Sparkle SAT Challenge 2018 (Extended Version)

Chuan Luo and Holger H. Hoos

LIACS
Universiteit Leiden
The Netherlands

FLoC 2018, Oxford, UK
12 July 2018

The state of the art in solving X ...

- ... is not defined by a single solver / solver configuration
- ... requires use of / interplay between multiple heuristic mechanisms
- ... has been substantially advanced by machine learning

Competitions ...

- ... have helped advance the state of the art in many fields (SAT, AI planning, machine learning, ...)
- ... are mostly focussed on single solvers, broad-spectrum performance
- ... often don't reflect state of the art
- ... typically don't provide effective incentive to improve state of the art

A new kind of competition:

- solvers submitted to competition platform
- state-of-the-art per-instande selector built based on all solvers
- solver contributions to overall performance assessed based on (relative) marginal contribution

  (Xu, Hutter, HH, Leyton-Brown 2012; Luo & Hoos – this event)
- full credit for contributions to selector performance goes to component solver authors

⇝ Sparkle (Luo & Hoos – this event)

1
2
3

XXXI Olimpíada Rio 2016

XXXI Olimpíada Rio 2016

Sparkle SAT Challenge 2018

- part of FLoC Olympic Games, coordinated with 2018 SAT Competition
- launched March 2018, leader board phase 5–15 April, final results now!
- 19 open-source solvers submitted, 4 hors-concours solvers included
- website: http://ada.liacs.nl/events/sparkle-sat-18

Training and testing sets

- training set: 1356 instances from 25 families,
  all solved instances (satisfiable + unsatisfiable) from main, application, crafted/hard-combinatorial tracks of 2014–2017 SAT Competitions + 2015 SAT Race
- testing set: 400 instances from 23 families,
  identical to testing set of main track of 2018 SAT Competition

Constructing the per-instance selector

- training set: 1356 instances from 25 families
- split training set into *core training set* and *validating set*
  - randomly select 15 instance families → core training set
  - remaining 10 families → validating set
- core training set: 893 instances from 15 families
- validating set: 463 instances from 10 families

- run AutoFolio (Lindauer *et al.* 2015) 100 times
  to obtain 100 per-instance selectors
  - train on core training set
  - choose selector with smallest PAR2 score on validating set

AutoFolio

- automatically configure flexible selector framework to find state-of-the-art, customised selectors

  (Lindauer, Hoos, Hutter, Schaub 2015)

- based on well-known, flexible per-instance algorithm selection framework: claspfolio 2 (Hoos & Lindauer & Schaub 2014)

- leverages state-of-the-art, general-purpose algorithm configurator: SMAC (Hutter, Hoos, Leyton-Brown 2011)

⇝ cutting-edge, robust algorithm selector construction in Sparkle

Assessing solver contributions

Given: set of solvers S; per-instance selector P based on S; instance set I

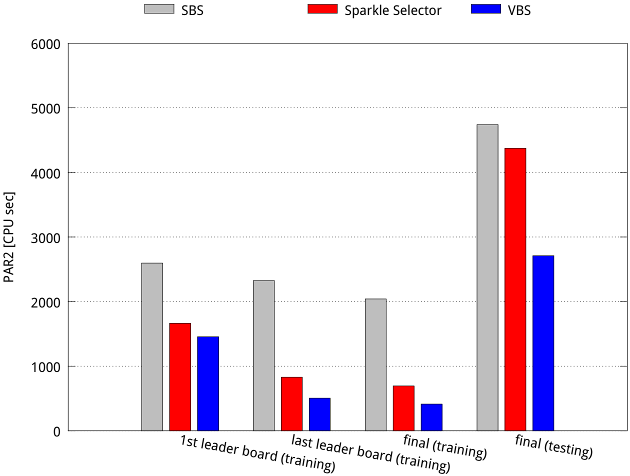
absolute marginal contribution (amc) of solver s on I:

$$amc(s, I) = \begin{cases} log_{10} \frac{PAR2(P \backslash \{s\}, I)}{PAR2(P, I)} & PAR2(P \backslash \{s\}, I) > PAR2(P, I) \\ 0 & \text{else} \end{cases}$$

