Locality in Random SAT Instances

Jesús Giráldez-Cru
KTH Royal Institute of Technology Stockholm, Sweden

Jordi Levy
IIIA, CSIC, Barcelona, Spain

IJCAI'17, Melbourne, Australia

SAT in Practice

- SAT is NP-complete
- Random SAT formulas require exponential tree-like refutations
- SAT solvers solve industrial instances with millions of clauses in seconds

SAT in Practice

- SAT is NP-complete
- Random SAT formulas require exponential tree-like refutations
- SAT solvers solve industrial instances with millions of clauses in seconds

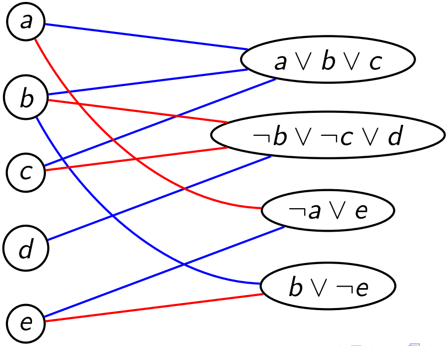
Objectives

- Study the structural properties of real-world SAT instances
- Propose new models of random formulas
- Exploit this knowledge to improve SAT solvers specialized in those kind of formulas

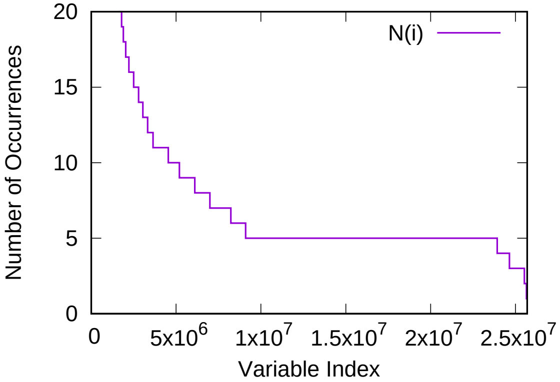
SAT in Practice

Objectives

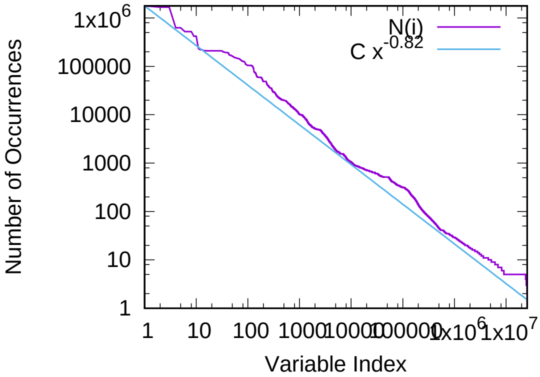
- Study the structural properties of real-world SAT instances
- Propose new models of random formulas
- Exploit this knowledge to improve SAT solvers specialized in those kind of formulas



Consider all variables of the SAT'08 Competition and sort them
$N(i)$ = number of occurrences of i-th most frequent variable
Most have 5 occurrences, although the average is 13.6!!!
A few have millions of occurrences!!!



Scale-free Formulas [Ansotegui, Bonet & Levy, IJCAI'09]



Expected number of occurrences of i-th most frequent variable

$$N(i) \sim i^{-0.82}$$

Seen as a graph, industrial SAT formulas are scale-free

Drawbacks of the (simple) Scale-free Model

- Since formulas are scale-free, the best variable branching heuristics is assigning most frequent variables.
- VSIDS heuristics: try to focus in some area of the formula.
- Scale-free formulas are too easy on practice (popular variables are too inter-connected)
- Real-world networks: scale-free structure (popularity) alone does not explain high clustering of networks

Drawbacks of the (simple) Scale-free Model

- Since formulas are scale-free, the best variable branching heuristics is assigning most frequent variables.
- VSIDS heuristics: try to focus in some area of the formula.
- Scale-free formulas are too easy on practice (popular variables are too inter-connected)
- Real-world networks: scale-free structure (popularity) alone does not explain high clustering of networks

In this paper

- There is a notion of locality in formulas
- This notion coincides with the notion of similarity used in complex networks

Popularity and Similarity Model

For every variable $i \in 1 \dots n$ and clause $j \in 1 \dots m$ assign a random angle (position) $\theta_i \in [0, 2\pi]$ and $\theta'_j \in [0, 2\pi]$.

