k literals uniformly and independently. Using Gableske's `kcnfgen` [14], we generated formulas with $n \in \{50, 60, 70, 80, 90\}$ variables and a clause-to-variable ratio r close to the *satisfiability threshold* [29] of $r \approx 4.267$. We checked each instance with `Glucose3` [2, 11] for satisfiability until we had 5 formulas of each size.
4. **Factoring:** These formulas encode the factoring problem in the interval $\{128, \dots, 256\}$ and were generated with [10].
5. **Coloring:** These formulas assert that a graph is colorable with 3 colors. We generated these formulas, using [21], over random graphs with n vertices and $m = 2.254n$ edges in expectation, which is slightly below the *non-colorability threshold* [20]. We obtained 32 satisfiable instances in 150 variables.

Our experiments investigated leading SLS solvers where the dominating component is based on the random walk procedure proposed in [36]. In this paper, Schönig’s Random Walk Algorithm **SRWA** was introduced, which is one of the solvers we used. The **probsAT** solver family [3] is based on this approach. One of these solvers won the random track of the SAT competition 2013 [4]. Another advancement of **SRWA** was implemented as **YALSAT** [7], which won the random track of the SAT competition 2017 [17]. These performances and similarities were reasons for choosing **SRWA**, **probsAT**, and **YALSAT** as SLS solvers for this paper. The connection to **GapSAT** [23] is another case in point.

We excluded the solvers **DCCAlm** [25] and **CSCCSat** [28] (combining **FrwCB** [26] and **DCCASat** [27]) as all of these depend on heuristics (like **CC**, **BM**, **CSDvars**, **NVDvars**, **SDvars**) that ultimately reduce the probabilistic nature when choosing the next variable to flip.

For **SRWA** we conducted most of our experiments: All instance types were tested, including different change values for the generation of the hidden solution. For **probsAT**, 55 hidden solution instances with $n \in \{50, 100, 150, 200, 300, 800\}$ were used. Since **YALSAT** can be regarded as a **probsAT** derivative, we tested **YALSAT** with 10 hidden solution instances with 300 variables each.

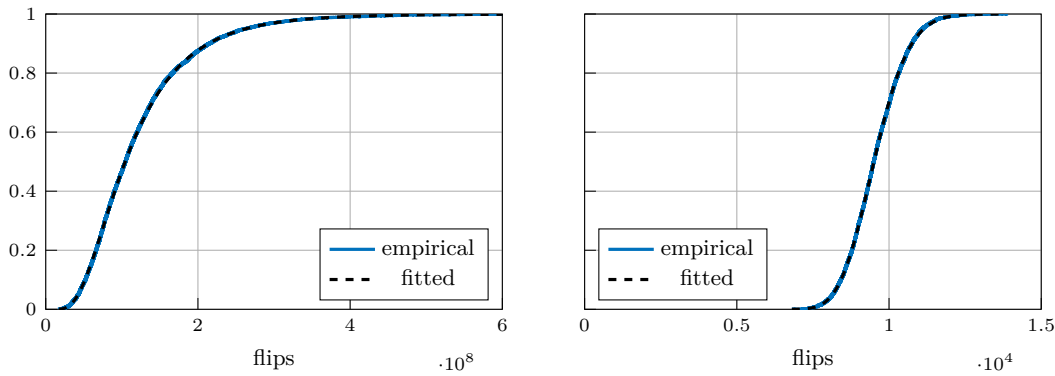
4.2 Experimental Results and Statistical Evaluation

The goal of this section is to explore the scenarios described above in more detail. We are particularly interested in how the hardness of an instance changes when logically equivalent clauses are added in the manner described above. To characterize this effect as accurately as possible, studying the ecdf is the most suitable method for this purpose. In turn, the ecdf can be described using well-known distribution types such as e. g., the normal distribution. In the following, we shall demonstrate that the three-parameter lognormal distribution, in particular, provides an exceptionally accurate description of the runtime behavior, and this is true for all considered problem domains and all solvers. The results are so compelling that we ultimately conjecture that the runtimes of **Alfa**-type algorithms all follow a lognormal distribution, regardless of the considered problem domain.

To illustrate this point, we first demonstrate our approach using two base instances. The first one is a factorization instance that was solved by **SRWA**. The second instance has a hidden solution and was solved by **probsAT**. For later reference, we refer to the first instance as *A* and to the second instance as *B*. As described above, we obtain 5000 samples for each base instance. Using these data points, we estimate the lognormal distribution’s three parameters by applying the maximum likelihood method (see [42]). After that, one can visually evaluate the suitability of the fitted lognormal distribution for describing the data. A useful method of visualizing the suitability is to plot the ecdf and the fitted cdf on the same graph.

Such a comparison is illustrated in Figure 1 for the two instances *A* and *B*. In both cases, no difference between the empirical data of the ecdf and the fitted distribution can be detected visually. In other words, the absolute error between the predicted probabilities from the fitted cdf versus the empirical probabilities from the ecdf is minuscule. Even though these are only two examples, it should be noted that these two instances are representative of the behavior of the investigated algorithms. Hardly any deviation could be observed in this plot type for all instances and all algorithms. All data is published under [42].

