Pseudo-Industrial Random SAT Generators

Jesús Giráldez-Cru[1] and Jordi Levy[1]

[1]Artificial Intelligence Research Institute, Spanish National Research Council (IIIA-CSIC),
{jgiraldez,levy}@iiia.csic.es

Introduction

Recent advances in SAT are focused on efficiently solving *real-world* or *industrial* problems. However, the reduced number of industrial SAT instances and the high cost of solving them condition the development and debugging processes of new techniques. This problem can be solved by defining new models of random SAT instance that capture *realistically* the main features of the real-world SAT instances. In this work, we review some of these models, and we define a new model in which we are working on.

The *Classical Random model* [8] was popularized to study the SAT/UNSAT phase transition phenomena, and the easy-hard-easy associated pattern, both dependent on the clause/variable ratio. In this model, k-CNF formulas consist of m independent clauses among the $2^k\binom{n}{k}$ clauses with k literals on n variables that are neither simplifiable nor tautologies. This model does not capture the main features of real-world problems, and SAT solvers perform very differently on random and industrial SAT instances.

One important feature of industrial SAT instance is the *scale-free structure* [2]. This means that the number of variables occurrences follows, in general, a power-law distribution $p(k) \propto k^{-\alpha}$, and these distributions are scale-free. This implies that there exists a big variability in the number of occurrences of variables. This kind of structure is very characteristic in real-world networks. Preferential Attachment [6] was proposed as a model to explain this behavior in growing networks. *Scale-free random SAT formulas* [3] were proposed as an alternative model that also reproduce this feature. This model is parametric in the exponent α, and generates formulas as set of independent clauses, where the clause with variables $\{i_1, \ldots, i_k\} \subseteq \{1, \ldots, n\}$ has probability $P(i_1 \vee \ldots \vee i_k) \sim \prod_{j=1}^{k} (i_j)^{\frac{1}{\alpha-1}}$.

Another important feature shared by the majority of industrial benchmarks is the *community structure* (or high *modularity*) [4, 5], meaning that they are characterized by a partition of communities of highly connected variables, i.e., they usually appear in clauses with variables of the same community. The *Community Attachment model* [7] was proposed to generate instances with this structure. It is parametric on a modularity Q and number of communities c. Formulas are also sets of independent clauses, where, with probability $Q + 1/c$, all their literals belong to the same community; otherwise, all of them belong to distinct communities. This generates formulas with modularity at least Q.

The Popularity and Similarity Model

SAT solvers *specialized* in industrial instances perform better in random instances generated with the scale-free or the community attachment models, than solvers *specialized* in classical random problems, and vice-versa. This means that industrial solvers exploit both the scale-free and the modular structure of SAT instances. However, both models have some drawbacks. Scale-free instances are not modular, and instances generated with the community attachment model are not scale-free.

Something similar has been observed in real-world networks. In [9], it is shown that properties of these growing networks are better explained considering two dimensions of attractiveness: *popularity* and *similarity*. Popularity means that new nodes prefer to connect to popular nodes, i.e. to nodes with a high degree (this is exactly preferential attachment), and similarity that new nodes also tend to connect to nodes similar to themselves, even if they are not popular. They propose a simple model where every node $i = 1, \dots, n$ has a creation time $t_i = i$, and a uniformly-random coordinate $\theta_i \in [0, 2\pi]$. Then, every new node i is connected to the m previous nodes $j < i$ that minimize $t_i \cdot ||\theta_i - \theta_j||$. The inverse of time creation t_i represents popularity, and the inverse of the angle $||\theta_i - \theta_j||$ represents similarity.

The direct adaptation of this model for SAT instances would be assigning coordinates to both variables and clauses, and connect every clause j with the k variables i's minimizing $t_i \cdot ||\theta_i - \theta_j||$. However, in this model the number of possible clauses grows linearly on n, instead of as $\binom{n}{k}$ like in other models.

In order to avoid the previous problem, we *randomize* the model. Every clause j with coordinate θ_j has provability $P_j(i) \sim t_i^{-\beta} \cdot ||\theta_i - \theta_j||^{-\delta}$ to contain variable i. Exponents β and δ regulate the weight of popularity and similarity in the model. Notice that with $\delta = 0$, and taking $\beta = \frac{1}{\alpha - 1}$ we obtain (pure) scale-free SAT formulas, as a particular case.

In this model, the scale-free feature is produced by the preferential attachment mechanism, and the community structure is a consequence of attaching nodes by similarity, which produces communities of similar nodes. It may also produce self-similar structure, another structure feature very common in real-world benchmarks [1].

As future work, we plan an exhaustive experimental analysis of this model. In particular, we want to focus on the relation between popularity and similarity in the case of real industrial SAT instances. Moreover, some questions remain open. For instance, what is the role of popular nodes within each community? Or, what is the relation between popularity/similarity and the communities organization? The answers to these questions will allow to correctly adequate this model to the particular case of *realistic* pseudo-industrial random SAT instances generation in an accurate manner.

References

[1] C. Ansótegui, M. L. Bonet, J. Giráldez-Cru, and J. Levy. The fractal dimension of SAT formulas. In *Proc. of IJCAR'14*, pages 107–121, 2014.

[2] C. Ansótegui, M. L. Bonet, and J. Levy. On the structure of industrial SAT instances. In *Proc. of CP'09*, pages 127–141, 2009.

[3] C. Ansótegui, M. L. Bonet, and J. Levy. Towards industrial-like random SAT instances. In *Proc. of IJCAI'09*, pages 387–392, 2009.

[4] C. Ansótegui, J. Giráldez-Cru, and J. Levy. The community structure of SAT formulas. In *Proc. of SAT'12*, pages 410–423, 2012.

[5] C. Ansótegui, J. Giráldez-Cru, J. Levy, and L. Simon. Using community structure to detect relevant learnt clauses. In *Proc. of SAT'15*, pages 238–254, 2015.

[6] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[7] J. Giráldez-Cru and J. Levy. A modularity-based random SAT instances generator. In *Proc. of IJCAI'15*, pages 1952–1958, 2015.

[8] D. Mitchell, B. Selman, and H. Levesque. Hard and easy distributions of SAT problems. In *Proc. of AAAI'92*, pages 459–465, 1992.

[9] F. Papadopoulos, M. Kitsak, M. Á. Serrano, M. Boguñá, and D. Krioukov. Popularity versus similarity in growing networks. *Nature*, 489:537–540, 2012.