Popularity and Similarity Model

For every variable $i \in 1 \dots n$ and clause $j \in 1 \dots m$ assign a random angle (position) $\theta_i \in [0, 2\pi]$ and $\theta'_j \in [0, 2\pi]$.
Define the energy of edge $i \leftrightarrow j$ (occurrence of variable i in clause j) as

$$e_{ij} = \underbrace{\beta \log i}_{\text{Popularity of variable } i} + \underbrace{\beta' \log j}_{\text{Popularity of clause } j} + \underbrace{\log \theta_{ij}}_{\text{Similarity between } i \text{ and } j}$$

(more energetic edges are less probable)
We do not allow multiple edges between the same pair of nodes (hence tautologies $a \vee \neg a \vee b$ or simplificable clauses $a \vee a \vee b$ are disallowed)

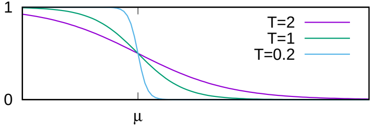
Popularity and Similarity Model

For every variable $i \in 1 \dots n$ and clause $j \in 1 \dots m$ assign a random angle (position) $\theta_i \in [0, 2\pi]$ and $\theta'_j \in [0, 2\pi]$.
Define the energy of edge $i \leftrightarrow j$ (occurrence of variable i in clause j) as

$$e_{ij} = \beta \log i + \beta' \log j + \log \theta_{ij}$$

Use the Fermi-Dirac probability distribution for fermions

$$E[n_{ij}] = \frac{1}{1+e^{\frac{e_{ij}-\mu}{k\,T}}}$$



where μ is the total chemical potential, k the Boltzmann's constant and T the temperature.

Popularity and Similarity Model

For every variable $i \in 1 \ldots n$ and clause $j \in 1 \ldots m$ assign a random angle (position) $\theta_i \in [0, 2\pi]$ and $\theta'_j \in [0, 2\pi]$.
Define the energy of edge $i \leftrightarrow j$ (occurrence of variable i in clause j) as

$$e_{ij} = \beta \log i + \beta' \log j + \log \theta_{ij}$$

Use the Fermi-Dirac probability distribution for fermions

$$E[n_{ij}] = \frac{1}{1+e^{\frac{e_{ij}-\mu}{k\,T}}}$$

where μ is the total chemical potential, k the Boltzmann's constant and T the temperature.
Redefining $T = k\,T$ and $\mu = \log R$, this results into

$$P(i \leftrightarrow j) = \frac{1}{1 + \left(\frac{i^{\beta} \cdot j^{\beta'} \cdot \theta_{ij}}{R}\right)^{1/T}}$$

Popularity and Similarity Model

$$P(i \leftrightarrow j) = \frac{1}{1 + \left(\frac{i^{\beta} \cdot j^{\beta'} \cdot \theta_{ij}}{R} \right)^{1/T}}$$

For $T = 0$ we have

$$P(i \leftrightarrow j) = \begin{cases} 1 & \text{if } e_{ij} < \mu \quad \text{i.e. } i^{\beta} \cdot j^{\beta'} \cdot \theta_{ij} < R \\ 0 & \text{if } e_{ij} > \mu \quad \text{i.e. } i^{\beta} \cdot j^{\beta'} \cdot \theta_{ij} > R \end{cases}$$

Fixed θ's, the model is deterministic

Popularity and Similarity Model

$$P(i \leftrightarrow j) = \frac{1}{1 + \left(\frac{i^{\beta} \cdot j^{\beta'} \cdot \theta_{ij}}{R}\right)^{1/T}}$$

In general, if k is the desired average size of clauses we compute the R satisfying:

$$\sum_{i=1}^{n} \sum_{j=1}^{m} P(i \leftrightarrow j) = k \cdot m$$

The chemical potential R depends on temperature T

Popularity and Similarity Model

$$P(i \leftrightarrow j) = \frac{1}{1 + \left(\frac{i^{\beta} \cdot j^{\beta'} \cdot \theta_{ij}}{R} \right)^{1/T}}$$