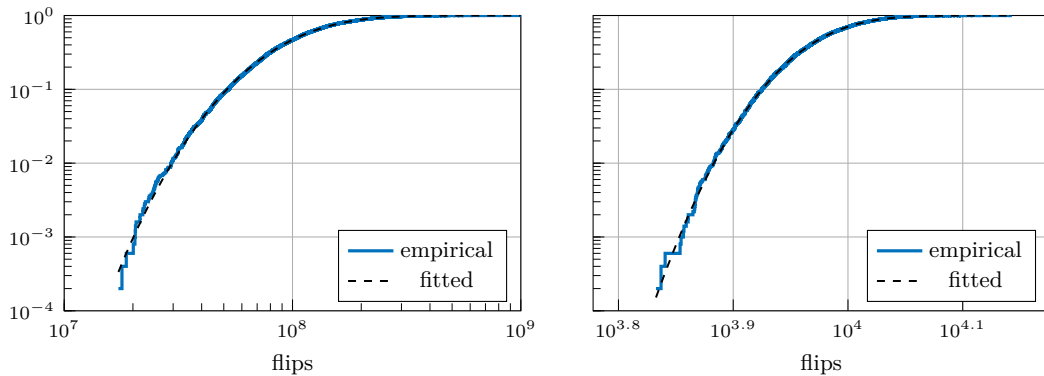
For the analysis, however, one should not confine oneself to this plot type. Although absolute errors can be observed easily, relative errors are more difficult to detect. Such a relative error may have a significant impact when used for decisions such as restarts. To illustrate this point, suppose that the true probability of a run of length ℓ is 0.0001. In contrast, the probability estimated based on a fit is 0.001. As can be seen, the absolute error



■ **Figure 1** The ecdf and fitted cdf of the hardness distribution of instance *A* (left) and *B* (right).

of 0.0009 is small, whereas the relative error of 10 is large. If one were to perform restarts after ℓ steps, the actual expected runtime would be ten times greater than the estimated expected runtime. Thus, the erroneous estimate of that probability would have translated into an unfavorable runtime. This example should illustrate the importance of checking the tails of a distribution for errors as well.

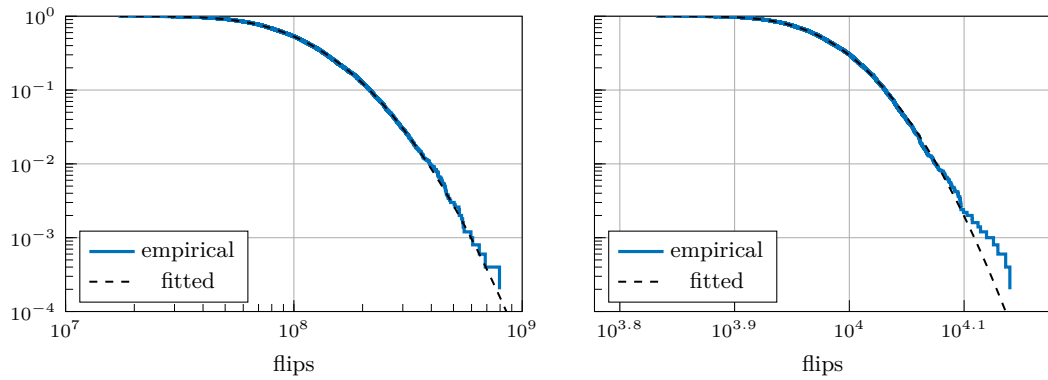
The left tail, i. e., the probabilities for very small values, can be checked visually by plotting the ecdf and fitted cdf with both axes logarithmically scaled. Thereby, the probabilities for extreme events (in this case, especially easy instances) can be measured accurately.



■ **Figure 2** Logarithmically scaled ecdf and fitted cdf of instances *A* (left) and *B* (right).

The two instances *A* and *B* are being examined in this manner in Figure 2. As can be observed, the lognormal fit accurately predicts the probabilities associated with very short runs. For the other instances, lognormal distributions were mostly also able to accurately describe the probabilities for short runs. However, the behavior of the ecdf and the fitted lognormal distribution differed very slightly in a few instances.

Lastly, the probabilities for particularly hard instances should also be checked. Any mistakes in this area could lead to underestimating the likelihood of encountering an exceptionally hard instance. For analyses of this type, the survival function S is a useful tool; if F is the cdf, $S(x) := 1 - F(x)$. Therefore, the survival function's value $S(x)$ represents the probability that an instance is (on average) harder than x in our case. If we plot the empirical survival function, i. e., $\hat{S}_n(x) := 1 - \hat{F}_n(x)$, and the fitted survival function together on a graph with logarithmically scaled axes, we can easily detect errors in the right tail.



■ **Figure 3** Logarithmically scaled empirical survival function and fitted survival function of instances A (left) and B (right).