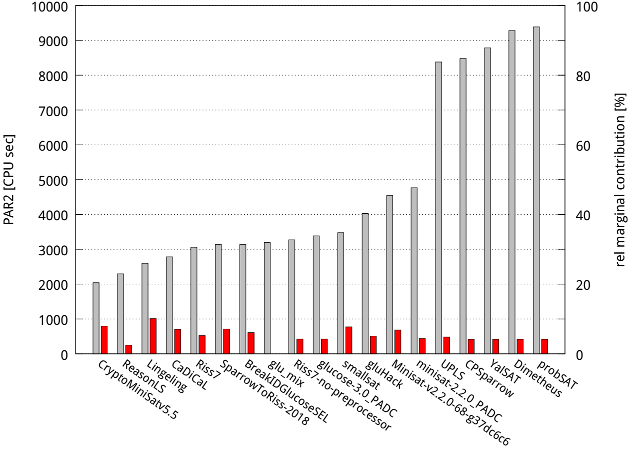
relative marginal contribution (rmc) of solver s of I:

$$rmc(s, I) = \frac{amc(s)}{\sum_{s' \in S} amc(s')}$$

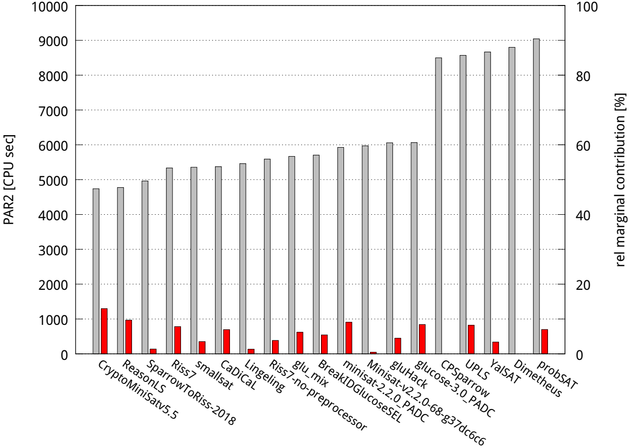
Improvement over time without hors-concours solvers



Stand-alone and relative marginal contribution on training set, without hors-concours solvers



Stand-alone and relative marginal contribution on testing set, without hors-concours solvers



Final results without hors-concours solvers, on testing set

PAR2 for SBS, VBS and Sparkle Selector

- SBS: 4740.02
- VBS: 2710.91
- Sparkle Selector: 4375.42

Official results: Ranking according to marginal contribution on testing set, without hors-concours solvers

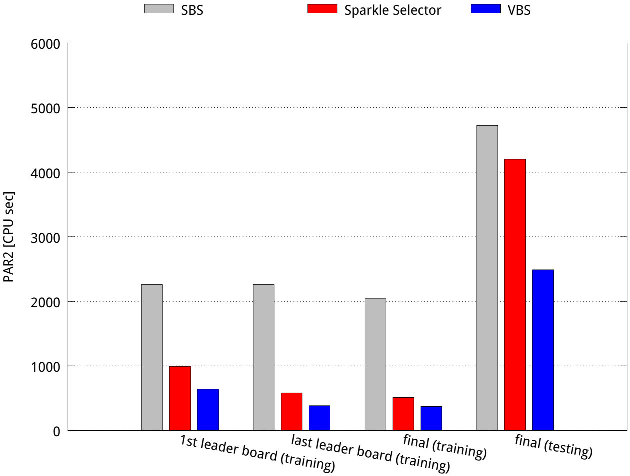
rank	solver (stand-alone rank)	rmc	amc
1	CryptoMiniSatv5.5 (1)	12.97%	0.0430
2	ReasonLS (2)	9.68%	0.0321
3	minisat-2.2.0_PADC (11)	9.07%	0.0301
4	glucose-3.0_PADC (14)	8.41%	0.0279
5	UPLS (16)	8.18%	0.0271
6	Riss7 (4)	7.81%	0.0259
7	probSAT (19)	6.99%	0.0232
8	CaDiCaL (6)	6.93%	0.0230
9	glu_mix (9)	6.20%	0.0205
10	BreakIDGlucoseSEL (10)	5.42%	0.0180

