MDPI

Article

Scale-Free Random SAT Instances

Carlos Ansótegui [1,†] , Maria Luisa Bonet [2,†] and Jordi Levy [3,*,†]

1 Departament d'Informàtica i Enginyeria Industrial, Universitat de Lleida, 25001 Lleida, Spain; carlos@diei.udl.cat
2 Computer Science Department, Universitat Politècnica de Catalunya, 08034 Barcelona, Spain; bonet@cs.upc.edu
3 Instituto de Investigación en Inteligencia Artificial (IIIA), CSIC, 08193 Bellaterra, Spain
* Correspondence: levy@iiia.csic.es
† These authors contributed equally to this work.

Abstract: We focus on the random generation of SAT instances that have properties similar to real-world instances. It is known that many industrial instances, even with a great number of variables, can be solved by a clever solver in a reasonable amount of time. This is not possible, in general, with classical randomly generated instances. We provide a different generation model of SAT instances, called *scale-free random SAT instances.* This is based on the use of a non-uniform probability distribution $P(i) \sim i^{-\beta}$ to select variable i, where β is a parameter of the model. This results in formulas where the number of occurrences k of variables follows a power-law distribution $P(k) \sim k^{-\delta}$, where $\delta = 1 + 1/\beta$. This property has been observed in most real-world SAT instances. For $\beta = 0$, our model extends classical random SAT instances. We prove the existence of a SAT–UNSAT phase transition phenomenon for scale-free random 2-SAT instances with $\beta < 1/2$ when the clause/variable ratio is $m/n = \frac{1-2\beta}{(1-\beta)^2}$. We also prove that scale-free random k-SAT instances are unsatisfiable with a high probability when the number of clauses exceeds $\omega(n^{(1-\beta)k})$. The proof of this result suggests that, when $\beta > 1 - 1/k$, the unsatisfiability of most formulas may be due to small cores of clauses. Finally, we show how this model will allow us to generate random instances similar to industrial instances, of interest for testing purposes.

Keywords: constraint satisfaction; SAT; scale-free networks; random models

1. Introduction

Over the last 20 years, SAT solvers have experienced a great improvement in their efficiency when solving practical SAT problems. This is the result of some techniques such as conflict-driven clause learning (CDCL), restarting, and clause deletion policies. The success of SAT solvers is surprising if we take into account that SAT is an NP-complete problem, and in fact, a big percentage of formulas need exponential size resolution proofs to be shown to be unsatisfiable. This has led some researchers to study what is the nature of real-world or industrial SAT instances that make them easy in practice. In parallel, most theoretical work on SAT has focused on uniform randomly selected instances. Nevertheless, nowadays we know that most industrial instances share some properties that are not present in most (uniform randomly chosen) SAT formulas. It is also well-known that solvers that perform well in industrial instances, do not perform well in random instances, and vice versa. Therefore, a new theoretical paradigm that describes the distribution of industrial instances is needed. Not surprisingly, generating random instances that are more similar to real-world instances is described as one of the ten biggest challenges in satisfiability [1–4].

Over the last 10 years, the analysis of the industrial SAT instances used in SAT solver competitions has allowed us to obtain a clear image of the structure of real-world instances. Ansótegui et al. [5] proved that most industrial instances, when represented as a graph,

have a scale-free structure. This kind of structure has also been observed in other real-world networks like the World Wide Web, Internet, some social networks like papers co-authorship or citation, protein interaction network, etc. Ansótegui et al. [6] show that these graph representations of industrial instances exhibit very high modularity. Modularity has been shown to be correlated with the runtime of CDCL SAT solvers [7] and has been used to improve the performance of some solvers [6,8,9]. It is also known that eigenvector centrality is correlated with the significance of variables [10].

Defining a model that captures all the properties observed in industrial instances is a hard task. Here, we focus on the scale-free structure. We will define a model and propose a random generator for *scale-free SAT formulas*, extending our work presented in the IJCAI'09 [11] conference. This model is parametric in terms of the size k of clauses and an exponent β. Formulas are sets of m independently sampled clauses of size k with possible repetitions. Clauses are sets of k independently sampled variables, without repetitions, where each variable x_i is chosen with probability $P(x_i) \sim i^{-\beta}$, and negated with probability $1/2$.

In this paper, we also study the SAT–UNSAT phase transition phenomena in this new model using percolation techniques of statistical mechanics. We prove that a random scale-free formula over n variables, exponent β and $\mathcal{O}(n^{(1-\beta)k})$ clauses of size k are unsatisfiable with a high probability (see Theorem 5). This means that, for big enough values of β, the number of clauses needed to make a formula unsatisfiable is sub-linear on the number of variables, contrary to the standard random SAT model. We also prove that scale-free random 2-SAT formulas with exponent $\beta < 1/2$ and a ratio of clause/variables $m/n > \frac{1-2\beta}{(1-\beta)^2}$ are also unsatisfiable with high probability (see Theorem 4). This last result, together with a coincident lower bound found by Friedrich et al. [12], allows us to conclude that scale-free random 2SAT formulas show a SAT–UNSAT phase transition threshold.

More recently, many new results related to the phase transition on scale-free random formulas have been found. Friedrich et al. [13] generalize the notion of scale-free random k-SAT formulas and prove that there exists an asymptotic satisfiability threshold (in the sense of [14]) for $\beta < 1 - 1/k$, when the number of clauses is linear in the number of variables. Friedrich and Rothenberger [15] find sufficient conditions for the sharpness of this threshold, generalizing the results of Friedgut [14] for uniform random formulas. Ref. [16] generalizes the notion of scale-free formulas to the notion of *non-uniform random formulas*, only assuming that variable x_i is selected with probability p_i (where $p_1 \geq p_2 \geq \cdots \geq p_n$) and determine the position of the threshold for $k = 2$. Cooper et al. [17] and Omelchenko and Bulatov [18,19,20] analyzed the *configuration model* for 2-SAT where, instead of fixing the probability of every variable, they fixed the degree of every variable. If these degrees follow a power-law distribution, the location of the satisfiability threshold (for $k = 2$) is the same as in our model. Giráldez-Cru and Levy [21,22] propose a new model of random formulas that, besides heterogeneity, also consider the notion of locality in formulas. Achlioptas et al. [23] computes the number of assignments of uniform random 2-SAT formulas, using the cavity method. Finally, Bläsius et al. [24] analyze the hardness of random instances according to their heterogeneity and locality.

This article proceeds as follows: In Section 2 we review some methods to generate scale-free random graphs. One of these methods is the basis of the definition of scale-free random formulas, introduced in Section 3. In Section 4, we summarize some properties of industrial or real-world SAT instances, described in detail in our work presented in [5]. We prove the existence of a SAT–UNSAT phase transition phenomenon in scale-free random 2-SAT instances in Section 5. This is achieved using percolation techniques. In Section 6, we prove that when the β parameter that regulates the scale-free structure of formulas exceeds a particular value, the SAT–UNSAT phase transition phenomena vanishes, and most formulas become unsatisfiable due to the small cores of unsatisfiable clauses.

2. Generation of Scale-Free Graphs

Generating scale-free formulas has an obvious relationship with the generation of scale-free graphs. In this section, we review some graph generation methods developed by researchers on complex networks.

A scale-free graph is a graph where node degrees follow a power-law distribution $P(k) \sim k^{-\gamma}$, at least asymptotically, where exponent γ is around 3. A preferential attachment [25] has been proposed as the natural process that makes scale-free networks prevalent in nature. This process can be used to generate scale-free graphs as follows: Given two numbers n and m, we start at time $t = m + 1$ with a clique of size $m + 1$ where all nodes have degree m (in the limit when n tends to infinity, the starting graph is not relevant). Then, at every time $t = m + 2, \ldots, n$, we add a new node (with index t), connected to m distinct and older nodes $s < t$, such that the probability that a node s gets a connection to this new node t is proportional to the degree of s at time t. This process generates a scale-free graph with asymptotic exponent $\gamma = 3$, average node degree $\mathrm{E}[k] = 2\,m$ and minimum degree $k_i \geq m$, for all nodes. We can also prove that the expected degree of node i is $\mathrm{E}[k_i] \sim i^{-\beta}$, for small values of i, where $\beta = 0.5$.

In order to explain the origin of scale-free networks where $\gamma \neq 3$, several models have been proposed [26]. One of these models is based on the *aging* of nodes [27]. This means that the probability of a node s (created at instant s) obtaining a new edge at instant t is proportional to the product of its degree and $(t - s)^{-\alpha}$, where $t - s$ is the *age* of the node. This model generates scale-free graphs when $\alpha < 1$. When $\alpha \to 0$, the exponents of the power-laws $P(k) \sim k^{-\gamma}$ and $\mathrm{E}[k_i] \sim i^{-\beta}$ are $\gamma = 3 + 4(1 - \ln 2)\alpha$ and $\beta = 1/2 - (1 - \ln 2)\alpha$, respectively. Therefore, the value of α may be used to tune the values of γ and β.

In the previous methods, *growth* in the number of nodes is essential. There are other methods, usually called *static*, where the number of nodes is fixed from the beginning and during the process we only add edges.

The simplest method, assuming uniform probability for all graphs with a scale-free degree distribution in the degree of nodes, is the *configuration method*, which can be implemented as outlined in the next paragraph.

Given a desired number of nodes n and exponent γ, for every node $i \in \{1, \ldots, n\}$, generate a degree k_i following the probability $P(k) = k^{-\gamma}/\zeta(\gamma)$, independently of i. Here, $\zeta(x) = \sum_{i=1}^{\infty} i^{-x}$ is the Riemann zeta function. Then, generate a graph with these node degrees, ensuring that all of them are generated with the same probability. This can be achieved, for instance, with an unfold–fold process: In the unfolding, we replicate node i, with degree k_i, into k_i new nodes with degree 1. Then, we randomly generate a graph where all nodes have degrees equal to one, ensuring that all 1-regular graphs with $\sum_{i=1}^{n} k_i$ nodes are generated with the same probability. Then, in the folding, we merge the k_i nodes that came from the replication of i, into the same node. When there is an edge between two nodes and these two nodes are merged, a self-loop is created. Similarly, when we have two edges $i_1 \leftrightarrow j$ and $i_2 \leftrightarrow j$ and i_1 and i_2 are merged, a duplicated edge is created. Therefore, we reject the resulting graph, if it contains self-loops or multiple edges between the same pair of nodes. Alternatively, we can also apply the Erdös-Rényi generation method to the unfolded set of nodes, with the average node degree equal to one. In the last case, we would ensure that after folding, node i has a degree close to k_i, since in the Erdös-Rényi model, node degrees follow a binomial distribution (a Poisson distribution $P(k) = e^{-z} \frac{z^k}{k!}$ in the infinite limit, where z is the average degree, $z = 1$ in our case).

The previous method has two problems. First, the resulting graph (after the unfold–fold process) will have an average node degree equal to:

$$\mathrm{E}[k] = \sum_{k=1}^{\infty} k\, P(k) = \sum_{k=1}^{\infty} k\, \frac{k^{-\gamma}}{\zeta(\gamma)} = \frac{\zeta(\gamma - 1)}{\zeta(\gamma)} \tag{1}$$

If we want to obtain a graph with a distinct average degree, we have to modify the probability $P(k)$ for small values of k and ensure that $P(k)$ follows a power-law distribution

only asymptotically for big values of k. In other words, we only require $P(k)$ to follow a *heavy-tail* distribution. Second, a great fraction of generated graphs will contain self-loops or multiple edges between the same pair of nodes after folding. This means that a great fraction of graphs will be rejected, which makes the method inefficient. However, the model can be useful to translate some properties of the Erdös-Rényi model to scale-free graphs via the unfolding–folding process and the *configuration model* [28,29].

The unfolding–folding procedure was described by Aiello et al. [30]. They, instead of assigning a random degree to each node, describe a model where, given two parameters α and γ (In the original paper, authors use the name β instead of γ.), we choose a random graph (with uniform probability, and allowing self-cycles) among all graphs satisfying that the number of nodes with degree x is e^{α}/x^{γ}. When $\gamma > 2$, the average node degree in this model is also $\frac{\zeta(\gamma-1)}{\zeta(\gamma)}$.