Figure 3 illustrates this type of plot for the instances A and B . Here, there is a discernible deviation between A and B . While for A , the lognormal fit provides an accurate description of the probabilities for long runs, in the case of B , the empirical survival function seems to approach 0 somewhat slower than the lognormal estimate. In the vast majority of cases, these extreme value probabilities are accurately reflected by the lognormal fit. In most other cases, the empirical survival function approaches 0 more slowly than the lognormal fit. Thus, in these cases, the likelihood of encountering an exceptionally hard instance is underestimated.

So far, we discussed the behavior of lognormal fits based on this visual inspection. Altogether, we concluded that lognormal distributions seem to be well suited for describing the data. Next, we shall concretize this through a statistical test. To be more precise, we apply the χ^2 -test as a goodness-of-fit test for each base instance. For each such instance, the fitted lognormal distribution used the 5000 data points, and afterwards, the χ^2 -test statistic is computed. Subsequently, the probability that such a value of the test statistic occurs under the assumption of the so-called null hypothesis is determined. We will refer to this probability as the p -value. In our case, the null hypothesis is the assumption that the data follow a lognormal distribution. If the fit is poor, then a small p -value will occur. If p is sufficiently small, the null hypothesis is rejected. We reject the null hypothesis if $p < 0.05$.

Two more remarks are due on this matter. First, from a high p -value, one cannot prove that the assumption that the data are lognormally distributed is correct. However, we use a sufficiently high p -value as a heuristic whether this assumption is reasonable.

Secondly, there is an obstacle that complicates statistical analysis by this method. As described, each of the 5000 data points is obtained by first sampling 100 runtimes of the corresponding instance and then calculating the mean. This means that we do not work with the actual expected values but only estimates. In other words, this implies that our data is noisy. The greater the variance in the respective instance, the greater the corresponding noise. If one were to apply the χ^2 -test to this noisy data, some cases would be incorrectly rejected, especially if the variance is large. To overcome this limitation, we additionally use a bootstrap-test, which is based on Cheng [9]. This test is presented in Algorithm 3.

Briefly summarized, this test simulates how our data points were generated, assuming the null hypothesis. For this purpose, particular attention should be paid to how the test sample is rendered noisy. Owing to the central limit theorem, it is reasonable to assume that the initial data's sample mean originates from a normal distribution around the true expected value. We use this assumption in the bootstrap-test using a noise signal drawn from a normal distribution with expected value 0. The variance of this normal distribution is determined from the initial data and divided by 100 (cf. central limit theorem).

■ **Algorithm 3** Bootstrap-test for noisy data.

Input: (noisy) random sample $\mathbf{y} = (y_1, y_2, \dots, y_n)$, integer N , significance $\alpha \in (0, 1)$

$\hat{\theta} \leftarrow \text{MLE}(\mathbf{y}, F)$, lognormal maximum likelihood estimation, F is the lognormal cdf

$X^2 \leftarrow \text{ChiSquare}(\mathbf{y}, \hat{\theta})$, Chi-squared goodness of fit test statistic

for $j = 1$ **to** N **do**

$\mathbf{y}' \leftarrow (y'_1, \dots, y'_n)$, where all y'_i are i. i. d. samples from the fitted lognormal distribution with parameters $\hat{\theta}$

$\mathbf{y}' \leftarrow \mathbf{y}' + \text{noise}$, where noise is sampled from an n -dimensional normal distribution

$\hat{\theta}' \leftarrow \text{MLE}(\mathbf{y}', F)$

$X_j^2 \leftarrow \text{ChiSquare}(\mathbf{y}', \hat{\theta}')$

Let $X_{(1)}^2 \leq X_{(2)}^2 \leq \dots \leq X_{(N)}^2$ be the sorted test statistics.

if $X_{(\lfloor (1-\alpha) \cdot N \rfloor)}^2 < X^2$ **then reject else accept;**

If there is a large difference between the respective p -values of the χ^2 - and bootstrap-tests, this suggests that the variance in the initial data is too high and that the number of samples used to calculate the sample mean should be increased. However, this case has only occurred twice, and we explicitly indicate it later. In all other cases, the p -values of the χ^2 - and the bootstrap-tests were similar. To illustrate this point, let us consider the p -values of the two instances A and B . The excellent representation of the runtime behavior over the entire support on instance A is also reflected in the two p -values. Using the χ^2 -test, one obtains a p -value of approximately 0.783, and using the bootstrap-test, one obtains a p -value of 0.76. Since these two p -values are both above 0.05, we conclude that the assumption that lognormal distributions describe the runtimes is reasonable.

In contrast, for instance B , we observed that while a lognormal distribution accurately describes the main part of the distribution, the probabilities for extremely long runs are inadequately represented. This observation is again reflected in the respective p -values. The χ^2 -test yields a p -value of ≈ 0.008 , and the bootstrap-test yields a p -value of 0.013. Since both p -values are below 0.05, the assumption that these data originate from a lognormal distribution is rejected.