Observations

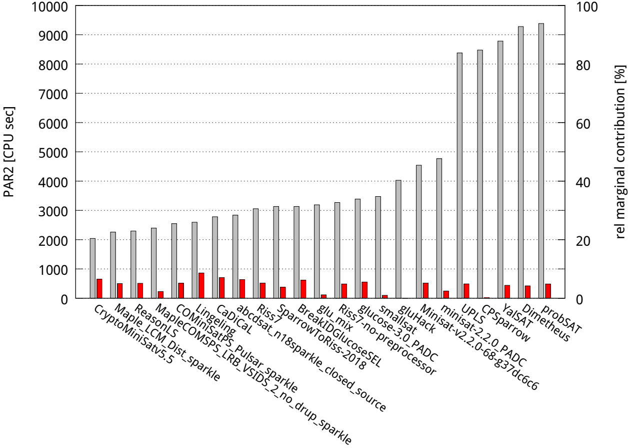
- Compared to SBS, the actual Sparkle selector shows considerably improved performance on the testing set, although the testing instances are quite different from the training instances (and harder for most solvers, Sparkle selector and VBS).
- Solver ranking according to marginal contribution is quite different from the traditional stand-alone ranking.
- Several solvers that don't have high stand-alone performance make considerable contribution to the Sparkle selector.

Additional insights:
Results with *hors-concours* solvers

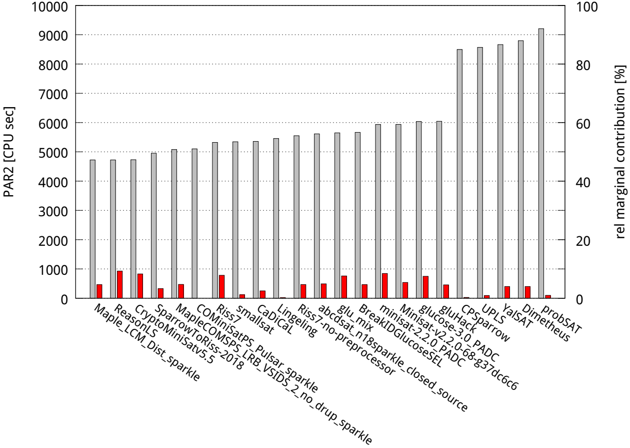
Improvement over time including hors-concours solvers



Stand-alone and relative marginal contribution on training set, with hors-concours solvers



Stand-alone and relative marginal contribution on testing set, with hors-concours solvers



Final results including hors-concours solvers, on testing set

PAR2 for SBS, VBS and Sparkle Selector

- SBS: 4724.03
- VBS: 2489.42
- Sparkle Selector: 4201.26

Ranking according to marginal contribution on testing set, with hors-concours solvers

Marginal contribution ranking (Top 10 solvers)

rank	solver (rank without hors-concours solvers, stand-alone rank)	rmc	amc
1	ReasonLS (2, 2)	9.30%	0.0533
2	minisat-2.2.0_PADC (3, 15)	8.46%	0.0485
3	CryptoMiniSatv5.5 (1, 3)	8.31%	0.0476
4	Riss7 (6, 7)	7.83%	0.0448
5	glu_mix (9, 13)	7.62%	0.0436
6	glucose-3.0_PADC (4, 17)	7.49%	0.0429
7	Minisat-v2.2.0-68-g37dc6c6 (17, 16)	5.40%	0.0309
8	*abcdsat_n18sparkle_closed_source (hors concours, 12)*	*4.91%*	*0.0281*
9	*MapleCOMSPS_LRB_VSIDS_2_no_drup_sparkle (hors concours, 5)*	*4.73%*	*0.0271*
10	Riss7-no-preprocessor (12, 11)	4.67%	0.0268

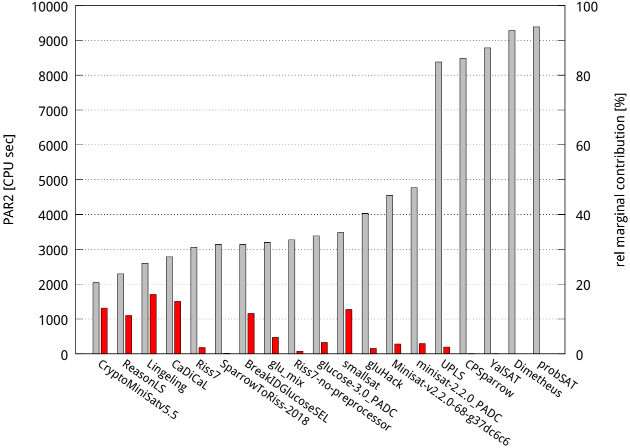
Observations

- Results are qualitatively similar to those without hors-concours solvers.