Alternatively, instead of fixing the degree of every node, we can fix the *expected degree* of every node $\mathrm{E}[k_i] = w_i$. In order to construct a graph where nodes have this expected degree $\mathrm{E}[k_i] \sim w_i$, we only need to generate edge $i \leftrightarrow j$ with probability $P(i \leftrightarrow j) \sim w_i w_j$. If we want to generate a scale-free graph where $P(k) \sim k^{-\delta}$, for sparse graphs, it suffices to fix $w_i = i^{1/(\delta-1)}$ [31,32] (see also Theorem 1).

Our scale-free formula generation method is based on this static scale-free graph generation model with fixed expected node degrees. Basically, nodes are replaced by variables. Then, instead of edges, we generate hyper-edges. By negating every variable connected by a hyper-edge with probability $1/2$, we obtain clauses.

3. Scale-Free Random Formulas

In this section, we describe the *scale-free random SAT formulas* model.

We consider k-SAT formulas over n variables, denoted by $x_1, \ldots, x_n$. A formula is a conjunction of m possibly repeated clauses, represented as a multiset. Clauses are disjunctions of k literals, noted as $l_1 \vee \cdots \vee l_k$, where every literal may be a variable x_i or its negation $\neg x_i$. We identify $\neg\neg x$ with x. We restrict clauses to not contain repeated occurrences of variables. This avoids simplifiable formulas such as $x \vee x \vee y$ and tautologies like $x \vee \neg x \vee y$. In general, we represent every variable by its index, and negation as a minus, writing i instead of x_i, and $-i$ instead of $\neg x_i$. In other words, a variable x is a number in $\{1, \ldots, n\}$, and a literal a number in $\{-n, \ldots, n\}$ distinct from zero. We use the notation $\pm x$ to denote either x or $\neg x$. The number of occurrences of literal l in a formula is denoted by k_l, and $K_x = k_x + k_{\neg x}$ denotes the number of occurrences of variable x. The size of a formula F is $|F| = m k$.

In the following, we will use the notation $P(x) \sim f(x)$ to indicate that random variable x follows the probability distribution $f(x)$. The notation $f(n) \approx g(n)$ indicates that $\lim_{n\to\infty} \frac{f(x)}{g(x)} = 1$ (Notice that $f(n) \approx g(n)$ is equivalent to $f(n) = g(n)(1 + o(1))$. When $g(n) = \Theta(1)$, then $f(n) = g(n)(1 + o(1))$ is equivalent to $f(n) = g(n) + o(1)$.).

Definition 1. *In the scale-free model, given n, m and β, to construct a random formula, we generate m clauses independently at random from the set of $2^k \binom{n}{k}$ clauses, sampling every valid clause with probability*

$$P(l_1 \vee \cdots \vee l_k) \sim \prod_{i=1}^{k} P(l_i)$$

where every literal l_i is sampled with probability

$$P(x) = P(\neg x) \sim x^{-\beta}$$

In practice, we generate a variable x with probability $P(x) = x^{-\beta}/\sum_{i=1}^{n} i^{-\beta}$, negate it with probability $1/2$, repeat the process k times, and reject clauses containing repeated variables.

Therefore, the probability of a clause satisfies the inequality

$$P(l_1 \vee \cdots \vee l_k) \geq \frac{k! \prod_{i=1}^{k} |l_i|^{-\beta}}{(2 \sum_{i=1}^{n} i^{-\beta})^k} \tag{2}$$

3.1. Some Properties of the Model

In the case of the graph generator, we reject self-loops and repeated edges between two nodes. This makes the distribution of degrees follow a power-law, only asymptotically and for sparse graphs. In our case, we reject clauses with repeated variables. This is the reason that the reverse direction is invalid in the previous inequality. It also makes formulas follow a power-law distribution in the number of variable occurrences only asymptotically (see Theorem 1). In the following, we will discuss when the approximation for $P(l_1 \vee \cdots \vee l_k) \approx \frac{k! \prod_{i=1}^{k} |l_i|^{-\beta}}{(2 \sum_{i=1}^{n} i^{-\beta})^k}$ is tight (Lemma 1).

Notice that $\sum_{i=1}^{n} i^{-\beta} = H_{n,\beta}$ are the *generalized harmonic numbers*. When n tends to infinity and $\beta \neq 1$, using the Euler–Maclaurin formula, they can be approximated as

$$\sum_{i=1}^{n} i^{-\beta} = \zeta(\beta) + \frac{1}{1-\beta} n^{1-\beta} + \frac{1}{2} n^{-\beta} + \mathcal{O}(n^{-\beta-1}) \tag{3}$$

where $\zeta(\beta)$ is the Riemann zeta function. When $\beta = 1$, we have

$$\sum_{i=1}^{n} i^{-1} = \gamma + \log n + \mathcal{O}(n^{-1}) \tag{4}$$

where γ is the Euler constant.

This means that, when n tends to infinity, the probability of sampling variable x_i is $P(x_i) = o(1)$, when $0 \leq \beta \leq 1$, and $P(x_i) = i^{-\beta}/\zeta(\beta) + o(1)$, when $\beta > 1$. Due to the fact that the probability of sampling a variable does not vanish, when the number of variables tend to infinity and $\beta > 1$, may be troublesome. In particular, the probability of generating clauses with duplicated variables does not vanish, even for constant clause sizes. Similarly, to avoid duplicated variables, we also have to impose an upper bound $k = o(n^{\min\{1/2, 1-\beta\}})$.

Lemma 1. *When $0 \leq \beta < 1$, the sizes of clauses are $k = o(n^{\min\{1/2, 1-\beta\}})$ and n tends to infinity, the probability of generating a clause with a duplicated variable tends to zero.*

In these conditions, the probability of a random variable and the probability of a random clause in a formula are

$$P(x = x_i) \approx \frac{i^{-\beta}}{\sum_{j=1}^{n} j^{-\beta}}$$

$$P(C = l_1 \vee \cdots \vee l_k) \approx \frac{k! \prod_{i=1}^{k} |l_i|^{-\beta}}{(2 \sum_{i=1}^{n} i^{-\beta})^k}$$

Proof. We will use a result known as *surname problem* [33], which generalizes the *birthday paradox*. Let $X_1, \ldots, X_k$ be independent random variables which have an identical discrete distribution $P(X = i) = p_i$, for $i \geq 1$. Let R_k be the coincidence probability that at least two X_j have the same value. Let $r_k = 1 - R_k$ be the non-coincidence probability. Then, r_k may be computed using the recurrence $r_0 = 1$ and

$$r_k = \sum_{j=1}^{k} (-1)^{j-1} \frac{(k-1)!}{(k-j)!} P_j r_{k-j}$$

where $P_k = \sum_{i \geq 1}(p_i)^k$. The coincidence probability can be computed as $R_1 = 0$ and

$$R_k = R_{k-1} + \sum_{j=2}^{k} (-1)^j \frac{(k-1)!}{(k-j)!} P_j (1 - R_{k-j})$$

In our case, we face the problem of choosing k independent variables, and we want to compute the probability of obtaining a duplicated variable, i.e., a rejected clause. When $\beta < 1$, we have

$$P_k = \frac{\sum_{i \geq 1} i^{-\beta k}}{(\sum_{i \geq 1} i^{-\beta})^k} = \frac{\frac{1}{1-\beta k} n^{1-\beta k} + \zeta(\beta k) + \mathcal{O}(n^{-\beta k})}{(\frac{1}{1-\beta} n^{1-\beta} + \mathcal{O}(1))^k} = \frac{(1-\beta)^k}{1-\beta k} n^{1-k} + \zeta(\beta k)(1-\beta)^k n^{-(1-\beta)k} + \mathcal{O}(n^{-k})$$

Depending on whether βk is greater or smaller than 1, the first or the second term of P_k will dominate.

Since $\frac{(k-1)!}{(k-j)!} < k^{j-1}$ and $R_{k-j} \geq 0$, we have

$$R_k \leq \sum_{i=2}^{k} \sum_{j=2}^{i} i^{j-1} P_j \leq k \max_{j=2,\dots,k} k^{j-1} P_j$$

In our case, assuming $k = \mathcal{O}(n^\alpha)$, and replacing the value of P_j, we obtain

$$R_k \leq \mathcal{O}(n^\alpha) \max_{j=2,\dots k} \mathcal{O}(n^{\alpha(j-1)}(n^{1-j} + n^{-(1-\beta)j})) = \max_{j=2,\dots k} \mathcal{O}(n^{1-(1-\alpha)j} + n^{-(1-\beta-\alpha)j})$$

Assuming $\alpha \leq 1 - \beta < 1$, the maximum is obtained for $j = 2$. In this situation, $R_k \leq \mathcal{O}(n^{-(1-2\alpha)} + n^{-1(1-\beta-\alpha)})$. Therefore, it suffices to assume that $\alpha < \min\{1/2, 1-\beta\}$ to ensure that $R_k = o(1)$. $\quad\square$

Lemma 2. *In a scale-free random formula over n variables and $m = C n$ clauses of size $k = \mathcal{O}(1)$, generated with exponent $0 < \beta < 1$, the expected number of occurrences of variable x_i is*

$$\mathrm{E}[K_i] \approx C k (1-\beta) \left(\frac{i}{n}\right)^{-\beta}$$

Proof. By Lemma 1 and Equation (3), since $0 < \beta < 1$ we have

$$\mathrm{E}[K_i] = P(i) |F| \approx \frac{i^{-\beta}}{\zeta(\beta) + \frac{1}{1-\beta} n^{1-\beta} + \mathcal{O}(n^{-\beta})} C k n \approx C k (1-\beta) \left(\frac{i}{n}\right)^{-\beta} \quad\square$$

The following theorem ensures that the formulas we obtain are scale-free, in the sense that the number of occurrences of variables follow a power-law distribution $P(K) \sim K^{-\delta}$, for big enough values of K.

Theorem 1. *In scale-free random formulas over n variables, with $m = C n$ clauses of size k, and generated with exponent $0 < \beta < 1$, when n tends to ∞, with C and k being constants, the probability that a variable has K occurrences, where $K = \Omega(\sqrt{n \log n})$ or $K = \Omega\left((n^2 \log n)^{\frac{\beta}{2+\beta}}\right)$, follows a power-law distribution $P(K) \sim K^{-\delta}$, where $\delta = 1/\beta + 1$.*

Proof. In the limit when $n \to \infty$, by Lemma 1, $P(x_i) \approx \mathfrak{C} i^{-\beta}$ is the probability of sampling a variable x_i, for some constant $\mathfrak{C} = 1/\sum_{j=1}^{n} j^{-\beta} \approx (1-\beta) n^{\beta-1}$ that depends on n. Let K_i be the number of occurrences of variable i in a randomly generated formula F. We have $\mathrm{E}[K_i] = |F| \mathfrak{C} i^{-\beta}$. Chernoff's or Hoeffding's bounds ensure that, under certain conditions that we will consider later, K_i is approximately $\mathrm{E}[K_i]$. Hence, $K_1 > K_2 > \cdots > K_n$ with high probability.

Now, we want to approximate the probability $F(K) = \int_{k=K}^{\infty} P(k) \, dk$ that a variable occurs at least K times. Given a value K, let i be the index of the variable satisfying $\mathrm{E}[K_i] = K$. Under these conditions, all variables with index smaller that i will have more than K occurrences, and those with indexes between $i+1$ and n have fewer than K occurrences. Therefore, $F(K) = i/n$, for the particular i defined above. From $\mathrm{E}[K_i] = K$ and $\mathrm{E}[K_i] = |F| \, C \, i^{-\beta}$, we obtain

$$F(K) = \frac{i}{n} = \frac{1}{n} \left(\frac{K}{|F| \, \mathfrak{C}} \right)^{-1/\beta}$$

Then, the probability $P(K)$ is

$$P(K) = -\frac{\partial}{\partial K} F(K) = \frac{(|F| \, \mathfrak{C})^{1/\beta}}{\beta \, n} K^{-1/\beta - 1}$$

Hence, we obtain a discrete power-law distribution with exponent $\delta = 1/\beta + 1$.