Overall, this example is intended to demonstrate that these two statistical tests can show an inadequate fit, even if the problems only arise for extreme values. Second, it should demonstrate that the p -values of the two tests are generally similar; if the p -values differ significantly, then more samples should be used to calculate the sample mean.

We now proceed by considering the adequacy of lognormal distributions for describing SRWA runtimes. The results of the statistical analysis are reported in Table 1 and can be found in [42].

■ **Table 1** Statistical goodness-of-fit results for Alfa+SRWA runtimes over various problem domains. The *rejected* row contains the number of instances where the lognormal distribution is not a good fit according to the χ^2 -test at a significance level of 0.05. To put these results into perspective, the second row contains the total number of instances of each domain. Out of a total of 230 instances, 5 got rejected.

	hidden	different chances	uniform	factoring	coloring	total
rejected	0	2	1	2	0	5
# of instances	20	120	25	33	32	230

82:10 Evidence for Long-Tails in SLS Algorithms

The first line in the table represents for how many instances the statistical tests rejected the lognormal distribution hypothesis. The second line indicates how many instances have been checked in total. It should be noted that the same number of instances was rejected by the χ^2 -test as was by the bootstrap-test. Thus, there is no need to distinguish between the tests here. It should also be mentioned that in statistical tests, there is always a possibility that a hypothesis will be rejected even though the null hypothesis holds (type 1 error). At a significance level of 0.05, this probability is at the most 5%. Accordingly, the total of 5 rejected instances may be attributed to so-called type 1 errors. This statement is also supported by the fact that no exceptionally low p -value was observed, i. e., no p -value that is unusual for a total of 230 samples.

For `probSAT`, on the other hand, the situation appears to be different. The results are summarized in Table 2 and [42]. The columns refer to the number of variables in the

■ **Table 2** Goodness-of-fit results for `Alfa+probSAT` over various hidden solution instance sizes. The *rejected* row contains the number of instances where the lognormal distribution is not a good fit according to the χ^2 -test at a significance level of 0.05. To put these results into perspective, the second row contains the total number of instances of each instance size.

number of variables	50	100	150	200	300	800	total
rejected	2	2	1	0	2	0	7
# of instances	10	10	10	10	10	5	55

corresponding SAT instances. The number of rejected instances is again identical regardless of whether the χ^2 - or bootstrap-test is applied. As can be seen, the lognormal distribution hypothesis was rejected for 7 of the 55 instances. This number can no longer be accounted for by type 1 errors at a significance level of 0.05. However, one can observe that the majority of rejected instances occur for a small number of variables. If one were to consider only the instances from 150 variables onwards, then the remaining rejected instances may be attributed to type 1 errors. This raises the suspicion that there may be a limiting process, i. e., that the lognormal distribution hypothesis is only valid for $n \rightarrow \infty$.

Lastly, a difference between the two static tests emerges for `YALSAT`. According to the χ^2 -test, 2 of the total 10 instances are rejected. However, using the bootstrap-test, the lognormal distribution hypothesis is not rejected for any instance. Therefore, one cannot rule out the possibility that lognormal distributions are the natural model to describe the instances, but more experiments are required to make a more precise statement.

In summary, the presumption that lognormal distributions are the appropriate choice for describing runtimes has been reinforced for `SRWA`. For `probSAT`, this appears plausible at least above a certain instance size. Likewise, the choice of lognormal distributions also seems reasonable for `YALSAT`. These observations lead us to the following conjecture.

► **Conjecture 6** (Strong Conjecture). *The runtime of `Alfa` with $SLS \in \{\text{SRWA}, \text{probSAT}, \text{YALSAT}\}$ follows a lognormal distribution.*

If this statement is true, then it would be intriguing in that one can infer how modifying the base instance affects the hardness of instances. This effect is likely to be the result of generating models for lognormal distributions. Just as the normal distribution is a natural model for the sum of i. i. d. random variables, the lognormal distribution is a natural model for the multiplication of i. i. d. random variables. Thus, one can hypothesize that each added clause exerts a small multiplicative effect on the instance's hardness.

Simultaneously, the three parameters of the lognormal distribution also provide insight into how the hardness of the instance changes. For example, the location parameter γ implies an inherent problem hardness that cannot be decreased regardless of the added clauses'

choice. At the same time, γ also serves as a numerical description for the value of this intrinsic hardness. Using Bayesian statistics, it is possible to infer the parameters while the solver is running. These estimated parameters can, for example, be used to schedule restarts. This would lead to a scenario similar to that discussed in [34].

Conjecture 6 is a strong statement. However, a small deviation of the probabilities, for example, at the left tail, would render the strong conjecture invalid from a strict mathematical point of view. Particularly, visual analyses revealed that the left tail's behavior, i. e., for extremely short runs, is occasionally not accurately reflected by lognormal distributions. Conversely, the right tail, i. e., the probabilities for particularly long runs, are usually either correctly represented by lognormal distributions or, occasionally, the corresponding probability approaches 0 even more slowly. We, therefore, rephrase our conjecture in a weakened form. Our observations fit a class of distributions known as long-tail distributions defined purely in terms of their behavior at the right tail.