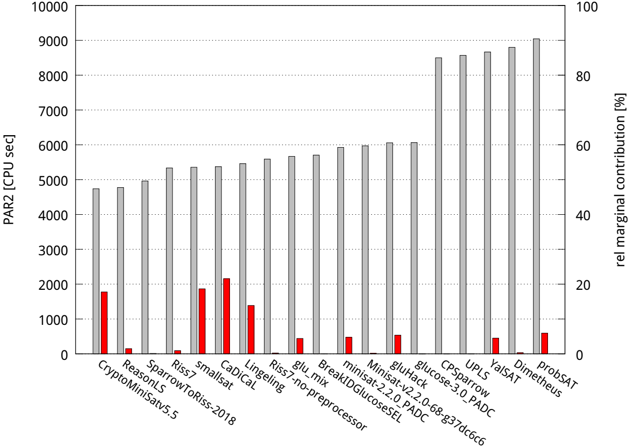
- Hors-concours (closed-source) solvers do contribute, but don't make a large difference.

Additional insights:
Results with respect to VBS

Stand-alone and relative marginal contribution to VBS on training set, without hors-concours solvers



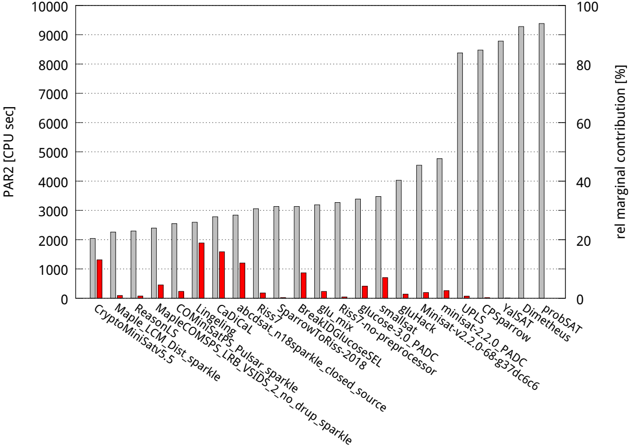
Stand-alone and relative marginal contribution to VBS on testing set, without hors-concours solvers



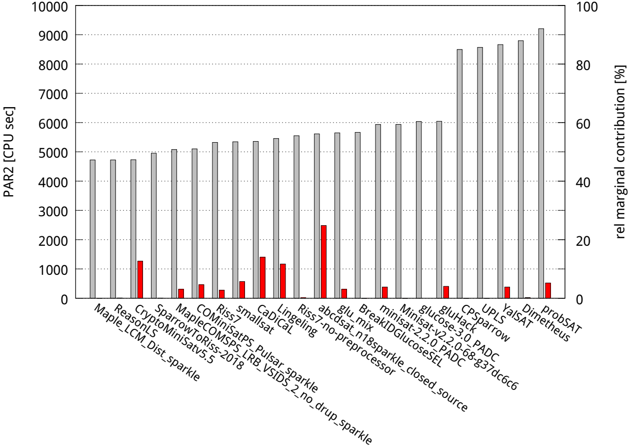
Ranking according to marginal contribution to VBS on testing set, without hors-concours solvers

rank	solver (stand-alone rank)	rmc	amc
1	CaDiCaL (6)	21.54%	0.0190
2	smallsat (5)	18.66%	0.0165
3	CryptoMiniSatv5.5 (1)	17.77%	0.0157
4	Lingeling (7)	13.88%	0.0122
5	probSAT (19)	5.94%	0.0052
6	gluHack (13)	5.36%	0.0047
7	minisat-2.2.0_PADC (11)	4.73%	0.0042
8	YalSAT (17)	4.51%	0.0040
9	glu_mix (9)	4.40%	0.0039
10	ReasonLS (2)	1.44%	0.0013

Stand-alone and relative marginal contribution to VBS on training set, with hors-concours solvers



Stand-alone and relative marginal contribution to VBS on testing set, with hors-concours solvers



Ranking according to marginal contribution to VBS on testing set, with hors-concours solvers