The problem is that $\mathrm{E}[K_i]$ is a good approximation of K_i only when i is small. For instance, when $i = \Omega(n)$, we have $P(x_i) = \Theta(n^{-1})$ and $\mathrm{E}[x_i] = \Theta(1)$. In this situation, when $n \to \infty$ being C and k constants, the number of occurrences K_i of the variable x_i follows a Poisson distribution with constant variance. This means that, even in the limit $n \to \infty$, we cannot assume that $i < j$ implies $K_i > K_j$, when $i = \Omega(n)$. In the following, we will find an upper bound for the index i of the variable (a lower bound for the value of K) ensuring that $\mathrm{E}[K_i]$ is a good approximation of K_i, when $n \to \infty$. We will use both Hoeffding's and Chernoff's bounds.

In what follows, let be C be the constant such that $|F| \approx C \, n$ is the size of the formula.

Hoeffding's bound states that, if $X = X_1 + \cdots + X_n$ is the sum of identical and independent Bernoulli variables, then

$$P(|X - \mathrm{E}[X]| \geq \epsilon \, n) \leq 2 \, e^{-2\epsilon^2 n}$$

taking $\epsilon = \sqrt{\frac{\log n}{n}}$ we obtain

$$P\left(|X - \mathrm{E}[X]| \geq \sqrt{n \log n} \right) \leq \frac{2}{n^2}$$

Given a value of K, let us fix two variables i and j such that

$$\mathrm{E}[K_i] = K$$
$$\mathrm{E}[K_j] = K - \sqrt{n \log n}$$

We have $P(K_j \geq K) \leq 2/n^2$, and for all variables r with bigger indexes

$$\sum_{r \geq j} P(K_r \geq K) = o(1)$$

We have already argued that $F(K) = P(k \geq K) \approx i/n$. Using j, we have a strict bound

$$P(k \geq K) \leq j/n + o(1)$$

By Lemma 2, we obtain

$$
\begin{aligned}
K = \quad & \mathrm{E}[K_i] \approx C(1-\beta)(i/n)^{-\beta} \\
K - \sqrt{n \log n} = \quad & \mathrm{E}[K_j] \approx C(1-\beta)(j/n)^{-\beta} = C(1-\beta)(j/i)^{-\beta}(i/n)^{-\beta} \approx K(j/i)^{-\beta}
\end{aligned}
$$

Therefore,

$$j \approx i\left(1 - \frac{\sqrt{n\log n}}{K}\right)^{1/\beta}$$

$$i \approx \left(\frac{C(1-\beta)}{K}\right)^{1/\beta}$$

By replacing the expressions for i and j, we obtain

$$P(k \geq K) \leq j/n + o(1) \approx \left(\frac{C(1-\beta)}{K}\left(1 - \frac{\sqrt{n\log n}}{K}\right)\right)^{1/\beta}$$

If $K = \Omega(\sqrt{n\log n})$, then

$$P(k \geq K) \leq \left(\frac{K}{C(1-\beta)}\right)^{-1/\beta} + o(1)$$

Similarly, we can prove the same lower bound $P(k \geq K) \geq \left(\frac{K}{C(1-\beta)}\right)^{-1/\beta} + o(1)$ now using the variable j such that $\mathrm{E}[K_k] = K + \sqrt{n\log n}$.

Alternatively, we can use the Chernoff's bound

$$P(|X - \mathrm{E}[X]| \leq \delta\mathrm{E}[X]) \leq 2\,e^{-\frac{\delta^2\mathrm{E}[X]}{3}}$$

where X is the sum of independent random variables in the range $[0,1]$. In order to ensure that the K_i's are sorted, in the limit $n \to \infty$, we have $P(K_i < K_{i+1}) = \mathcal{O}(n^{-1})$. We take the value of δ that satisfies

$$\delta\mathrm{E}[K_i] = \frac{\mathrm{E}[K_i] - \mathrm{E}[K_{i+1}]}{2}$$

By Lemma 2 and the Taylor expansion $(1+x)^a = 1 + a\,x + \mathcal{O}(x^2)$, this value of δ, when $i \to \infty$, is

$$\delta \approx 1/2 - 1/2\left(\frac{i+1}{i}\right)^{-\beta} \approx \frac{\beta}{2i}$$

and, for this value of δ, we impose

$$2\,e^{-\frac{\delta^2\mathrm{E}[K_i]}{3}} \approx 2\exp\left(-\frac{(\beta/2i)^2 C(1-\beta)(i/n)^{-\beta}}{3}\right) = \mathcal{O}(n^{-1})$$

From this, we obtain the minimum value of i, for which $P(K_i < K_{i+1}) = \mathcal{O}(n^{-1})$.

$$i = \mathcal{O}\left(n^{\beta/(2+\beta)} / \log^{1/(2+\beta)} n\right)$$

The value of $K = \mathrm{E}[K_i]$ corresponding to this variable x_i gives us a value from which on we can expect to observe the power-law distribution in $P(K)$.

$$K = \Omega\left((n^2\log n)^{\beta/(2+\beta)}\right) \quad \square$$

3.2. Implementation of the Generator

The generation method is formalized in Algorithm 1.

Algorithm 1: Scale-free random k-SAT formula generator.

Input: n, m, k, β

Output: a k-SAT instance with n variables and m clauses

1 $F = \varnothing$;

2 **foreach** $i = 1, \dots, m$ **do**

3   **repeat**

4          $C_i = \square$;

5   **foreach** $j = 1, \dots, k$ **do**

6   v = sampleVariable(β,n);

7              $C_i = C_i \vee (-1)^{rand(2)} \cdot v$;

8   **until** C_i *does not contain repeated variables*;

9      $F = F \cup \{C_i\}$

The function sampleVariable(β,n) may be implemented in two ways.

We can compute a vector p such that $p[i] = \sum_{j=1}^{i} j^{-\beta} / \sum_{j=1}^{n} j^{-\beta}$ at the beginning of the algorithm. Then, every time we call sampleVariable, we compute a random number r uniformly distributed in $[0, 1)$, using a dichotomic search, look for the smallest i such that $p[i] > r$, and return such i.

Alternatively, if n is big we can use the following approximated algorithm. If we want to generate numbers x with probability density $f(x)$, we can integrate $F(x) = \int f(x)\,dx$, find the inverse function, and compute $F^{-1}(y)$, where y is a uniformly random number in $[0, 1]$. Our probability function is discrete. However, when $0 < \beta < 1$, and both $X \to \infty$ and $n \to \infty$, we can approximate it as

$$P(x \le X) = \frac{\sum_{i=1}^{X} i^{-\beta}}{\sum_{i=1}^{n} i^{-\beta}} \approx \frac{\zeta(\beta) + 1/(1-\beta)\,X^{1-\beta}}{\zeta(\beta) + 1/(1-\beta)\,n^{1-\beta}} \tag{5}$$

Therefore, computing the inverse, sampleVariable may be computed as

$$X = \left\lfloor \left(\left(n^{1-\beta} + (1-\beta)\zeta(\beta)\right) Y - (1-\beta)\,\zeta(\beta) \right)^{1/(1-\beta)} \right\rfloor + 1 \tag{6}$$

where Y is a uniform random variable in $[0, 1)$. This way, avoiding the use of the vector p and the dichotomic search, we save a $\mathcal{O}(\log n)$ factor in the time-complexity and a $\mathcal{O}(n)$ factor in the space-complexity of the generator.

4. Industrial SAT Instances

In the previous section, we introduced scale-free random SAT instances. We want this model to generate formulas as close as possible to industrial ones. Therefore, we want to compute the value of β that best fits industrial instances. For this purpose, we have studied the 100 benchmarks (all industrial) used in the SAT Race 2008. All together, they contain n = 25,693,792 variables, with a total of $\sum_{i=1}^{n} K_i$ = 349,760,681 occurrences. Therefore, the average number of occurrences per variable is $E[K_i] = \sum_{i=1}^{n} K_i / n = 13.6$. If we used the classical (uniform) random model to generate instances with this average number of occurrences, most of the variables would have a number of occurrences very close to 13.6. However, in the analyzed industrial instances, close to 90% of the variables have less than this number of occurrences, and more than 60% have 6 or fewer occurrences. The big value of the average is produced by a small fraction of the variables that have a huge number of occurrences. This indicates that the number of occurrences could be better modeled with a power-law distribution. This was already suggested by Boufkhad et al. [34].

In order to check if those industrial instances (all together) are scale-free SAT formulas, and estimate the value of β, we compute the number of occurrences of each variable of each industrial instance. Then, we rename the indexes of such variables such that $K_i \ge K_{i+1}$,

for $i = 1, \ldots, n - 1$. Now, before comparing K_i with $i^{-\beta} / \sum_{j=1}^{n} j^{-\beta}$, we renormalize both functions such that both are defined in $[0, 1]$ and its integral in this range is 1. Hence, we define for the empirical K_i, the *empirical* function ϕ^{ind} as

$$\phi^{ind}(x) =_{def} \frac{n}{\sum_{j=1}^{n} K_j} K_{\lfloor n\,x \rfloor} \tag{7}$$

and, for the theoretical function $P(i)$, the *theoretical* function $\phi(x; \beta, n)$ as

$$\phi(x; \beta, n) = \frac{n}{\sum_{j=1}^{n} j^{-\beta}} \lfloor n\,x \rfloor^{-\beta} \approx \frac{n}{\zeta(\beta) + \frac{1}{1-\beta} n^{1-\beta}} (n\,x)^{-\beta} = \frac{1 - \beta}{\frac{(1-\beta)\zeta(\beta)}{n^{1-\beta}} + 1} x^{-\beta}$$

when $\beta < 1$, we obtain

$$\phi(x; \beta) = \lim_{n \to \infty} \phi(x; \beta, n) = (1 - \beta) x^{-\beta} \tag{8}$$

In Figure 1, we represent both functions with normal axes and with double-logarithmic axes. Notice that in double logarithmic-axes, the slope of $\phi^{ind}(x)$ allows us to estimate the value of $\beta = 0.82$.

Theorem 1 allows us to ensure that the distribution of frequencies on the number of occurrences of variables follows a power-law distribution, with exponent $\delta = 1/0.82 + 1 = 2.22$.

Finally, we have generated a scale-free random 3-SAT formula with $n = 10^7$ variables, $m = 2.5 \times 10^7$ clauses and $\beta = 0.82$. In Figure 2, we show the frequencies of occurrences of variables of this formula and compared it with those obtained for the SAT Race 2008, and the line with slope $\alpha = 1/0.82 + 1 = 2.22$.
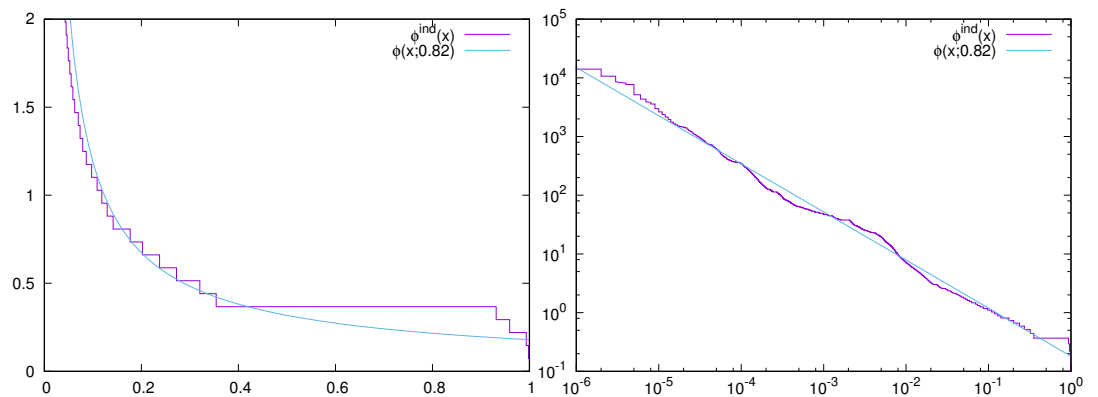
Figure 1. Estimated industrial function $\phi^{ind}(x)$ (in red) and power-law function $\phi(x; 0.82) = (1 - 0.82)\,x^{-0.82}$ (in blue), with normal axes (**left**) and double-logarithmic axes (**right**).
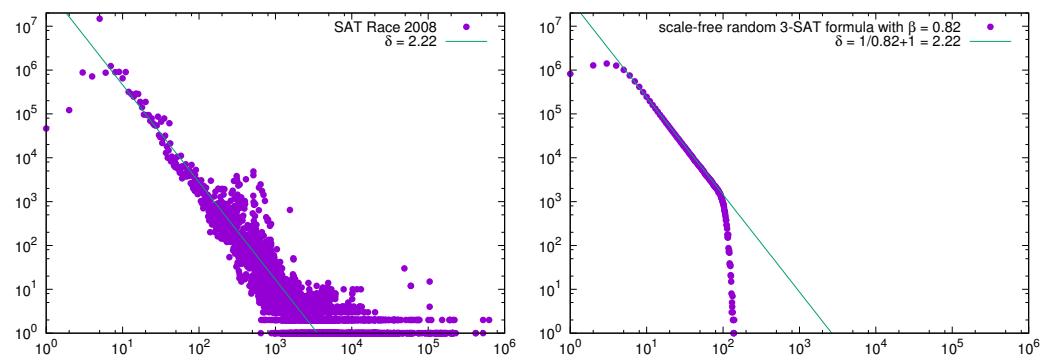
Figure 2. Comparison of the frequencies of variable occurrences obtained for the whole set of instances used in the SAT Race 2008, and for a scale-free random 3-SAT formula generated with

$\beta = 0.82$, $n = 10^7$ and $m = 2.5 \times 10^7$. In both cases, the x-axis represents the number of occurrences, and the y-axis the number of variables with this number of occurrences. Both axes are logarithmic. It also shows the line with slope $\alpha = 1/0.82 + 1 = 2.22$, corresponding to the function $f(x) = C x^{-2.22}$ in double-logarithmic axes.