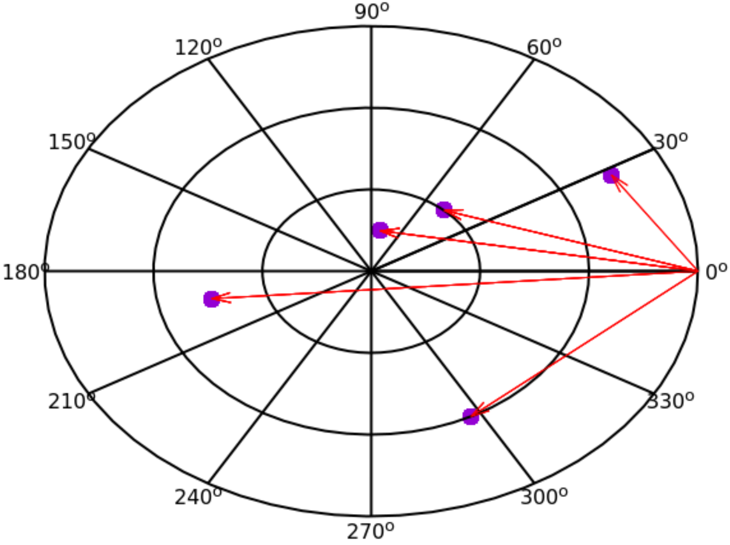
In general, if k is the desired average size of clauses we compute the R satisfying:

$$\sum_{i=1}^{n} \sum_{j=1}^{m} P(i \leftrightarrow j) = k \cdot m$$

Lemma

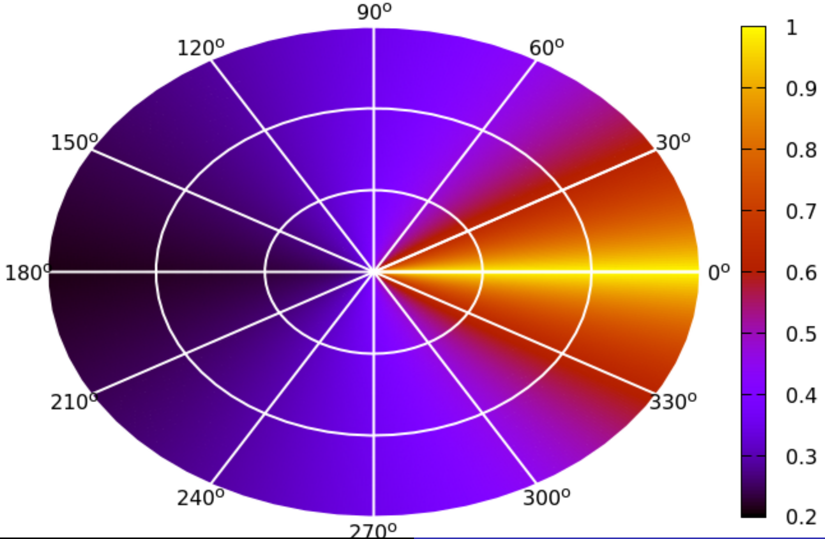
If $P(i \leftrightarrow j) = f(i^{\beta}\, j^{\beta'}\, \theta_{ij})$, and f decreases fast enough, then the resulting SAT instance is scale-free with variable occurrences $P(k) \sim k^{-\delta}$, where $\delta = 1 + 1/\beta$ and clauses sizes $P(s) \sim s^{-\delta'}$ where $\delta' = 1 + 1/\beta'$.

