Brief report of the results of the experiments on the randomly generated set of problems

April 2017

The tests were run on an Intel Core i3 3.3 Ghz with 12Gb RAM, running under Linux (Debian distribution, release 8.7). When ran on the randomly generated tests, the provers were given 1 minute time.

1 Evaluation of the optimization techniques

Semantic branching seems not to affect significantly the performances of the prover. On the whole random test set, in fact, the prover either with the default settings or by disabling semantic branching could not solve about 4% of the problems in the allowed one minute time. On the contrary, disabling backjumping doubles the percentage of timeouts.

The diagram in Figure 1 plots the percentage of timeouts against the problem modal depth. They confirm that semantic branching has almost no impact on the prover performances, that instead improve significantly with backjumping.

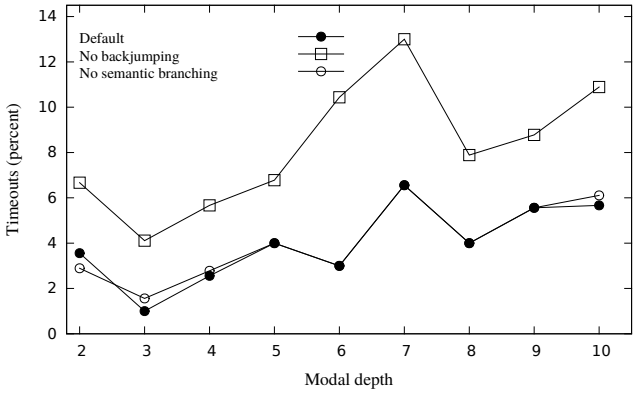


Figure 1: Impact of backjumping and semantic branching on Sibyl behaviour

2 Impact of the number of assertions on Sibyl's performances

The diagram in Figure 2 plots the percentage of timeouts on the sets of tests against the modal depth. It clearly shows that the number of assertions has no significant impact on the behaviour of the prover.

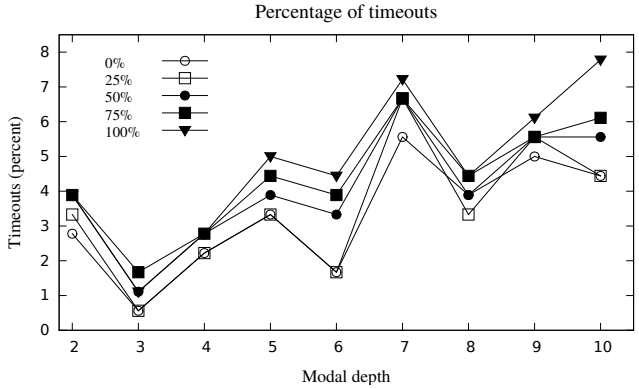


Figure 2: Impact of the number of assertions on Sibyl behaviour

3 Comparison with the SPASS prover

Considering the whole test set, SPASS (v. 3.5) failed to solve about 12% of the 8100 problems in the allowed one minute time.

The diagram of Figure 3 is the analogous of that of Figure 2 but for the SPASS prover. Comparing the two diagrams, it is apparent that SPASS is to be much more sensible than Sibyl to the presence of transitivity assertions and relation hierarchies: it reaches more than 30% timeouts in the basic set of tests, while Sibyl always stays below 8%. A direct comparison is represented by the diagram in Figure 4, that plots the percentage of timeouts against the number of assertions in each test set. While in the tests with no assertions SPASS performs better than Sibyl, the performances of the former considerably degrade when the number of assertions increases, reaching about 21% of timeout cases, against Sibyl's 4.75%.

A clearer view of the relative behaviour of the provers is obtained by analyzing the relation between the percentage of failures and the problem modal degree separately in the test sets with different numbers of assertions. The diagrams of Figure 5 consider the sets with no assertions (a), 50% of the assertions w.r.t. the basic set (b), and the basic set itself (c). Each of them plots the percentage of timeouts of the two provers against the problem modal depth in each group of problems.