5. Phase Transition in Scale-Free Random 2-SAT Formulas

Chvátal and Reed [35] proved that a random formula with $(1 + o(1))cn$ clauses of size 2 over n variables is satisfiable with probability $1 - o(1)$, when $c < 1$, and unsatisfiable with probability $1 - o(1)$, when $c > 1$, where $o(1)$ represents a quantity tending to zero as n tends to infinity.

As we will see in this section, a similar result for scale-free random 2-SAT formulas can be obtained using percolation and mean-field techniques.

Percolation theory describes the behavior of connected components in a graph when we remove edges randomly. Erdös and Rényi [36] are considered the initiators of this theory. In this seminal paper on graph theory, they proposed a random graph model $G(n, m)$ where all graphs with n nodes and m edges are selected with the same probability. Gilbert [37] proposed a similar model $G(n, p)$ where n is also the number of nodes, and every $\binom{n}{2}$ possible edge is selected with probability p. For not very sparse graphs (when $p\,n^2 \to \infty$), both models have basically the same properties, taking $m = \binom{n}{2}p$. Erdös and Rényi [38] also studied the connectivity on these graphs and proved that

- when $n\,p < 1$, i.e., $m < n/2$, a random graph almost surely has no connected component larger than $\mathcal{O}(\log n)$;
- when $n\,p = 1$, i.e., $m = n/2$ a largest component of size $n^{2/3}$ almost surely emerges;
- when $n\,p > 1$, i.e., $m > n/2$, the graph almost surely contains a unique giant component with a fraction of the nodes and no other component contains more than $\mathcal{O}(\log n)$ nodes.

Phase transition is a phenomenon that has been observed and studied in many AI problems. Many problems have an order parameter that separates a region of solvable and unsolvable problems, and it has been observed that hard problems occur at critical values of this parameter. Mitchell et al. [39] found this phenomenon in 3-SAT when the ratio between the number of clauses and variables is $m/n \approx 4.3$. Gent and Walsh [40] observed the same phenomenon with clauses of mixed length.

There is a close relationship between SAT problems and graphs. Both, percolation on graphs and phase transition in SAT (or other AI problems) are critical phenomena and both can be studied using mean-field techniques from statistical mechanics. Percolation theory has been used and inspired works in the literature about random SAT and satisfiability threshold, e.g., in Achlioptas et al. [41], to determine the satisfiability threshold of 1-in-k SAT and NAE 3-SAT formulas. Some results on graphs have been previously extended to 2-SAT. For instance, Sinclair and Vilenchik [42] adapted Achlioptas processes for graphs into formulas. Bollobás et al. [43] investigated the scaling window of the 2-SAT phase transition, finding the critical exponent of the order parameter and proving that the transition is continuous, adapting results of Bollobás [44] for Erdös-Rényi graphs. The relationship between percolation in random graphs and phase transition in random 2-SAT formulas is suggested in many other works. For instance, Monasson et al. [45] when studying the phase transition in $2 + P$-SAT (a mixture of $(1 - p)m$ clauses of size 2 and $p\,m$ clauses of size 3) mentioned that *"It is likely that the 2SAT transition results from percolation of these loops..."*. Cooper et al. [17] use the emergence of a giant component in a graph to prove the existence of a phase transition in 2-SAT random formulas with prescribed degrees, using the configuration model. They find, for this model, the same criterion as Friedrich et al. [12] and us in Theorem 2.

Given a random 2-SAT formula with m clauses over n variables, we can construct an Erdös-Rényi graph where the $2\,n$ literals are nodes, and the m clauses are edges. At the percolation point $m = (2\,n)/2$ of this graph, a *giant component* emerges. Just at the same

point $m = n$, the 2-SAT phase transition threshold is located. However, despite the coincidence in the point, the relation between both facts is not direct: a giant component in the graph is not the same as a giant (hence unsatisfiable) loop of implications in the SAT formula. The connection between two edges $a \leftrightarrow b$ and $b \leftrightarrow c$ in the graph is given by a common node (literal) b, whereas, in the SAT formula, the resolution between $a \vee b$ and $\neg b \vee c$ is through a variable b that is affirmed in one clause and negated in the other. In this section, we elaborate on the relation of giant components in graphs and unsatisfiability proofs in 2-SAT formulas.

5.1. A Criterion for Phase Transition in 2-SAT

Unsatisfiability proofs of 2-SAT formulas are characterized by *bicycles*. Let F be a 2-SAT formula. Any sequence of literals $x_1, \ldots, x_s$ satisfying $\neg x_i \vee x_{i+1} \in F$, for any $i = 1, \ldots, s - 1$, is called an *implication sequence*. We say that y *implies* y', if there exists an implication sequence of the form $y, x_1, \ldots, x_n, y'$. Any implication sequence of the form $x_1, \ldots, x_s, x_1$ is called a *cycle*. A *bicycle* is a cycle $x_1, \ldots, x_n, x_1$ such that there exists a variable a satisfying $\{a, \neg a\} \subseteq \{x_1, \ldots, x_n\}$.

A 2-SAT formula is unsatisfiable if, and only if, it contains a bicycle [35,46].

We will also consider random graphs with n nodes and m edges (We will deal with distinct models of random graphs where every graph has a distinct probability of being chosen.), and connected components, defined as subsets of nodes such that any pair of them is connected by a path inside the component. A random graph of size n is said to contain a *giant connected component* if almost surely (Almost surely means that, in the model of random graphs, as $n \to \infty$, the probability tends to one.) it contains a connected component with a positive fraction of the nodes. Given a model of random graphs, we say that c is the percolation threshold if any random graph with n nodes and more than $c\,n$ edges almost surely contains a giant component. In a random graph, the degree of a node x, noted as k_x, is a random variable. The random variable k represents the degree of a random node chosen with a uniform probability. (In some of the models of random graphs that we will consider, not all degrees of nodes follow the same probability distribution. Therefore, we will distinguish between k and k_x.)

As we commented above, we can represent any 2-SAT formula as a graph where nodes are literals, and clauses $a \vee b$ are edges between literals a and b. In classical 2-SAT random formulas, since literals are chosen independently with uniform probability, the generated graph will be an Erdös-Rényi graph following the model $G(2\,n, m)$. However, a connected component in the graph is not necessarily an unsatisfiability proof of the formula.

First, in a random SAT formula, we may have repeated clauses, which means that from m clauses we will obtain less than m edges. However, for a linear number of clauses, when $\beta < 1/2$, there are $(1 - o(1))\,m$ distinct clauses or edges. In the classical case, in the limit $n \to \infty$, with a linear number of clauses $m = \mathcal{O}(n)$, and a quadratic number of possible clauses, the probability of any clause is $\mathcal{O}(n^{-2})$, and the probability of being repeated $m\,\mathcal{O}(n^{-2}) = \mathcal{O}(n^{-1})$. Therefore, the fraction of repeated clauses is negligible. For scale-free 2-CNF formulas, in Theorem 5, we will see that, if $\beta < 1/2$, then clauses have probability $o(n^{-1})$. Precisely, the most probable 2-CNF clause is $x_1 \vee x_2$. This means, that after generating $\mathcal{O}(n)$ clauses, the probability that a newly generated clause has already been generated previously is bounded by

$$P(x_1 \vee x_2)\,\mathcal{O}(n) \approx \frac{2!\,1^{-\beta}2^{-\beta}}{(2\sum_{i=1}^{n} i^{-\beta})^2}\,\mathcal{O}(n) = \frac{2^{-1-\beta}}{(\zeta(\beta) + \frac{n^{1-\beta}}{1-\beta} + \mathcal{O}(n^{-\beta}))^2}\,\mathcal{O}(n) = \mathcal{O}(n^{2\beta-1})$$

This probability bounds the value of the fraction of repeated clauses, which is meaningless when $\beta < 1/2$.

Second, connected components and cycles have not the same structure. Therefore, the existence of a giant connected component and the existence of a giant cycle are independent facts.

Molloy and Reed [47,48] and Cohen et al. [49,50] have studied the existence of giant components in random graphs with heterogeneous and fixed node degrees. Molloy and Reed [47] prove that the critical point is at

$$Q(\lambda) = \sum_{i>0} i(i-2)\lambda_i = 0 \tag{9}$$

where λ_i is the fraction of nodes with degree i. Cohen et al. [49] independently prove (but in a much more informal way) that the critical point is characterized by

$$\frac{\mathrm{E}[k^2]}{\mathrm{E}[k]} = 2 \tag{10}$$

where k is the degree of a random node, and E denotes expectation. It is easy to see that both criteria are exactly the same. Interestingly, the criterion depends not only on the expected degree of nodes, but also on the expected square degree of the nodes, and hence on the variability of node's degrees. The variability on the node degrees plays an important role in the location of the percolation threshold. For instance, in the Erdös-Rényi model, the percolation threshold is located at $m/n = 1/2$; hence, the expected degree of nodes is $1/2$. However, the expected degree of nodes belonging to the same connected component (recall that minimally connected components are trees, where the number of edges is equal to the number of nodes minus one) of size r is, at least, $(r-1)/r \approx 1$. This discrepancy is only possible if the variability in the node degrees is high. This also explains why, in regular random formulas, where we impose variables to occur exactly the same number of times (instead of the same average number of times), we find distinct phase transition thresholds.

Cohen et al. [49] starts assuming that loops of connected nodes may be neglected. In this situation, the percolation transition takes place when a node i, connected to a node j in the connected component, is also connected in average to at least one other node, i.e., when $\mathrm{E}[k_i \mid i \leftrightarrow j] = \sum_{k_i} k_i P(k_i \mid i \leftrightarrow j) = 2$.

Molloy and Reed [47] give a more detailed proof that we will try to summarize. Given the list of fixed degrees k_i of every node, they describe a random algorithm that constructs (exposes) all graphs compatible with these degrees with the same probability, exposing connected components one by one:

Let c_i be the degree of node i on the partially exposed graph. Initially, set $c_i = 0$, for every node. Then, until $c_i = k_i$, for all nodes, repeat the following actions. If, for some node i, we have $0 < c_i < k_i$, then (case A) select it; otherwise, (case B) choose freely a node i such that $c_i = 0$. Then, in both cases, choose another node $j \neq i$ with probability $P(j) \sim k_j - c_j$. Expose the edge $i \leftrightarrow j$, and increase c_i and c_j. Notice that every time we execute case B, we start the exposition of a new connected component of the graph.