Hyperbolic Geometry



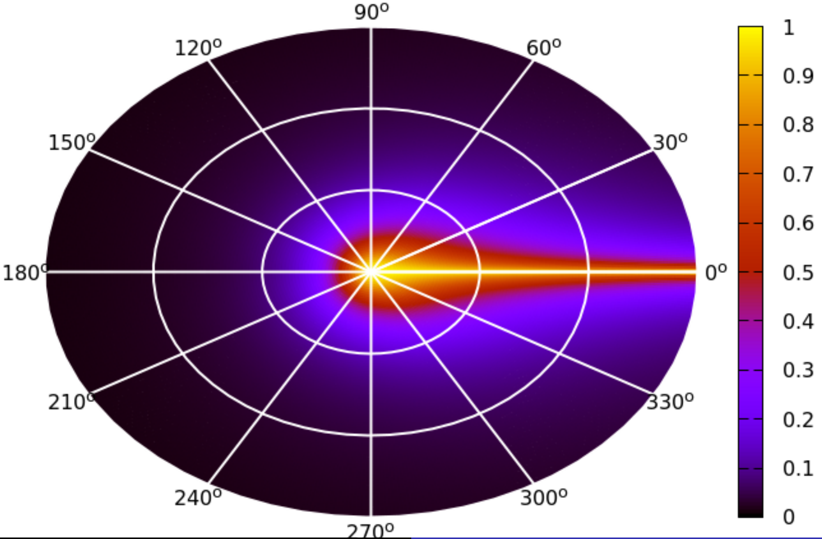
Probability of Connection

$\beta = 0.1$ prefer similar nodes $(T = 1)$



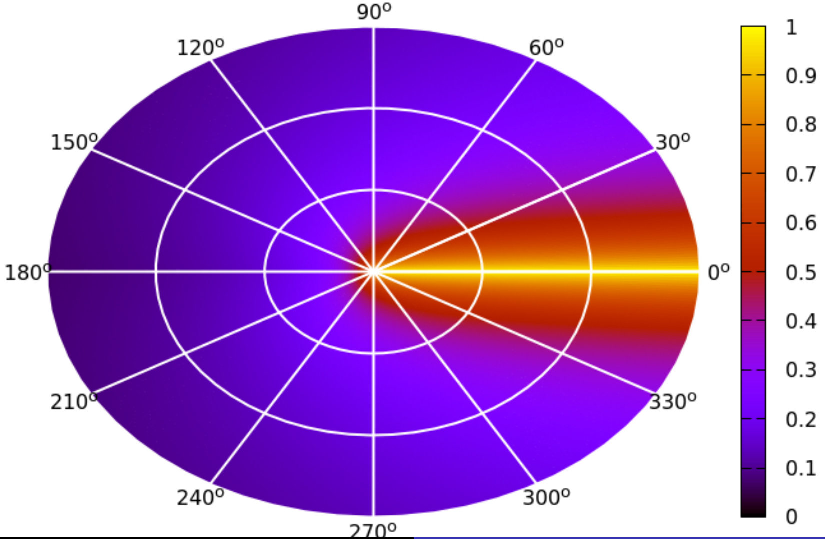
Probability of Connection

$\beta = 2$ prefer popular nodes $(T = 1)$



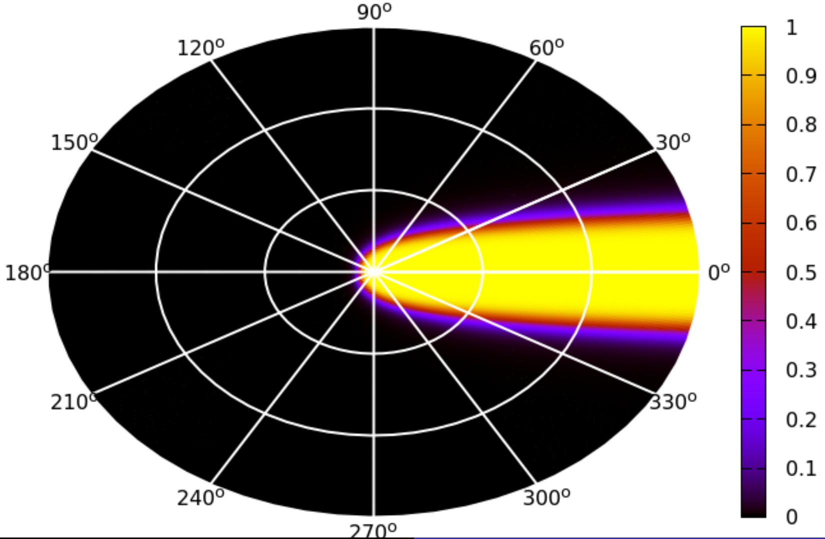
Probability of Connection

$\beta = 0.8$ balance similarity-popularity $(T = 1)$



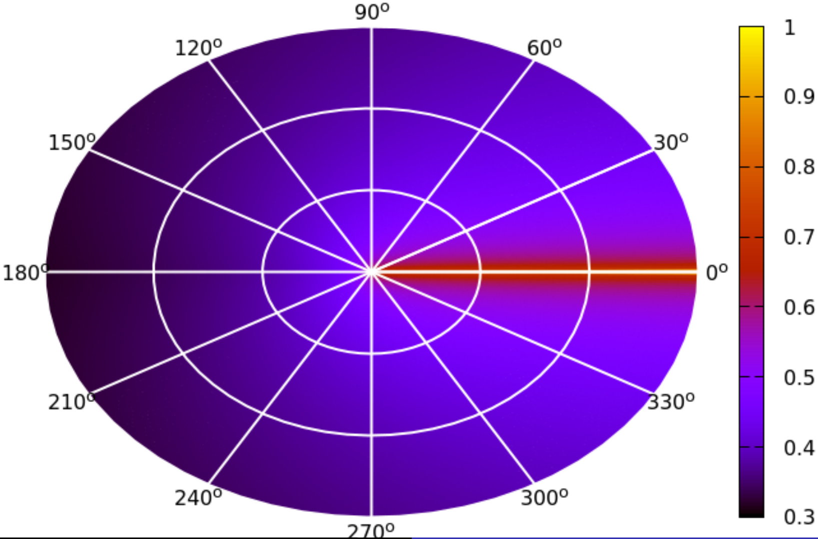
Probability of Connection

$T = 0.1$ connect only to closest nodes ($\beta = 0.8$)