Marginal contribution ranking (Top 10 solvers)

rank	solver (rank with respect to VBS without hors-concours solvers, stand-alone rank)	rmc	amc
1	*abcdsat_n18sparkle_closed_source (hors concours, 12)*	*24.84%*	*0.0280*
2	CaDiCaL (1, 9)	14.03%	0.0158
3	CryptoMiniSatv5.5 (3, 3)	12.68%	0.0143
4	Lingeling (4, 10)	11.70%	0.0132
5	smallsat (2, 8)	5.70%	0.0064
6	probSAT (5, 23)	5.21%	0.0059
7	*COMiniSatPS_Pulsar_sparkle (hors concours, 6)*	*4.62%*	*0.0052*
8	gluHack (6, 18)	4.06%	0.0046
9	YalSAT (8, 21)	3.84%	0.0043
10	minisat-2.2.0_PADC (7, 15)	3.82%	0.0043

Observations

- Solver ranking according to marginal contribution is quite different from traditional stand-alone ranking.

- Several solvers that don't have high stand-alone performance make considerable contribution to the VBS.

  - Results without hors-concours solvers: (*CaDiCaL*), which makes largest marginal contribution, ranks 9th according to stand-alone performance.

  - Results with hors-concours solvers: (*abcdsat_n18sparkle_closed_source*), which makes largest marginal contribution, ranks 12th according to stand-alone performance.

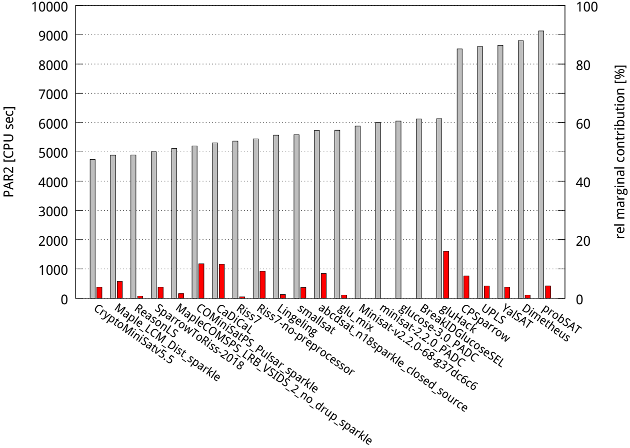
Additional insights:
Training & testing
on similar instances

What if we could train on instances from the same families as testing instances?

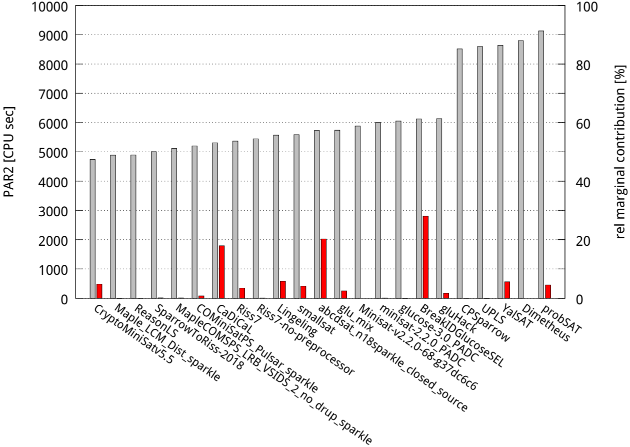
- testing set: 400 instances, 23 instance families
- split evaluation set from main track of 2018 SAT Competition into disjoint training and testing sets
  - for each instance family, ≈50% of instances → training set, remaining instances → testing set
- new training set: 195 instances from 23 families
- new testing set: 205 instances from 23 families

- PAR2 on new testing set:
  - SBS: 4739.87
  - VBS: 2498.68
  - Sparkle Selector: 3317.72 (75.3% of gap closed)

Stand-alone and relative marginal contribution on new testing set, with hors-concours solvers



Stand-alone and relative marginal contribution to VBS on new testing set, with hors-concours solvers



Observations

- The performance gap between Sparkle selector and VBS on similar training and testing instances is much smaller than for different training and testing instances.
- Solver rankings according to marginal contribution are quite different from the traditional stand-alone ranking.
- Several solvers that don't have high stand-alone performance make considerable contribution to the Sparkle selector and the VBS.

Advantages of Sparkle challenge over traditional competition:

- makes it easier to gain recognition for specialised techniques
- better reflects and makes accessible state of the art
- provides incentive to improve true state of the art

Further use of Sparkle:

- Sparkle Planning Challenge 2019:
  http://ada.liacs.nl/events/sparkle-planning-19
- continuous solver evaluation (as community service)
- experimentation platform for algorithm selection, configuration, programming by optimisation (PbO)