Let X_r be the random variable representing the number of *open vertexes* in partially exposed nodes, i.e., $X_r = \sum_{c_i>0} k_i - c_i$, after the rth edge $i \leftrightarrow j$ has been exposed. Notice that we execute case B when we have $X_{r-1} = 0$, and we obtain $X_r = k_i + k_j - 2$. When we execute case A, there are two situations: (case A1) if node j is a partially exposed node (i.e., $0 < c_j$), then $X_r = X_{r-1} - 2$, and (case A2) if node j has never been exposed (i.e., $c_j = 0$), then $X_r = X_{r-1} + k_j - 2$.

Suppose that cases B and A1 do not happen very often. Then, the expected change in X_r is

$$\mathrm{E}[X_r - X_{r-1}] \approx \frac{\sum_j k_j(k_j - 2)}{\sum_j k_j} = \frac{Q(\lambda)}{\mathrm{E}[k]}$$

and, since $X_r - X_{r-1} \geq -1$, a standard result of random walk theory ensures that if $Q(\lambda) > 0$ then, after $\Theta(n)$ steps, X_r is almost surely of order $\Theta(n)$; and if $Q(\lambda) < 0$, then X_r returns to zero fairly quickly. In the first case, we generate a giant connected component of size $\Theta(n)$, and in the second case, no component is larger than $\mathcal{O}(\log n)$. In order to prove that executions of case A1 do not hurt, Molloy and Reed prove that the probability of

choosing a partially exposed node (a node with $c_j > 0$) is negligible unless we have already exposed a fraction $\Theta(n)$ of the nodes in the current connected component.

Theorems 2 and 3 establish a similar criterion for the existence of a giant set of implied literals from a given one. This almost surely implies the unsatisfiability of the formula. The proof of the theorem resembles Molloy and Reed's and Cohen et al.'s proofs. In Theorem 2 we fix the number of occurrences of every literal, whereas in Theorem 3, we fix the number of occurrences of variables. Compared with the definition of Q in Molloy and Reed's, we observe that in Theorem 3, the 2 is replaced by a 3. In Theorem 2, we combine the number of literals k_i with the number of their negated k_{-i}, and the constant is a 1 instead of a 2. Notice that the condition, in this case, is equal to the condition found by Cooper et al. [17] for the configuration method and prescribed literal degrees.

Theorem 2. *Let F be a 2-CNF formula generated in a random model with variables $\{x_1, \ldots, x_n\}$, where every literal x_i (resp $\neg x_i$) is selected with probability p_i (resp p_{-i}), and literals in clauses are not correlated (i.e., $P(x_i \vee x_j) \approx p_i p_j$). Assume that $p_i = o(1)$ and $m = \mathcal{O}(n)$. Let $k_i = 2 m p_i$ be the expected number of occurrences of literal x_i. Then, if $\sum_{i=-n}^{n} k_i(k_{-i} - 1) > 0$, then almost surely F is unsatisfiable.*

Proof. The proof resembles Molloy and Reed's proof for the percolation threshold on graphs. This proof is quite long, and our proof does not differ very much. Therefore, we will only sketch it.

In our case, we do not deal with connected components. In fact, we do not expose the random formula with our algorithm. We assume that we already have the formula, and we describe in Algorithm 2 how to enumerates the set of literals implied by a given initial literal x.

Algorithm 2: Algorithm for finding literals implied by x

> **Input:** F, x
> 1 **foreach** y **do**
> 2 $c_y = 0; o_y = false$
>
> 3 $o_x = true;$
> 4 **while** $(\exists y.o_{\neg y} = true \wedge c_y < k_y) \wedge \neg(\exists y.o_y = o_{\neg y} = true)$ **do**
> 5   select $y \vee z \in F$ such that $o_{\neg y} = true;$
> 6 $o_z := true;$
> 7 $c_y := c_y + 1;$
> 8 $c_z := c_z + 1;$
> 9 $F := F \setminus \{y \vee z\}$
> 10 **return** $\{z \mid o_z = true\}$

The Boolean variable o_y denotes if the literal has been reached from the initial literal x and the counter c_y denotes the number of clauses containing y that we have already removed from the formula. Therefore, $k_y - c_y$ is the number of clauses containing y that still remains in F. When x implies y and $\neg y$, for some variable y, we say that x implies a contradiction. In this case, x also implies $\neg x$. The algorithm returns the set of literals implied by x or a contradiction (in this second case, we abort, since we already have $x \rightarrow \neg x$ that is what we want to check). Notice also that $c_y > 0$ implies $o_y = true \vee o_{\neg y} = true$.

Notice that this algorithm is quite similar to Molloy and Reed's algorithm for exposing connected components of a random graph. Counter c_y has a similar meaning and we only require the Boolean variable o_y to denote the condition of *open vertex* (expressed as $c_y > 0$ in Molloy and Reed's algorithm). The algorithm is deterministic if you consider the formula given. However, for a random formula, the algorithm performs exactly the same steps and can be seen as a random algorithm. Similarly, we can define the random variable $X_r = \sum_{o_{\neg x} = true} k_x - c_x$ after iteration r. At every iteration, this variable satisfies:

(case A) $X_r = X_{r-1} - 1$, if $o_z = true$ and $o_{\neg z} = false$;

(case B) $X_r = X_{r-1} - 1 + k_{\neg z}$, if $o_z = o_{\neg z} = false$;

(case C) $X_r = X_{r-1} - 2 + k_{\neg z} - c_{\neg z}$, if $o_z = false$ and $o_{\neg z} = true$.

Notice that line 7 decreases X_r in 1, line 8 decreases X_r in 1, when $o_{\neg z} = true$, and line 6 increases X_r in $k_{\neg z} - c_{\neg z}$, when $o_z = false$. However, if both o_z and $o_{\neg z}$ are false, then $c_{\neg z}$ is zero. After case C, we find a contradiction and finish. In case B, the expected gain in X_r is

$$E[X_r - X_{r-1}] \approx \frac{\sum_z k_z \, (k_{\neg z} - 1)}{\sum_z k_z}$$

In case A, the random variable only decreases by one. Like Molloy and Reed's work, we can also argue that the case A is negligible, unless we have already added to the set of implied literals a constant fraction of them.

Therefore, reproducing all the lemmas of Molloy and Reed's proof, we can conclude that, when $\sum_z k_z \, (k_{\neg z} - 1) > 0$, almost surely there exists a constant $0 < c < 1$ such that for a fraction c of initial literals x, the set of literals implied by x is a fraction c of all literals or contains a contradiction, and hence, X implies $\neg x$. For a particular variable x, the probability that x implies $\neg x$ and $\neg x$ implies x is at least c^4. The probability that the formula is satisfiable is at most $(1 - c^4)^n$, which tends exponentially to zero as n tends to infinity. $\quad\square$

Theorem 3. *Let F be a 2-CNF formula generated in a random model with variables* $\{x_1, \dots, x_n\}$, *where every variable* x_i *is selected with probability* P_i *and negated with probability* $1/2$, *and variables in clauses are not correlated. Assume that* $P_i = o(1)$ *and* $m = \mathcal{O}(n)$. *Let* $K_i = 2\, m\, P_i$ *be the expected number of occurrences of variable* x_i. *Then, if* $\sum_{i=1}^n K_i(K_i - 3) > 0$, *then almost surely F is unsatisfiable.*

The condition $\sum_{i=1}^n K_i(K_i - 3) > 0$ *is equivalent to* $E[K^2]/E[K] > 3$

Proof. The proof is, like in Theorem 2, based on the proof of Molloy and Reed [47]. In this case, however, the expected gain in the random variable X_r is given by:

$$E[X_r - X_{r-1}] = \frac{\sum_{i=1}^n \frac{K_i}{2}\left(\frac{K_i - 1}{2} - 1\right)}{\sum_{i=1}^n \frac{K_i}{2}}$$

since $\frac{K_i}{2}$ is the expected value of k_i and $\frac{K_i - 1}{2}$ is the expected value of k_{-i} conditioned to the existence of one positive occurrence of x_i. Then, the condition $E[X_r - X_{r-1}] > 0$ is equivalent to $\sum_{i=1}^n K_i(K_i - 3) > 0$. $\quad\square$

For the proof of Theorem 3, we could also use the argument of Cohen et al. [49]. In the case of graphs, we obtain a giant connected component when a node i, connected to a node j, is also connected in average to at least one other node. Formally, when the expected degree of i, conditioned to the fact that i and j are connected, is $E[k_i \mid i \leftrightarrow j] = 2$.

In our case, in order for a giant cycle to emerge, when there is a clause $x \vee y$, we have to find, at least, another clause containing $\neg x$. In this situation, the expected number of other clauses containing x is 2, that added to the original clause $x \vee y$, gives a minimum number of 3 clauses containing x. Given a pair of literals x and y, let $\pm x \vee y$ express the fact: "$x \vee y \in F$ or $\neg x \vee y \in F$. Formally, our criterion can be written as

$$E[K_x \mid \pm x \vee y] > 3 \tag{11}$$

This criterion is the necessary and sufficient condition to continue the construction of a set of clauses, ensuring that the probability that this set contains a fraction of the literals tends to one.

Using Bayes, we have

$$
\begin{aligned}
\mathrm{E}[K_x \mid \pm x \vee y] &= \sum_{k=0}^{\infty} k\, P(K_x = k \mid \pm x \vee y) \\
&= \sum_{k=0}^{\infty} k\, \frac{P(K_x = k \wedge \pm x \vee y)}{P(\pm x \vee y)} = \sum_{k=0}^{\infty} k\, \frac{P(\pm x \vee y \mid K_x = k) P(K_x = k)}{P(\pm x \vee y)}
\end{aligned}
$$

Given a pair of literals x and y, the probability that either $x \vee y$ or $\neg x \vee y$ are one of the clauses of the formula, conditioned by the fact that the number of occurrences of variable x is k (and assuming that clauses are not repeated) is: $P(\pm x \vee y \mid K_x = k) = \frac{k}{2(n-1)}$ and, the probability of the same fact without condition: $P(\pm x \vee y) = \frac{\mathrm{E}[K_x]}{2(n-1)}$. Therefore,

$$
\begin{aligned}
\mathrm{E}[K_x \mid \pm x \vee y] &= \sum_{k=0}^{\infty} k\, \frac{\frac{k}{2(n-1)} P(K_x = k)}{\frac{\mathrm{E}[K_x]}{2(n-1)}} = \frac{\sum_{k=0}^{\infty} k^2\, P(K_x = k)}{\mathrm{E}[K_x]} \\
&= \frac{\mathrm{E}[K^2]}{\mathrm{E}[K]} > 3
\end{aligned}
\tag{12}
$$

defines the unsatisfiability threshold.

The previous theorems ensure that, when the criterion is satisfied, there is a giant bicycle containing a fraction of the literals, and the formula is unsatisfiable. However, if the formula is unsatisfiable, it can be due to a *small* bicycle. Therefore, the reverse implication is not necessarily true. In other words, Theorems 2 and 3 establish a sufficient (but not necessary) condition for unsatisfiability of random 2-SAT formulas, which results in an upper bound for the phase transition point. However, we conjecture that, either giant bicycles are more probable than small bicycles and the percolation threshold (obtained with the criterion) is equal to the phase transition point, or, if small bicycles are more probable, the phase transition point is at $c = 0$.

5.2. Classical 2-SAT Formulas

Theorems 2 and 3 may be used to find the phase transition point in terms of the number of clauses divided by the number of variables. In this subsection, we apply the technique to (classical) random 2-SAT formulas.

We start with a formula (or graph), not necessarily at the critical threshold. Then, we apply a percolation process where a fraction $1 - p$ of randomly selected clauses (edges) are removed, such that the remaining p fraction of edges are in the critical threshold. If we start with the complete formula with all possible $2^2 \binom{n}{2}$ clauses over n variables, and remove clauses with uniform probability, this process generates a (classical) random 2-SAT formula in the SAT–UNSAT transition point (except for the lack of repeated clauses).