Probability of Connection

$T = 3$ connect to random nodes ($\beta = 0.8$)



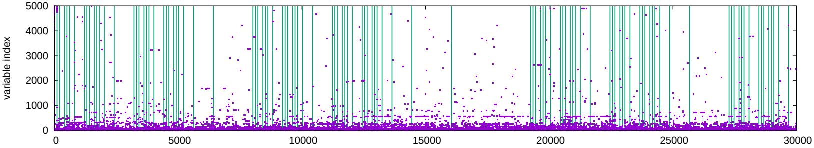
Some Problems (solved in the paper)

- How to compute R:
  It can be analytically approximated for $T \approx 0$.
  For big temperature we use an algorithm based on Newton-Raphson method.
- We present a simplified model where the probability of connection is

$$P(i \leftrightarrow j) = \min\left\{1\,,\ \frac{R}{(i^{\beta} j^{\beta'}\,\theta_{ij})^{1/T}}\right\}$$

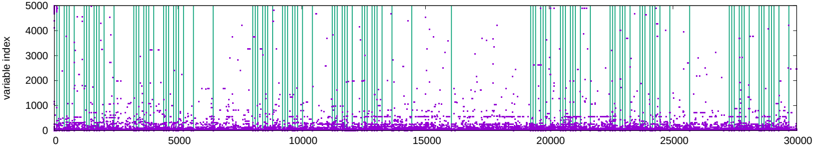
- Proliferation of small clauses make most formulas trivially unsatisfiable. A minimum size of clauses is proposed as a solution

$\beta = 0.8$ $T = 1.5$ (vars. ordered by popularity)

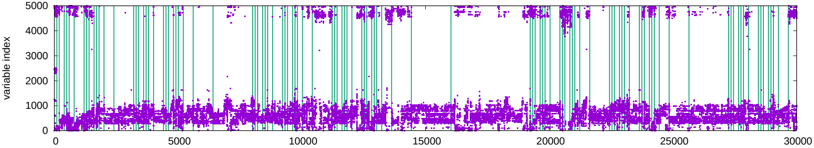


Decided Variable

$\beta = 0.8$ $T = 1.5$ (vars. ordered by popularity)

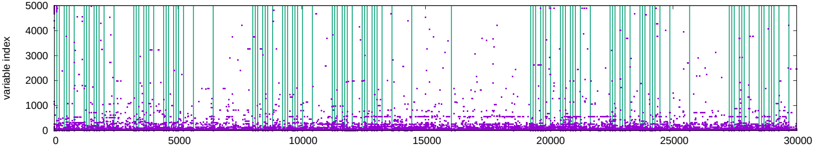


$\beta = 0.1$ $T = 0.75$ (vars. ordered by similarity)

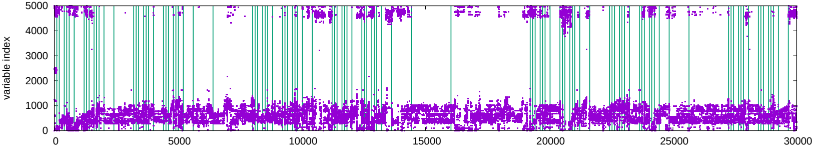


Decided Variable

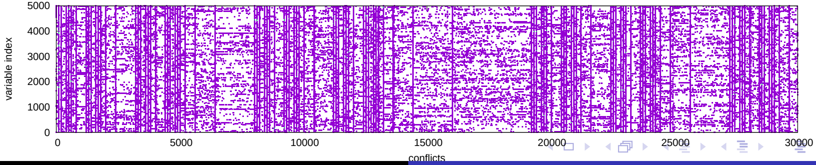
$\beta = 0.8$ $T = 1.5$ (vars. ordered by popularity)



$\beta = 0.1$ $T = 0.75$ (vars. ordered by similarity)



$T = 100$ ($\beta = 0.1$ and vars. ordered by similarity)



Conclusions and Further Work

- An equilibrium between the forces of popularity and similarity defines the structure of industrial SAT instances
- Modern SAT solvers exploit both structures
- Explicit computation of variable coordinates may lead to better branching heuristics
- Analysis of the temperature of formulas may characterize their difficulty