► **Definition 7** ([12]). *A positive, real-valued random variable X is long-tailed, if and only if*

$$\forall x \in \mathbb{R}^+ : \Pr[X > x] > 0 \quad \text{and} \quad \forall y \in \mathbb{R}^+ : \lim_{x \rightarrow \infty} \frac{\Pr[X > x + y]}{\Pr[X > x]} = 1.$$

► **Conjecture 8** (Weak Conjecture). *The runtime of Alfa with $SLS \in \{SRWA, probSAT, YaLSAT\}$ follows a long-tailed distribution.*

It should be noted that lognormal distributions have the long-tail property [12, 30]. That is, if the Strong Conjecture holds, the Weak Conjecture is implied. The reverse is, however, not true. In the next section, we show an important consequence in case the Weak Conjecture holds.

5 Restarts Are Useful For Long-Tailed Distributions

If the Strong Conjecture holds, i. e., if the runtimes are lognormally distributed, then restarts are useful [22]. This section extends this result and mathematically proves that restarts are useful even if only the Weak Conjecture holds. This will be achieved by showing that restarts are useful for long-tailed distributions.

A condition for the usefulness of restarts, as defined in Definition 1, was proven in [22]. We will show the result using this theorem that is restated below.

► **Theorem 9** ([22]). *Let X be a positive, real-valued random variable having quantile function Q , then restarts are useful if and only if there is a quantile $p \in (0, 1)$ such that*

$$R(p, X) := (1 - p) \cdot \frac{Q(p)}{\mathbb{E}[X]} + \frac{\int_0^p Q(u) \, du}{\mathbb{E}[X]} < p.$$

Even if the quantile function and the expected value are unknown, $R(p, X)$ can be characterized for large values of p .

► **Lemma 10.** *Consider a positive, real-valued random variable X with pdf f and quantile function Q such that $\mathbb{E}[X] < \infty$. Also, assume that the limit $\lim_{t \rightarrow \infty} t^2 \cdot f(t)$ exists. Then,*

$$\lim_{p \rightarrow 1} R(p, X) = \lim_{p \rightarrow 1} \left((1 - p) \cdot \frac{Q(p)}{\mathbb{E}[X]} + \frac{\int_0^p Q(u) \, du}{\mathbb{E}[X]} \right) = 1.$$

82:12 Evidence for Long-Tails in SLS Algorithms

Proof. In the following, let F and f be the cdf and pdf of X , respectively. We start by specifying the derivative of Q with respect to p as a preliminary consideration. From $F = Q^{-1}$ and the application of the inverse function theorem [35] it follows:

$$Q'(p) := \frac{d}{dp}Q(p) = \frac{1}{f(Q(p))}. \quad (1)$$

As the first step in our proof, we consider the limiting value of the second summand of $R(p, X)$. This value can be determined by integration by substitution with $x = Q(u)$ followed by applying the change of variable method with $p = F(t)$:

$$\lim_{p \rightarrow 1} \frac{\int_0^p Q(u) du}{E[X]} = \lim_{p \rightarrow 1} \frac{\int_0^{Q(p)} x \cdot f(x) dx}{E[X]} = \lim_{t \rightarrow \infty} \frac{\int_0^t x \cdot f(x) dx}{E[X]} = 1.$$

The last equality holds because the numerator matches the definition of the expected value.

Next, we examine the limit of $(1-p)Q(p)/E[X]$. Since $\lim_{p \rightarrow 1}(1-p) = 0$, the limit of $(1-p) \cdot Q(p)/E[X]$ needs to be examined more closely. For this purpose, L'Hospital's rule is applied twice as well as the change of variable method with $p = F(t)$ is used in the following:

$$\lim_{p \rightarrow 1} (1-p) \cdot Q(p) = \lim_{p \rightarrow 1} Q(p)^2 \cdot f(Q(p)) = \lim_{t \rightarrow \infty} t^2 \cdot f(t).$$

It is well-known that if $\liminf_{t \rightarrow \infty} t^2 \cdot f(t) > 0$ were to hold, then the expected value $E[X]$ would be infinite (this statement is, for example, implicitly given in [12]). This would contradict the premise of the lemma; therefore, $\liminf_{t \rightarrow \infty} t^2 \cdot f(t) = 0$. Moreover, since, by assumption, $\lim_{t \rightarrow \infty} t^2 \cdot f(t)$ exists, we may conclude that

$$\lim_{t \rightarrow \infty} t^2 \cdot f(t) = \limsup_{t \rightarrow \infty} t^2 \cdot f(t) = \liminf_{t \rightarrow \infty} t^2 \cdot f(t) = 0. \quad \blacktriangleleft$$

A frequently used tool for the description of distributions is the hazard rate function.

► **Definition 11** ([31]). *Let X be a positive, real-valued random variable having cdf F and pdf f . The hazard rate function $r: \mathbb{R}^+ \rightarrow \mathbb{R}^+$ of X is given by*

$$r(t) := \frac{f(t)}{1 - F(t)}.$$

In particular, there is an interesting relationship between the long-tail property and the hazard rate function's behavior.