If k_x' is the number of occurrences of literal x in the original graph, then, after removing the $(1 - p)$ fraction, the new distribution on the number of occurrences is $P(k_x) = \sum_{k_x' = k_x}^{\infty} P(k_x') \binom{k_x'}{k_x} p^{k_x} (1 - p)^{k_x' - k_x}$. Using this binomial distribution we obtain the moments $\mathrm{E}[k_x] = p\, \mathrm{E}[k_x']$ and $\mathrm{E}[k_x^2] = p^2\, \mathrm{E}[(k_x')^2] + p(1 - p)\mathrm{E}[k_x']$ for any literal x. Since $K_x = k_x + k_{\neg x}$, and k_x and $k_{\neg x}$ are independent variables with the same distribution, we have $\mathrm{E}[K_x] = 2\, \mathrm{E}[k_x]$ and $\mathrm{E}[K_x^2] = 2\, \mathrm{E}[k_x^2] + 2\, E^2[k_x]$. If we impose the criterion of Theorem 3 to this new formula, we obtain

$$
\frac{\mathrm{E}[K^2]}{\mathrm{E}[K]} = \frac{2\, p^2\, \mathrm{E}[(k')^2] + 2\, p(1 - p)\mathrm{E}[k'] + 2\, p^2\, E^2[k']}{2\, p\, \mathrm{E}[k']} = 3
$$

Hence,

$$
p = \frac{2}{\frac{\mathrm{E}[(k')^2]}{\mathrm{E}[k']} + \mathrm{E}[k'] - 1}
$$

For the complete formula we have $k'_x = 2(n-1)$ for any literal, therefore $p = 1/(2n - 5/2)$. The expected number of clauses in the phase transition threshold is

$$\mathrm{E}[m] = 2^2 \binom{n}{2} p = \frac{2n(n-1)}{2n - 5/2} = n + \mathcal{O}(1)$$

This proves that the clause/variable fraction at the 2-SAT phase transition threshold is at most $m/n = 1$, reproducing the results of Chvátal and Reed [35].

For the expected moments, we obtain $\mathrm{E}[k] \approx 1$ and $\mathrm{E}[k^2] \approx 2$, for the number of occurrences of literals, and $\mathrm{E}[K] \approx 2$ and $\mathrm{E}[K^2] \approx 6$, for the number of occurrences of variables.

Now, consider the case of (classical) regular random 2-SAT formulas. These are random formulas where the number of occurrences of a literal minus the number of occurrences of another literal is, at most, one. Assume that all literals have exactly the same number of occurrences $k_x = m/n$. Applying Theorem 2, without any need of percolation process, we obtain $\sum_{i=-n}^{n} k_i(k_{-i} - 1) = 2n \frac{m}{n}(\frac{m}{n} - 1) = 0$. Therefore, $m/n = 1$ is an upper bound for the phase transition point, reproducing the results of Boufkhad et al. [51]. Notice that, in this case, the conditions of Theorem 3 are not fulfilled: k_x and $k_{\neg x}$ are not independent random variables. If we consider the proof of this Theorem, since in a random regular formula $k_x = k_{\neg x}$, if this formula contains a clause $x \lor y$, we only need to require that $\mathrm{E}[K_x \mid x \lor y] = 2$ in order to ensure that there is another clause containing $\neg x$. With this new criterion, and reproducing the proof of Theorem 3, we obtain that the threshold in a regular random formula is $\frac{\mathrm{E}[K^2]}{\mathrm{E}[K]} = 2$.

5.3. Scale-Free 2-SAT Formulas

Recently, Friedrich et al. [12] proved that scale-free random 2-SAT formulas with exponent $\delta > 3$ and clause/variable ratio $m/n < \frac{(\delta-1)(\delta-3)}{(\delta-2)^2}$ are satisfiable with probability $1 - o(1)$. (In their paper, they write β instead of δ, but we prefer to use β with the same meaning as in [11].) This gives a lower bound for a possible phase transition point, in terms of δ. They conjecture that this bound is tight and that this phase transition exists. Replacing $\delta = 1/\beta + 1$ (according to Theorem 1) in this inequality, we obtain:

Scale-free random 2-SAT formulas with exponent $\beta < 1/2$ and clause/variable ratio $m/n < \frac{1-2\beta}{(1-\beta)^2}$ are satisfiable with probability $1 - o(1)$.

In the first statement of the following theorem, we prove that when the clause/variable ratio exceeds this value, formulas are almost surely unsatisfiable.

Theorem 4. *(1) Scale-free random 2-SAT formulas with exponent $\beta < 1/2$ and clause/variable ratio*

$$m/n > \frac{1 - 2\beta}{(1 - \beta)^2}$$

are unsatisfiable with probability $1 - o(1)$.
(2) Scale-free random 2-SAT formulas over n variables, exponent $\beta = 1/2$ and more that

$$4 n \log^{-1} n + \mathcal{O}(n^{1/2} \log^{-1} n)$$

distinct clauses, or exponent $1/2 < \beta < 1$, and more than

$$\frac{1}{(1 - \beta)^2 \zeta(2\beta)} n^{2(1-\beta)} + \mathcal{O}(n^{1-\beta})$$

distinct clauses, are unsatisfiable with probability $1 - o(1)$.

Proof. In the case of scale-free formulas, we cannot start the percolation process from the complete formula, since the uniform-random deletion of clauses does not give rise to

scale-free formulas. Therefore, we can simply impose the criterion on the original formula. We will do all the computations using the number of occurrences of variables K_x, instead of the number of occurrences of the literal k_x, and applying Theorem 3.

Since $\beta < 1$, by Lemma 1, repetitions of variables in clauses may be neglected, and the probability that a particular literal in the formula corresponds to the variable x is given by $P(x) \approx \frac{x^{-\beta}}{\sum_{i=1}^{n} i^{-\beta}}$. Since the election of every variable for every possible literal of the formula is independent, the number of occurrences of x follows a binomial distribution

$$P(K_x = K) \approx \binom{2m}{K} \left(\frac{x^{-\beta}}{\sum_{i=1}^{n} i^{-\beta}} \right)^K \left(1 - \frac{x^{-\beta}}{\sum_{i=1}^{n} i^{-\beta}} \right)^{2m-K}$$

In the limit $m \to \infty$, the distribution approaches a Poisson distribution where

$$E[K_x] \approx \frac{x^{-\beta}}{\sum_{i=1}^{n} i^{-\beta}} 2m$$

$$E[K_x^2] \approx \left(\frac{x^{-\beta}}{\sum_{i=1}^{n} i^{-\beta}} 2m \right)^2 + \frac{x^{-\beta}}{\sum_{i=1}^{n} i^{-\beta}} 2m$$

Recall that in scale-free formulas K_x follows a distinct probability distribution for every variable x; therefore, we have to average over all variables

$$E[K] \quad = \frac{1}{n} \sum_{x=1}^{n} E[K_x] \approx \frac{1}{n} \sum_{x=1}^{n} \frac{x^{-\beta}}{\sum_{i=1}^{n} i^{-\beta}} 2m = \frac{2m}{n}$$

$$E[K^2] \quad = \frac{1}{n} \sum_{x=1}^{n} E[K_x^2] \approx \frac{1}{n} \sum_{x=1}^{n} \left[\left(\frac{x^{-\beta}}{\sum_{i=1}^{n} i^{-\beta}} 2m \right)^2 + \frac{x^{-\beta}}{\sum_{i=1}^{n} i^{-\beta}} 2m \right]$$

$$\approx \frac{4m^2}{n} \frac{\sum_{x=1}^{n} x^{-2\beta}}{\left[\sum_{i=1}^{n} i^{-\beta} \right]^2} + \frac{2m}{n}$$

Imposing the criterion $E[K^2]/E[K] = 3$, we obtain

$$m \approx \frac{\left[\sum_{i=1}^{n} i^{-\beta} \right]^2}{\sum_{x=1}^{n} x^{-2\beta}}$$

Applying Equations (3) and (4), we obtain

$$m = \begin{cases} \dfrac{1-2\beta}{(1-\beta)^2} n + \mathcal{O}(n^\beta) & \text{if } \beta < 1/2 \\[2ex] 4\,n \log^{-1} n + \mathcal{O}(n^{1/2} \log^{-1} n) & \text{if } \beta = 1/2 \\[2ex] \dfrac{1}{(1-\beta)^2 \zeta(2\beta)} n^{2(1-\beta)} + \mathcal{O}(n^{1-\beta}) & \text{if } 1/2 < \beta < 1 \\[2ex] \dfrac{1}{\zeta(2)} \log^2 n + \mathcal{O}(n^{-1} \log n) & \text{if } \beta = 1 \\[2ex] \dfrac{\zeta^2(\beta)}{\zeta(2\beta)} + \mathcal{O}(n^{1-\beta}) & \text{if } 1 < \beta \end{cases} \tag{13}$$

The last two cases are meaningless since we have assumed that $\beta < 1$ in other parts of the proof. The first three possibilities prove the two statements of the theorem. In the second and third cases, since we cannot prove that the fraction of repeated clauses is meaningless, we obtain a bound on the number of distinct clauses. $\qquad \square$

From Friedrich et al. [12] and Theorem 4, we can conclude:

Corollary 1. *Scale-free 2-SAT formulas over n variables and exponent $\beta < 1/2$ have a SAT–UNSAT phase transition threshold when the variable/clauses ratio is*

$$m/n = \frac{1-2\beta}{(1-\beta)^2}$$

We have experimentally analyzed the fraction of satisfiable random scale-free 2-SAT formulas depending on the parameter β and fraction of clause/variable m/n. The results are plotted in Figure 3, for formulas with $n = 10^5$ variables. We observe that the phase transition predicted by Theorem 4 is quite precise, except when $\beta \gtrsim 1/2$. In the limit $n \to \infty$, the fraction of satisfiable formulas with n variables and $c\,n$ clauses tends to zero when $c > 0$. However, as the number of clauses needed to make the formula unsatisfiable grows as $n^{2(1-\beta)}$, when β is close to $1/2$ the confluence is very slow.
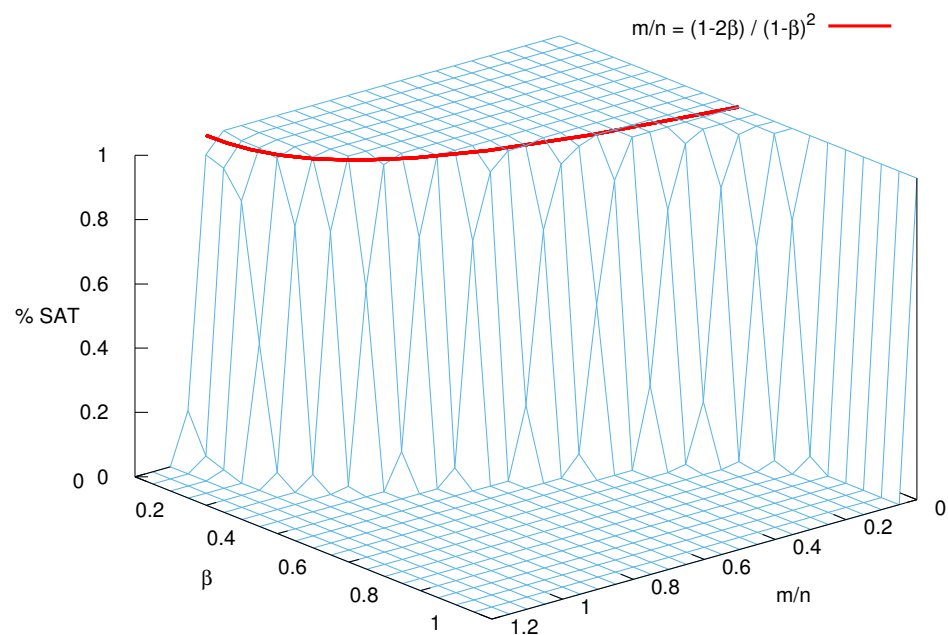
Figure 3. Fraction of satisfiable formulas as a function of parameter β and fraction of clause/variables m/n. The number of variables is $n = 10^5$ and the fraction is approximated repeating the experiment for 10 formulas at every point. We also draw the theoretical threshold $m/n = \frac{1-2\beta}{(1-\beta)^2}$.

