An Anecdote on Evaluating QBF Solvers and Quantifier Alternations*

Florian Lonsing[1] and Uwe Egly[2]

[1] Computer Science Department, Stanford University, Stanford, CA 94305, USA
[2] Institute of Logic and Computation, TU Wien, 1040 Vienna, Austria

Abstract. On the occasion of the 25th anniversary of the *International Conference on Principles and Practice of Constraint Programming (CP)*, we are glad to present the history of our paper entitled *Evaluating QBF Solvers: Quantifier Alternations Matter* that was presented at CP 2018. Our paper was finally accepted at CP 2018 after an 18-month odyssey, where it was rejected three times (in different versions) from other top-tier conferences.

1 Scientific Context

In our CP 2018 [14] paper[3] (quoting the abstract verbatim in the following), *we present an experimental study of the effects of quantifier alternations on the evaluation of quantified Boolean formula (QBF) solvers. The number of quantifier alternations in a QBF in prenex conjunctive normal form (PCNF) is directly related to the theoretical hardness of the respective QBF satisfiability problem in the polynomial hierarchy. We show empirically that the performance of solvers based on different solving paradigms substantially varies depending on the numbers of alternations in PCNFs. In related theoretical work, quantifier alternations have become the focus of understanding the strengths and weaknesses of various QBF proof systems implemented in solvers. Our results motivate the development of methods to evaluate orthogonal solving paradigms by taking quantifier alternations into account. This is necessary to showcase the broad range of existing QBF solving paradigms for practical QBF applications. Moreover, we highlight the potential of combining different approaches and QBF proof systems in solvers.*

In contrast to the satisfiability problem of propositional logic (SAT), which is NP-complete, the satisfiability problem of QBFs with arbitrarily nested quantifiers is PSPACE-complete, cf. [11]. Moreover, while implementations of the CDCL algorithm [17] dominate the field of SAT solving, the landscape of paradigms used to solve QBFs is much more diverse.

* The work that resulted in our CP 2018 paper [14], which is the topic of this note, was carried out while the first author was employed at the Institute of Logic and Computation, TU Wien, Austria, and was supported by the Austrian Science Fund (FWF) under grant S11409-N23.

[3] A self-archived version of our paper with an appendix containing additional experimental results is available on arXiv [15].

One of the main goals of our CP 2018 paper was to point out the importance of reflecting the diversity of QBF solving paradigms in empirical evaluations. The choice of benchmark problems may have an impact on how solving paradigms with complementary strengths are reflected in experimental results. This phenomenon was already pointed out by John N. Hooker in his paper *Testing Heuristics: We Have It All Wrong* [9] that is one of the key references relevant for our work. We quote from Hooker's paper [9]: *Once a set of canonical problems has become accepted, new methods that have strengths complementary to those of the old ones are at a disadvantage on the accepted problem sets. They are less likely to be judged successful by their authors and less likely to be published. So algorithms that excel on the canon have a selective advantage.*

It is an intriguing coincidence that John N. Hooker was also the PC chair of CP 2018, where our paper finally was accepted.

2 The History of our Paper: Fourth Time's a Charm

In the following anecdote, we would like to tell the history and origin of our paper from setting up an early sketch of ideas up to the final publication at CP 2018. In the process towards the final publication, we went through three rejections of our paper (in different versions) from top-tier conferences. We think that our story gives an example of how the belief in one's own ideas combined with perseverance, patience, dedication, resilience, and—most importantly—a lot of hard work can lead to a publication at a top-tier venue like CP.

Prequel. The first concept of our later paper came up in September 2016 in the wake of a QBF survey talk[4] that the first author held at the *Dagstuhl seminar 16381: SAT and Interactions* [4],[5] which was co-organized by the second author.

In the presentation of experimental results for the talk, we highlighted the diversity observed empirically between QBF solvers based on expansion [1,6,10] and on QCDCL [8,12,18]. QCDCL solvers tended to perform better on QBFs with many alternations while expansion performed better on QBFs with few alternations. As an open problem, we stated the question of how the empirical hardness of instances with a certain number of alternations could be better understood, and what the role of alternations in general in the hardness of instances is. We had a vague gut feeling that alternations should play a role but at that time had no ideas how to explain their importance in empirical nor in theoretical hardness. (In related work on QBF proof complexity, focus was put on alternations [2,5,7]. As a recent result [3], the empirical observations we made in our CP 2018 paper related to the diversity of QCDCL and expansion were confirmed theoretically.)

[4] QBF survey talk slides (last accessed in September 2019): `http://www.florianlonsing.com/talks/Lonsing-Dagstuhl-2016-talk.pdf`

[5] Dagstuhl seminar *SAT and Interactions* website (last accessed in September 2019): `https://www.dagstuhl.de/de/programm/kalender/semhp/?semnr=16381`

Following the Dagstuhl seminar, we came up with the idea to conduct a comprehensive empirical study based on the benchmarks, solvers, and preprocessors from the 2016 QBFEVAL competition, which were the most recent tools and benchmarks available back then. We also planned to carry out a virtual best solver analysis (VBS) and to analyze solver performance on instances divided into classes based on their numbers of alternations. This way, we aimed to highlight performance diversity of solvers implementing different solving paradigms on instances having different numbers of alternations. In fall 2016, we ran numerous experiments and gathered plenty of experimental data and started to turn our results into a paper.