► **Lemma 12** ([30]). *Let X be a positive, real-valued random variable with hazard rate function r such that the limit $\lim_{t \rightarrow \infty} r(t)$ exists. Then, the following three statements are equivalent:*

1. X is long-tailed.
2. $\lim_{x \rightarrow \infty} \int_x^{x+y} r(t) dt = 0, \quad \forall y > 0.$
3. $\lim_{t \rightarrow \infty} r(t) = 0.$

Proof. The full length-version of this paper [41] contains a proof of this lemma since the manuscript [30] was still unpublished at the time of writing. ◀

With the help of these preliminary considerations, we are now ready to show that restarts are useful for long-tailed distributions.

► **Theorem 13.** *Consider a positive, long-tailed random variable X with continuous pdf f and hazard rate function r . Also assume that either $E[X] = \infty$ holds or the limits $\lim_{t \rightarrow \infty} r(t)$ and $\lim_{t \rightarrow \infty} t^2 \cdot f(t)$ both exist. In both cases, restarts are useful for X .*

Proof. Let F be the cdf and Q the quantile function of X . We begin with the case $E[X] = \infty$. According to Theorem 9, restarts are useful if and only if

$$(1-p) \cdot \frac{Q(p)}{E[X]} + \frac{1}{E[X]} \cdot \int_0^p Q(u) \, du < p$$

for some $p \in (0, 1)$. However, if the expected value $E[X]$ is infinite, then the left side of this inequality is zero and the inequality is obviously satisfied. Hence, the statement follows.

Secondly, we assume that $E[X] < \infty$ and that both $\lim_{t \rightarrow \infty} r(t)$ and $\lim_{t \rightarrow \infty} t^2 \cdot f(t)$ exist. Equation (1) can now be used to calculate the following derivative:

$$\frac{d}{dp} (R(p, X) - p) = \frac{d}{dp} \left((1-p) \cdot \frac{Q(p)}{E[X]} + \frac{\int_0^p Q(u) \, du}{E[X]} - p \right) = \frac{1-p}{E[X] \cdot f(Q(p))} - 1.$$

Consider the limit of this expression for $p \rightarrow 1$. Once again, the change of variable method is applied with $p = F(t)$, resulting in:

$$\lim_{p \rightarrow 1} \frac{1-p}{E[X] \cdot f(Q(p))} - 1 = \lim_{t \rightarrow \infty} \frac{1-F(t)}{E[X] \cdot f(t)} - 1 = \lim_{t \rightarrow \infty} \frac{1}{E[X] \cdot r(t)} - 1.$$

By assumption, X has a long-tail distribution and the limit of $\lim_{t \rightarrow \infty} r(t)$ exists. For this reason, $\lim_{t \rightarrow \infty} r(t) = 0$ follows as a result of Lemma 12. Furthermore, since $E[X] < \infty$ holds, we may conclude that

$$\lim_{p \rightarrow 1} \frac{1-p}{E[X] \cdot f(Q(p))} - 1 = \lim_{t \rightarrow \infty} \frac{1}{E[X] \cdot r(t)} - 1 = \infty. \quad (2)$$

The condition from Theorem 9 can be rephrased in such a way that restarts are useful if and only if $R(p, X) - p < 0$. According to Lemma 10, the left-hand side of this inequality approaches 0 for $p \rightarrow 1$. However, as has been shown in Equation (2), the derivative of $R(p, X) - p$ approaches infinity for $p \rightarrow 1$. These two observations imply that there is a $p \in (0, 1)$ satisfying $R(p, X) - p < 0$. Consequently, restarts are useful for X . ◀

It should be noted that the conditions of this theorem are not restrictive since all naturally occurring long-tail distributions satisfy these conditions (see also [30]).

► **Conjecture 14** (Corollary of the Weak Conjecture). *Restarts are useful for Alfa with $SLS \in \{SRWA, probSAT, YaLSAT\}$.*

If Conjecture 8 is true, then this statement follows immediately by Theorem 13.

6 Conclusion

We have provided compelling evidence that the runtime of **Alfa** follows a long-tailed or lognormal distribution. According to [38], the usefulness of restarts is a necessary, however not a sufficient, condition to obtain super-linear speedups by parallelization. Since we have shown that the necessary condition is (presumably) satisfied, this immediately raises the question of whether super-linear speedups are obtained by parallelizing **Alfa**-type algorithms.

We additionally want to pose the question whether some of the Conjectures 6 or 8 can be theoretically proven. A first line of attack would be to analyze the special case of an solver like SRWA whose runtime was already theoretically analyzed.

The technique of analyzing the runtime distribution of **Alfa** could be further developed to help better understand the behavior of CDCL solvers. These kind of solvers heavily employ the technique of adding new clauses and deleting some clauses. This can be thought of as solving a new logically equivalent formula of the base instance.

Preliminary results on the solvers excluded for heuristic reasons seem to suggest that the **Alfa**-method forces the runtime of the base solver to exhibit a multimodal behavior. Thus, the lognormal distribution is not a good fit in this case. However, an initial visual inspection of the data indicates an even heavier tail.