In order to test experimentally the second statement of Theorem 4, we have analyzed the fraction of satisfiable formulas with respect to m/α, where $\alpha = \frac{1}{(1-\beta)^2 \zeta(2\beta)} n^{2(1-\beta)}$. In Figure 4, we show the results for $\beta = 0.7$. We observe that, for distinct values of n, the transition between SAT and UNSAT is around α. However, for increasing values of n the transition does not seem to become more abrupt.
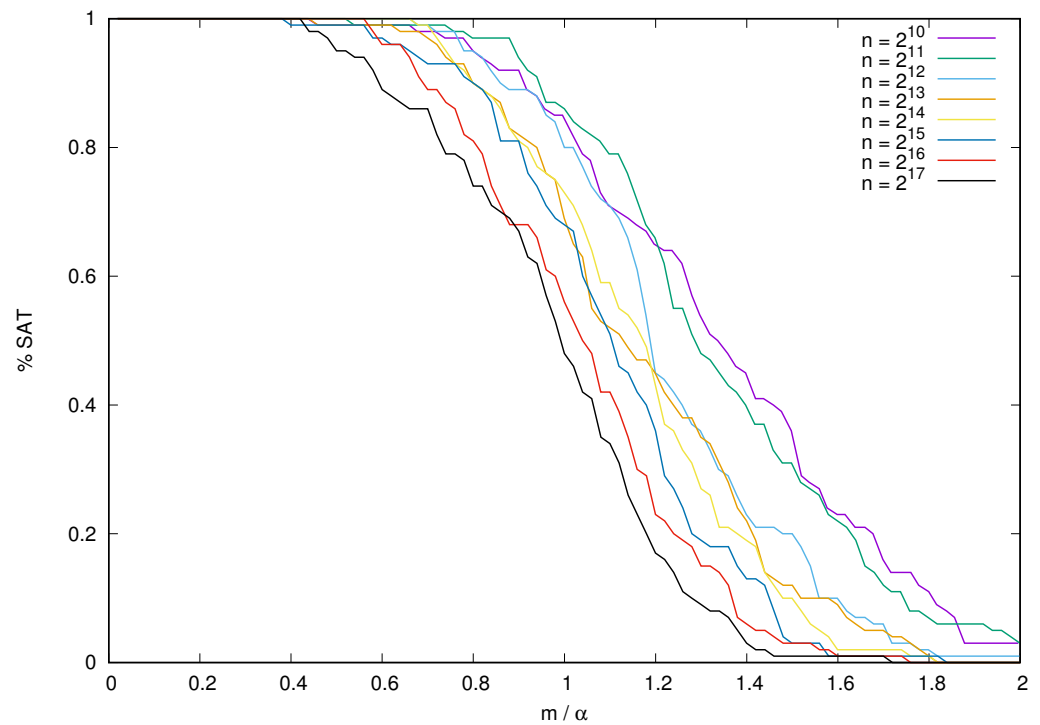
Figure 4. Fraction of satisfiable formulas as a function of m/α, where $\alpha = \frac{1}{(1-\beta)^2 \zeta(2\beta)} n^{2(1-\beta)}$ for $\beta = 0.7$, and distinct values of n between 2^{10} and 2^{17}. Every point is computed repeating the experiment for 100 formulas, and checking how many of them are satisfiable.

6. Unsatisfiability by Small Cores

In the proof of Theorem 4 we have already seen that when $\beta > 1/2$ the number of clauses needed to make a 2-SAT formula unsatisfiable is sub-linear. Therefore, the phase transition factor—understood as a constant c such that, on the limit $n \to \infty$, formulas with less than $c\,n$ clauses are satisfiable, and those with more than $c\,n$ clauses are unsatisfiable— is zero. In this section, we will prove that, when β exceeds a certain bound, scale-free formulas become unsatisfiable due to a small subset of clauses containing variables with small indexes. Moreover, this result holds for clauses of any size.

Theorem 5. *A random scale-free formula over n variables, exponent $0 \le \beta < 1$ and $\omega(n^{(1-\beta)k})$ clauses of size k is unsatisfiable with probability $1 - o(1)$.*

Proof. The probability of a clause only containing the smallest k variables is

$$P(x_1 \vee \cdots \vee x_k) \approx \frac{k!\, 1^{-\beta} \cdots k^{-\beta}}{\left(2 \sum_{i=1}^{n} i^{-\beta}\right)^k}$$

This inequality would be an equality, if we allow tautologies and simplifiable clauses (i.e., repeated variables) in formulas.

Using Eqaution (3), we obtain

$$P(x_1 \vee \cdots \vee x_k) \approx \frac{k!\, 1^{-\beta} \cdots k^{-\beta}}{\left(2 \sum_{i=1}^{n} i^{-\beta}\right)^k} = \frac{(k!)^{1-\beta}}{2^k \left(\zeta(\beta) + \frac{n^{1-\beta}}{1-\beta} + \mathcal{O}(n^{-\beta})\right)^k}$$

In the limit $n \to \infty$, the probability of generating the clause $x_1 \vee \cdots \vee x_k$ after generating $n^{(1-\beta)k}$ independent clauses is

$$1 - \left(1 - \frac{(k!)^{1-\beta}}{2^k \left(\frac{n^{1-\beta}}{1-\beta}\right)^k}\right)^{n^{(1-\beta)k}} = 1 - \left(1 - \frac{\left(\frac{1-\beta}{2}\right)^k (k!)^{1-\beta}}{n^{(1-\beta)k}}\right)^{n^{(1-\beta)k}} \approx 1 - e^{-\left(\frac{1-\beta}{2}\right)^k (k!)^{1-\beta}}$$

Therefore, the probability of generating the clause $x_1 \vee \cdots \vee x_k$ is $1 - o(1)$ when the number of clauses is $m = \omega(n^{(1-\beta)k})$. The same applies for other 2^k clauses with distinct signs, and, if $k = \mathcal{O}(1)$, to a refutation of the formula only using these set of clauses. $\square$

As in classical random formulas, the expected number of truth assignments that satisfy a scale-free random formula is $2^n (1 - 2^{-k})^m$. This imposes a linear upper bound on the number of clauses of satisfiable scale-free formulas, i.e., a random scale-free formula with $m = c\,n$ clauses of size k over n variables such that $c > 2^k \log 2$ is unsatisfiable with probability $1 - o(1)$. Therefore, the bound in Theorem 5 only *improves* this other linear bound when $(1 - \beta)k < 1$, hence when $\beta > 1 - 1/k$.

Figure 5 shows an experimental estimation of how many clauses are needed to make unsatisfiable 50% of the random formulas generated with distinct values of β and $k = 3$, as a function of the number of variables.
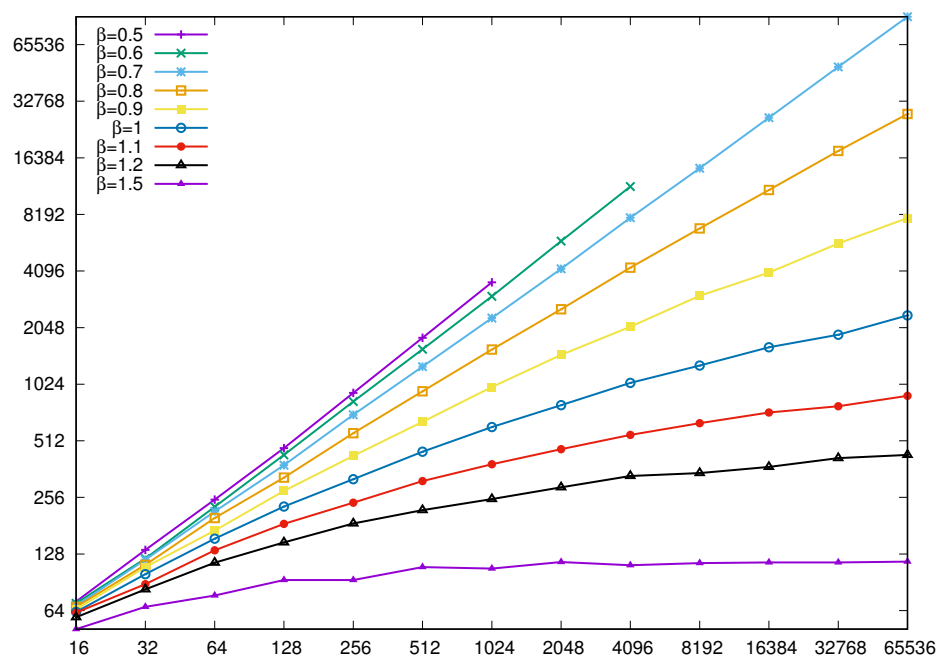

Figure 5. Estimation of the number of clauses that are needed to make unsatisfiable 50% of the formulas generated for distinct values of β and $k = 3$ as a function of the number of variables.

Theorem 5 predicts that the number of clauses in a satisfiable scale-free 2-SAT formula cannot grow faster than $\mathcal{O}(n^{2(1-\beta)})$, due to the emergence of small cores. When $1/2 < \beta < 1$, the second statement of Theorem 4, predicts exactly the same exponent $2(1-\beta)$ for the emergence of a giant bicycle. This suggests that, in this range of β, the probability of the existence of a small and a giant unsatisfiable core of clauses is similar. However, experimental results (see Figure 4) suggest that the SAT–UNSAT transition is quite smooth, like in classical 1-SAT. This suggests that small cores are, in fact, more prominent. Another argument in this direction is as follows:

Let $C(V)$ be the subset of clauses only containing variables of the subset V of variables. The greater $|C(V)|/|V|$ is, the higher is the probability to have an unsatisfiable core inside

$C(V)$. In the case of scale-free random k-SAT formulas, let C_r be the set of clauses only containing variables $\{1, \ldots, r\}$. We can estimate

$$E\left[\frac{|C_r|}{r}\right] = \frac{m}{r}\left(\frac{\sum_{i=1}^{r} i^{-\beta}}{\sum_{i=1}^{n} i^{-\beta}}\right)^k \approx \frac{m}{r}\left(\frac{r^{1-\beta}-1}{n^{1-\beta}-1}\right)^k$$

For $(1-\beta)k \geq 1$, i.e., $\beta < 1 - 1/k$, the maximum of this function is $r = \infty$. For $(1-\beta)k < 1$, i.e., $\beta > 1 - 1/k$, the maximum is finite:

$$r = (1-(1-\beta)k)^{-1/(1-\beta)}$$

Notice that $(1-\beta)k$ is the exponent predicted by Theorem 5, and that for 2-SAT, $1 - 1/k = 1/2$. Therefore, we obtain another proof that at $\beta = 1 - 1/k$ we obtain a change in the behavior of scale-free random k-SAT formulas. When $n \to \infty$, for $\beta \leq 1 - 1/k$ the most probable is to obtain a very large core that involves a fraction of the whole set of clauses. For $\beta > 1 - 1/k$ the most probable is to obtain a small core only involving a finite set of clauses and variables $\{1, \ldots, (1-(1-\beta)k)^{-1/(1-\beta)}\}$.

7. Conclusions

We have proposed a new model of generation of random SAT formulas that better mimic the properties observed in *real-world* formulas. In particular, the number of occurrences of variables follows a power-law distribution, as observed in the industrial SAT instances used in competitions. This is obtained by assigning a distinct probability $P(i) \sim i^{-\beta}$ to every variable $i \in \{1, \ldots, n\}$, where β is a parameter. This model generalizes (classical) random SAT formulas by taking $\beta = 0$.

We prove the existence of a SAT–UNSAT phase transition for 2-CNF formulas. This result is obtained using a novel technique based on percolation techniques. For arbitrary k-CNF formulas, we prove that formulas with more than $\omega(n^{(1-\beta)k})$ clauses are unsatisfiable with probability $1 - o(1)$. More precisely, when $\beta > 1 - 1/k$, formulas are unsatisfiable due to a small set of clauses that only involve the most frequent variables.

Author Contributions: Investigation, C.A., M.L.B. and J.L.; Writing—original draft, C.A., M.L.B. and J.L. All authors contributed equally to this work. All authors have read and agreed to the published version of the manuscript.