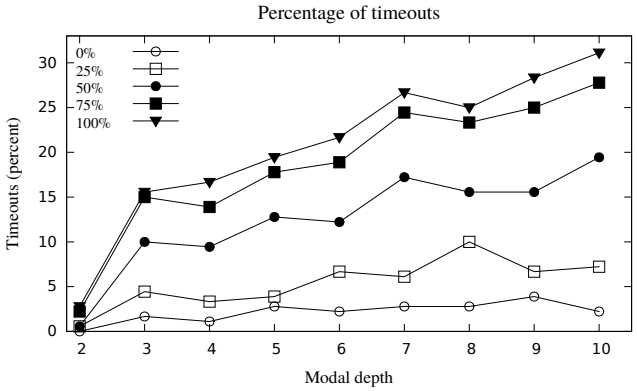


Figure 3: Impact of the number of assertions on the behaviour of SPASS

Considering the number of assertions as the leading parameter in analyzing the results is not very accurate, since tests with the same percentage of assertions are grouped together, but the number of formulae and the modal degree differ in tests belonging to the same group. Actually, each group of tests with the same percentage of assertions and the same modal depth should be analyzed separately.

An example can be found in Figure 6, whose diagrams analyze the midway case of tests with modal degree 6 and 50% assertions w.r.t the basic set. Diagram (a) shows the sat/unsat partition of the test set and diagram (b) shows the percentage of timeouts of the provers. Apparently, no significant new insight on the behaviour of the provers seems to be gained by such a more detailed analysis.

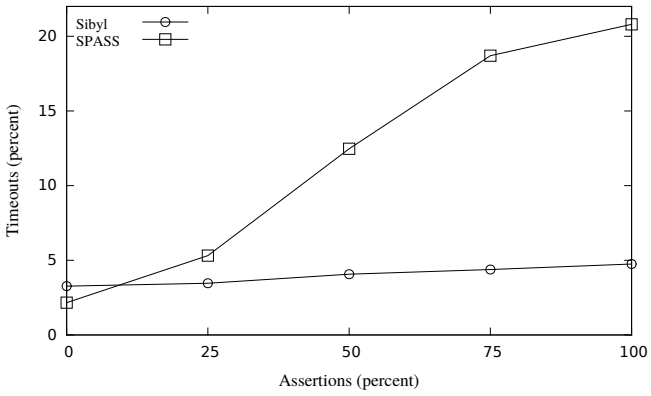
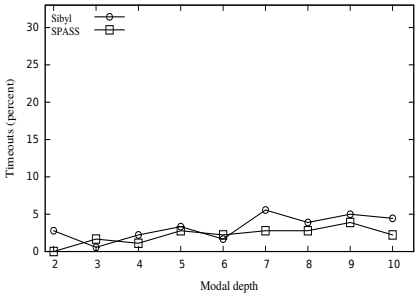
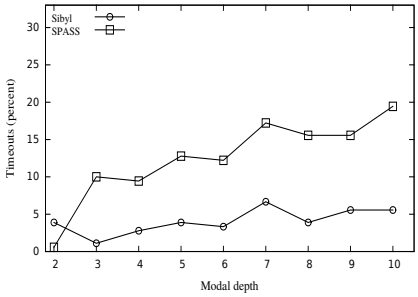


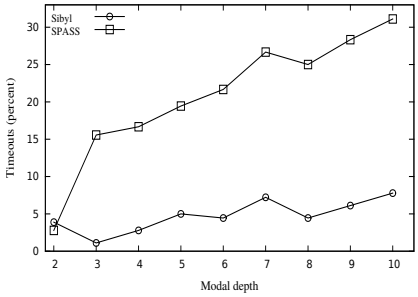
Figure 4: Sibyl and SPASS performances: timeouts vs number of assertions



(a) No assertions at all

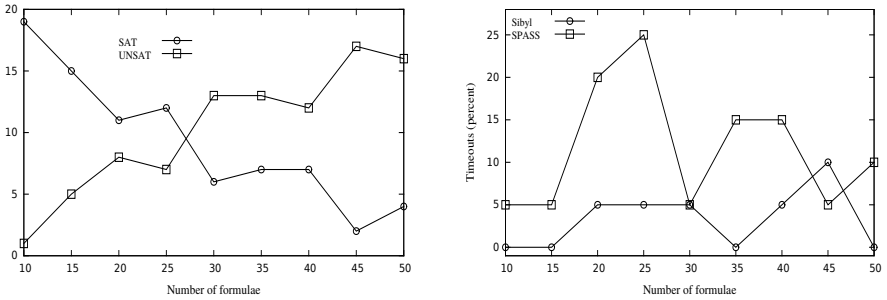


(b) 50% assertions



(c) 100% assertions

Figure 5: Sibyl and SPASS performances



(a) Sat/unsat partition (b) Percentage of timeouts

Figure 6: The test set with 50% assertions and modal depth 6