References

- 1 Alejandro Arbelaez, Charlotte Truchet, and Philippe Codognet. Using sequential runtime distributions for the parallel speedup prediction of SAT local search. *Theory and Practice of Logic Programming*, 13(4-5):625–639, 2013.
- 2 Gilles Audemard and Laurent Simon. Predicting learnt clauses quality in modern SAT solvers. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI '09)*, pages 399–404, 2009.
- 3 Adrian Balint. Original implementation of *probSAT*, 2015. Available at <https://github.com/adrianopolus/probSAT>.
- 4 Adrian Balint, Anton Belov, Marijn J. H. Heule, and Matti Järvisalo. SAT Competition 2013: Results. URL: <http://satcompetition.org/2013/results.shtml>.
- 5 Tomáš Balyo and Lukáš Chrpa. Using algorithm configuration tools to generate hard SAT benchmarks. In *Proceedings of the 11th International Symposium on Combinatorial Search (SOCS '18)*, pages 133–137. AAAI Press, 2018.
- 6 Wolfgang Barthel, Alexander K. Hartmann, Michele Leone, Federico Ricci-Tersenghi, Martin Weigt, and Riccardo Zecchina. Hiding solutions in random satisfiability problems: A statistical mechanics approach. *Physical review letters*, 88(188701):1–4, 2002.
- 7 Armin Biere. Yet another local search solver and Lingeling and friends entering the SAT Competition 2014. In *Proceedings of SAT Competition 2014: Solver and Benchmark Descriptions*, volume B-2014-2, pages 39–40. University of Helsinki, 2014.
- 8 Archie Blake. *Canonical Expressions in Boolean Algebra*. PhD thesis, University of Chicago, 1937.
- 9 Russell Cheng. *Non-standard parametric statistical inference*. Oxford University Press, 2017.
- 10 Maximilian Diemer. Source code of *GenFactorSat*, 2021. Newest version available at <https://github.com/madiemer/gen-factor-sat/>.
- 11 Niklas Eén and Niklas Sörensson. An extensible SAT-solver. In *Selected Revised Papers of the 6th International Conference on Theory and Applications of Satisfiability Testing (SAT '03)*, volume 2919 of *Lecture Notes in Computer Science*, pages 502–518. Springer, 2004.
- 12 Sergey Foss, Dmitry Korshunov, and Stan Zachary. *An Introduction to Heavy-Tailed and Subexponential Distributions*, volume 6. Springer, 2011.
- 13 Daniel Frost, Irina Rish, and Lluís Vila. Summarizing CSP hardness with continuous probability distributions. In *Proceedings of the 14th National Conference on Artificial Intelligence and 9th Innovative Applications of Artificial Intelligence Conference (AAAI/IAAI '97)*, pages 327–333, 1997.
- 14 Oliver Gableske. Source code of *kcnfgen (version 1.0)*, 2015. Retrieved from https://www.gableske.net/downloads/kcnfgen_v1.0.tar.gz.
- 15 Carla P. Gomes and Bart Selman. Algorithm portfolio design: Theory vs. practice. In *Proceedings of the 13th Conference on Uncertainty in Artificial Intelligence (UAI '97)*, pages 190–197, 1997.