References

1. Selman, B.; Kautz, H.A.; McAllester, D.A. Ten Challenges in Propositional Reasoning and Search. In Proceedings of the 15th International Joint Conference on Artificial Intelligence (IJCAI 1997), Nagoya, Japan, 23–29 August 1997; pp. 50–54.
2. Selman, B. Satisfiability Testing: Recent Developments and Challenge Problems. In Proceedings of the 15th Annual IEEE Symposium on Logic in Computer Science (LICS 2000), Santa Barbara, CA, USA, 26–29 June 2000; p. 178.
3. Kautz, H.A.; Selman, B. Ten Challenges Redux: Recent Progress in Propositional Reasoning and Search. In Proceedings of the 9th International Conference on Principles and Practice of Constraint Programming (CP 2003), Kinsale, Ireland, 29 September–3 October 2003; pp. 1–18.
4. Kautz, H.A.; Selman, B. The state of SAT. *Discret. Appl. Math.* **2007**, *155*, 1514–1524. [CrossRef]
5. Ansótegui, C.; Bonet, M.L.; Levy, J. On the Structure of Industrial SAT Instances. In Proceedings of the 15th International Conference on Principles and Practice of Constraint Programming, CP 2009, Lisbon, Portugal, 20–24 September 2009; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2009; Volume 5732, pp. 127–141. [CrossRef]

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

7. Newsham, Z.; Ganesh, V.; Fischmeister, S.; Audemard, G.; Simon, L. Impact of Community Structure on SAT Solver Performance. In Proceedings of the 17th International Conference on Theory and Applications of Satisfiability Testing (SAT 2014), Vienna, Austria, 14–17 July 2014; pp. 252–268.

8. Sonobe, T.; Kondoh, S.; Inaba, M. Community Branching for Parallel Portfolio SAT Solvers. In Proceedings of the 17th International Conference on Theory and Applications of Satisfiability Testing (SAT 2014), Vienna, Austria, 14–17 July 2014; pp. 188–196.

9. Martins, R.; Manquinho, V.M.; Lynce, I. Community-Based Partitioning for MaxSAT Solving. In Proceedings of the 16th International Conference on Theory and Applications of Satisfiability Testing (SAT 2013), Helsinki, Finland, 8–12 July 2013; pp. 182–191.

10. Katsirelos, G.; Simon, L. Eigenvector Centrality in Industrial SAT Instances. In Proceedings of the 18th International Conference on Principles and Practice of Constraint Programming (CP 2012), Quebec, QC, Canada, 8–12 October 2012; pp. 348–356.

11. Ansótegui, C.; Bonet, M.L.; Levy, J. Towards Industrial-Like Random SAT Instances. In Proceedings of the 21st International Joint Conference on Artificial Intelligence, IJCAI 2009, Newark, NJ, USA, 2–5 November 2009; pp. 387–392.

12. Friedrich, T.; Krohmer, A.; Rothenberger, R.; Sutton, A.M. Phase Transitions for Scale-Free SAT Formulas. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, AAAI 2017, San Francisco, CA, USA, 4–9 February 2017; AAAI Press: Palo Alto, CA, USA, 2017; pp. 3893–3899.

13. Friedrich, T.; Krohmer, A.; Rothenberger, R.; Sauerwald, T.; Sutton, A.M. Bounds on the Satisfiability Threshold for Power Law Distributed Random SAT. In Proceedings of the 25th Annual European Symposium on Algorithms, ESA 2017, Vienna, Austria, 4–6 September 2017; LIPIcs; Schloss Dagstuhl—Leibniz-Zentrum für Informatik: Wadern, Germany, 2017; Volume 87, pp. 37:1–37:15. [CrossRef]

14. Friedgut, E. Sharp Thresholds of Graph properties, and the k-SAT Problem. *J. Am. Math. Soc.* **1998**, *12*, 1017–1054. [CrossRef]

15. Friedrich, T.; Rothenberger, R. Sharpness of the Satisfiability Threshold for Non-uniform Random k-SAT. In Proceedings of the 21st International Conference Theory and Applications of Satisfiability Testing, SAT 2018, Oxford, UK, 9–12 July 2018; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2018; Volume 10929, pp. 273–291. [CrossRef]

16. Friedrich, T.; Rothenberger, R. The Satisfiability Threshold for Non-Uniform Random 2-SAT. In Proceedings of the 46th International Colloquium on Automata, Languages, and Programming, ICALP 2019, Patras, Greece, 8–12 July 2019; LIPIcs; Schloss Dagstuhl—Leibniz-Zentrum für Informatik, Wadern, Germany, 2019; Volume 132, pp. 61:1–61:14. [CrossRef]

17. Cooper, C.; Frieze, A.; Sorkin, G.B. Random 2-SAT with Prescribed Literal Degrees. *Algorithmica* **2007**, *48*, 249–265. [CrossRef]

18. Omelchenko, O.; Bulatov, A.A. Satisfiability Threshold for Power Law Random 2-SAT in Configuration Model. In Proceedings of the 22nd International Conference on Theory and Applications of Satisfiability Testing, SAT 2019, Lisbon, Portugal, 9–12 July 2019; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2019; Volume 11628, pp. 53–70. [CrossRef]

19. Omelchenko, O.; Bulatov, A. Satisfiability and Algorithms for Non-uniform Random k-SAT. In Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Vancouver, BC, Canada, 2–9 February 2021; AAAI Press: Palo Alto, CA, USA, 2021; pp. 3886–3894.

20. Omelchenko, O.; Bulatov, A.A. Satisfiability threshold for power law random 2-SAT in configuration model. *Theor. Comput. Sci.* **2021**, *888*, 70–94. [CrossRef]

21. Giráldez-Cru, J.; Levy, J. Locality in Random SAT Instances. In Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017. ijcai.org, Melbourne, Australia, 19–25 August 2017; pp. 638–644. [CrossRef]

22. Giráldez-Cru, J.; Levy, J. Popularity-similarity random SAT formulas. *Artif. Intell.* **2021**, *299*, 103537. [CrossRef]

23. Achlioptas, D.; Coja-Oghlan, A.; Hahn-Klimroth, M.; Lee, J.; Müller, N.; Penschuck, M.; Zhou, G. The number of satisfying assignments of random 2-SAT formulas. *Random Struct. Algorithms* **2021**, *58*, 609–647. [CrossRef]

24. Bläsius, T.; Friedrich, T.; Göbel, A.; Levy, J.; Rothenberger, R. The Impact of Heterogeneity and Geometry on the Proof Complexity of Random Satisfiability. In Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms, SODA 2021, SIAM 2021, Virtual Conference, 10–13 January 2021; pp. 42–53. [CrossRef]

25. Barabási, A.L.; Albert, R. Emergence of scaling in random networks. *Science* **1999**, *286*, 509–512. [CrossRef]

26. Dorogovtsev, S.N.; Mendes, J.F.F. *Evolution of Networks: From Biological Nets to the Internet and WWW (Physics)*; Oxford University Press, Inc.: New York, NY, USA, 2003.

27. Dorogovtsev, S.N.; Mendes, J.F.F. Evolution of networks with aging of sites. *Phys. Rev. E* **2000**, *62*, 1842–1845. [CrossRef]

28. Bender, E.A.; Canfield, E. The asymptotic number of labeled graphs with given degree sequences. *J. Comb. Theory Ser. A* **1978**, *24*, 296–307. [CrossRef]

29. Bollobás, B. *Random Graphs*; Cambridge Studies in Advanced Mathematics, Cambridge University Press: Cambridge, UK, 2001.

30. Aiello, W.; Chung, F.; Lu, L. A Random Graph Model for Massive Graphs. In Proceedings of the Thirty-Second Annual ACM Symposium on Theory of Computing, Portland, OR, USA, 21–23 May 2000; pp. 171–180.

31. Chung, F.; Lu, L. Connected Components in Random Graphs with Given Expected Degree Sequences. *Ann. Comb.* **2002**, *6*, 125–145. [CrossRef]

32. Goh, K.I.; Kahng, B.; Kim, D. Universal behavior of load distribution in scale-free networks. *Phys. Rev. Lett.* **2001**, *87*, 278701. [CrossRef] [PubMed]

33. Mase, S. Approximations to the birthday problem with unequal occurrence probabilities and their application to the surname problem in Japan. *Ann. Inst. Stat. Math.* **1992**, *44*, 479–499.

34. Boufkhad, Y.; Dubois, O.; Interian, Y.; Selman, B. Regular Random *k*-SAT: Properties of Balanced Formulas. *J. Autom. Reason.* **2005**, *35*, 181–200. [CrossRef]

35. Chvátal, V.; Reed, B.A. Mick Gets Some (the Odds Are on His Side). In Proceedings of the 33rd Annual Symposium on Foundations of Computer Science, FOCS 1992, Pittsburgh, PA, USA, 244–27 October 1992; pp. 620–627.

36. Erdös, P.; Rényi, A. On Random Graphs I. *Publ. Math.* **1959**, *6*, 290–297.

37. Gilbert, E.N. Random Graphs. *Ann. Math. Stat.* **1959**, *30*, 1141–1144. [CrossRef]

38. Erdös, P.; Rényi, A. On the evolution of random graphs. *Publ. Math. Inst. Hungary. Acad. Sci.* **1960**, *5*, 17–61.

39. Mitchell, D.G.; Selman, B.; Levesque, H.J. Hard and Easy Distributions of SAT Problems. In Proceedings of the 10th National Conference on Artificial Intelligence (AAAI 1992), San Jose, CA, USA, 12–16 July 1992; pp. 459–465.

40. Gent, I.P.; Walsh, T. The SAT Phase Transition. In Proceedings of the 11th European Conference on Artificial Intelligenc (ECAI 1994), Amsterdam, The Netherlands, 8–12 August 1994; pp. 105–109.

41. Achlioptas, D.; Chtcherba, A.D.; Istrate, G.; Moore, C. The phase transition in 1-in-k SAT and NAE 3-SAT. In Proceedings of the 20th Annual Symposium on Discrete Algorithms, SODA 2001, Washington, DC, USA, 7–9 January 2001; pp. 721–722.

42. Sinclair, A.; Vilenchik, D. Delaying satisfiability for random 2SAT. *Random Struct. Algorithms* **2013**, *43*, 251–263. [CrossRef]

43. Bollobás, B.; Borgs, C.; Chayes, J.T.; Kim, J.H.; Wilson, D.B. The Scaling Window of the 2-SAT Transition. *Random Struct. Algorithms* **2001**, *18*, 201–256. [CrossRef]

44. Bollobás, B. The evolution of random graphs. *Trans. Am. Math. Soc.* **1984**, *286*, 257–274. [CrossRef]

45. Monasson, R.; Zecchina, R.; Kirkpatrick, S.; Selman, B.; Troyansky, L. 2+p-SAT: Relation of typical-case complexity to the nature of the phase transition. *Random Struct. Algorithms* **1999**, *15*, 414–435. [CrossRef]

46. Aspvall, B.; Plass, M.F.; Tarjan, R.E. A Linear-Time Algorithm for Testing the Truth of Certain Quantified Boolean Formulas. *Inf. Process. Lett.* **1979**, *8*, 121–123. [CrossRef]

47. Molloy, M.; Reed, B. A critical point for random graphs with a given degree sequence. *Random Struct. Algorithms* **1995**, *6*, 161–180. [CrossRef]

48. Newman, M.E.J. *Networks: An Introduction*; Oxford University Press: Oxford; NY, USA, 2010.

49. Cohen, R.; Erez, K.; ben Avraham, D.; Havlin, S. Resilience of the Internet to Random Breakdowns. *Phys. Rev. Lett.* **2000**, *85*, 4626–4628. [CrossRef] [PubMed]

50. Cohen, R.; Havlin, S.; ben Avraham, D. Structural properties of scale-free networks. In *Handbook of Graphs and Networks*; John Wiley & Sons, Ltd.: Hoboken, NJ, USA, 2002; Chapter 4, pp. 85–110. [CrossRef]

51. Boufkhad, Y.; Dubois, O.; Interian, Y.; Selman, B. Regular Random k-SAT: Properties of Balanced Formulas. In Proceedings of the 8th International Conference on Theory and Applications of Satisfiability Testing (SAT 2005), St Andrews, UK, 19–23 June 2005; pp. 181–200.