- 16 Carla P. Gomes, Bart Selman, Nuno Crato, and Henry A. Kautz. Heavy-tailed phenomena in satisfiability and constraint satisfaction problems. *Journal of Automated Reasoning*, 24:67–100, 2000. Related version in *CP '97*.
- 17 Marijn J. H. Heule, Matti Järvisalo, and Tomáš Balyo. SAT Competition 2017: Results. URL: <https://baldur.iti.kit.edu/sat-competition-2017/index.php?cat=results>.
- 18 Holger H. Hoos and Thomas Stützle. Evaluating Las Vegas algorithms: Pitfalls and remedies. In *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence (UAI '98)*, pages 238–245, 1998.
- 19 Norman L. Johnson, Samuel Kotz, and Narayanaswamy Balakrishnan. *Continuous Univariate Distributions, Volume 1*. Wiley Series in Probability and Statistics. John Wiley & Sons, 2nd edition, 1994.
- 20 Alexis C. Kaporis, Lefteris M. Kirousis, and Yannis C. Stamatiou. A note on the non-colorability threshold of a random graph. *The Electronic Journal of Combinatorics*, 7(1), 2000.
- 21 Massimo Lauria, Jan Elffers, Jakob Nordström, and Marc Vinyals. CNFgen: A generator of crafted benchmarks. In *Proceedings of the 20th International Conference on Theory and Applications of Satisfiability Testing (SAT '17)*, pages 464–473, 2017.
- 22 Jan-Hendrik Lorenz. Runtime distributions and criteria for restarts. In *Proceedings of the 44th International Conference on Current Trends in Theory and Practice of Computer Science (SOFSEM '18)*, pages 493–507. Springer, 2018.
- 23 Jan-Hendrik Lorenz and Florian Wörz. On the effect of learned clauses on stochastic local search. In *Proceedings of the 23rd International Conference on Theory and Applications of Satisfiability Testing (SAT '20)*, volume 12178 of *Lecture Notes in Computer Science*, pages 89–106. Springer, 2020. Implementation and statistical tests of GapSAT available at Zenodo doi:10.5281/zenodo.3776052.
- 24 Jan-Hendrik Lorenz and Florian Wörz. Source code of *concealSATgen*, 2021. Newest version available at <https://github.com/FlorianWoerz/concealSATgen/>.
- 25 Chuan Luo, Shaowei Cai, and Kaile Su. DCCAlm in SAT Competition 2016. In *Proceedings of SAT Competition 2016: Solver and Benchmark Descriptions*, volume B-2016-1. University of Helsinki, 2016.
- 26 Chuan Luo, Shaowei Cai, Wei Wu, and Kaile Su. Focused random walk with configuration checking and break minimum for satisfiability. In *Proceedings of the 19th International Conference on Principles and Practice of Constraint Programming (CP '13)*, volume 8124 of *Lecture Notes in Computer Science*, pages 481–496. Springer, 2013.
- 27 Chuan Luo, Shaowei Cai, Wei Wu, and Kaile Su. Double configuration checking in stochastic local search for satisfiability. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence*, pages 2703–2709. AAAI Press, 2014.
- 28 Chuan Luo, Shaowei Cai, Wei Wu, and Kaile Su. CSCCSat in SAT Competition 2016. In *Proceedings of SAT Competition 2016: Solver and Benchmark Descriptions*, volume B-2016-1. University of Helsinki, 2016.
- 29 Stephan Mertens, Marc Mézard, and Riccardo Zecchina. Threshold values of random k -SAT from the cavity method. *Random Structures & Algorithms*, 28(3):340–373, 2006.
- 30 Jayakrishnan Nair, Adam Wierman, and Bert Zwart. The fundamentals of heavy tails: Properties, emergence, and estimation. Preprint, California Institute of Technology, 2020.
- 31 Marvin Rausand, Anne Barros, and Arnljot Hoyland. *System reliability theory: models, statistical methods, and applications*. John Wiley & Sons, 2nd edition, 2003.
- 32 Irina Rish and Daniel Frost. Statistical analysis of backtracking on inconsistent CSPs. In *Proceedings of the 3rd International Conference on Principles and Practice of Constraint Programming (CP '97)*, pages 150–162, 1997.
- 33 John Alan Robinson. A machine-oriented logic based on the resolution principle. *Journal of the ACM*, 12(1):23–41, 1965.

- 34 Yongshao Ruan, Eric Horvitz, and Henry A. Kautz. Restart policies with dependence among runs: A dynamic programming approach. In *Proceedings of the 8th International Conference on Principles and Practice of Constraint Programming (CP '02)*, volume 2470 of *Lecture Notes in Computer Science*, pages 573–586. Springer, 2002.
- 35 Walter Rudin. *Principles of mathematical analysis*, volume 3. McGraw-Hill New York, 1964.
- 36 Uwe Schöning. A probabilistic algorithm for k -SAT based on limited local search and restart. *Algorithmica*, 32(4):615–623, 2002. Preliminary version in *FOCS '99*.
- 37 Uwe Schöning and Jacobo Torán. *The Satisfiability Problem: Algorithms and Analyses*, volume 3 of *Mathematics for Applications (Mathematik für Anwendungen)*. Lehmanns Media, 2013.
- 38 Oleg V. Shylo, Timothy Middelkoop, and Panos M. Pardalos. Restart strategies in optimization: parallel and serial cases. *Parallel Computing*, 37(1):60–68, 2011.
- 39 Gunnar Völkel, Ludwig Lausser, Florian Schmid, Johann M. Kraus, and Hans A. Kestler. Sputnik: *ad hoc* distributed computation. *Bioinformatics*, 31(8):1298–1301, 2015.
- 40 Sven Dag Wicksell. On logarithmic correlation with an application to the distribution of ages at first marriage. *Meddelanden från Lunds Astronomiska Observatorium*, 84:1–21, 1917.
- 41 Florian Wörz and Jan-Hendrik Lorenz. Evidence for long-tails in SLS algorithms. Technical Report 2107.00378, arXiv.org, 2021. This is the full-length version of the present paper.
- 42 Florian Wörz and Jan-Hendrik Lorenz. Supplementary Data for “Evidence for Long-Tails in SLS Algorithms”, 2021. We have provided all data of this paper. All base instances, resolvents, and modifications can be found under [doi:10.5281/zenodo.4715893](https://doi.org/10.5281/zenodo.4715893). Visual and statistical evaluations can be found under https://github.com/FlorianWoerz/SLS_Evidence_long_tail, where all evaluations take place in the files `./evaluation/jupyter/evaluate_*.ipynb`. A permanent version of this repository has been preserved under [doi:10.5281/zenodo.5026180](https://doi.org/10.5281/zenodo.5026180).