At that time, we were ready to enter what would turn out as an 18-month odyssey, resulting in our submitted papers being rejected three times from top-tier conferences and finally accepted at CP 2018.

First Submission—Reject! In early 2017, we submitted our results as an 8-page short paper to a top-tier conference. We decided to submit a short paper because we thought we would be able to convey our main message, i.e., solver performance diversity with respect to alternations, in a concise way. (In hindsight, that decision may have been detrimental at that time since our paper got accepted as a regular, 15-page paper at CP 2018, and the additional pages were crucial to add important discussions and more in-depth analysis.) At the time of submission, we posted an extended version of the submitted paper on arXiv with supplementary experimental results that we could not include due to the space constraints. We updated that version on arXiv when we re-submitted our revised paper to other conferences, as described below.

Our paper was rejected. However, the reviews were not discouraging and did not point out severe technical issues in our evaluation. One concern that was raised was that the results may not come as a surprise.

Second Submission—Reject! We decided to revise our paper based on the feedback received with the first rejection and tried to sharpen our message. Shortly after the rejection of the first submission, in mid 2017, we submitted the revised, 8-page short paper to another top-tier conference. That submitted version still was based on benchmarks and solvers from QBFEVAL 2016, the most recent solvers and benchmarks at that time. Our paper was again rejected, but with much more negative feedback, including the non-constructive feedback in one review that our contribution was *inadequate for a conference presentation*, without providing arguments that would justify such inadequacy.

We were quite frustrated and decided to take a break. Submitting our paper to a low-quality venue was by no means an option for us. Instead, we planned to extend our short paper into a long one to be submitted to a journal.

We started working on the journal version in the fall of 2017. At that time, the benchmarks and solvers from the 2017 QBFEVAL competition were available. Therefore, we ran all experiments from scratch using the new benchmarks and a larger set of solvers. Unfortunately, after having put in a lot of work in running experiments and analyzing data, the journal paper never materialized. At the

same time we had been getting excited with another project, i.e., our later IJCAR 2018 paper [16], and soon entirely focused on that new project. Consequently, as time went by, we abandoned the planned journal paper and actually were not expecting to be able to get back to it again.

Third Submission—Reject! In early 2018, after we had completed working on our IJCAR 2018 paper [16], during a social dinner we chatted with a colleague who was interested in our abandoned work and inspired us to revisit it. We gave it another try. As the submission deadline was coming up, we again prepared an 8-page short paper that was based on the new data using benchmarks and solvers from QBFEVAL 2017 that we had gathered for our planned journal paper. We submitted the paper to the same conference as in the second submission. Our paper was again rejected, and like with the second submission, we were very frustrated with non-constructive feedback provided by one review.

Fourth Submission—ACCEPT! Being quite desperate from the third rejection, we decided to give it one last try because we were running out of time: the funding of the position of the first author would expire in the end of September 2018, with the potential risk of finally dropping out of academia. Given these circumstances, we prepared a regular 15-page paper to have more space to present an in-depth analysis. (As conjectured above, our previous submissions may have been more successful if we had prepared a long paper from the very beginning instead of a short one.) We decided to submit to CP, where we already had made a nice experience before [13]. We were very glad that our paper was finally accepted and were delighted to receive highly constructive feedback in the reviews.

Acknowledgments. We would like to thank Eugene Freuder for organizing the CP Anniversary Volume and for inviting us to contribute this note to the volume.

References

1. Ayari, A., Basin, D.A.: QUBOS: Deciding Quantified Boolean Logic Using Propositional Satisfiability Solvers. In: FMCAD. LNCS, vol. 2517, pp. 187–201. Springer (2002)
2. Beyersdorff, O., Blinkhorn, J., Hinde, L.: Size, Cost and Capacity: A Semantic Technique for Hard Random QBFs. In: ITCS. LIPIcs, vol. 94, pp. 9:1–9:18. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik (2018)
3. Beyersdorff, O., Chew, L., Clymo, J., Mahajan, M.: Short Proofs in QBF Expansion. In: SAT. LNCS, vol. 11628, pp. 19–35. Springer (2019)
4. Beyersdorff, O., Creignou, N., Egly, U., Vollmer, H.: SAT and Interactions (Dagstuhl Seminar 16381). Dagstuhl Reports **6**(9), 74–93 (2016), `https://doi.org/10.4230/DagRep.6.9.74`
5. Beyersdorff, O., Hinde, L., Pich, J.: Reasons for Hardness in QBF Proof Systems. In: FSTTCS. LIPIcs, vol. 93, pp. 14:1–14:15. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik (2017)
6. Biere, A.: Resolve and Expand. In: SAT. LNCS, vol. 3542, pp. 59–70. Springer (2004)

7. Chen, H.: Proof Complexity Modulo the Polynomial Hierarchy: Understanding Alternation as a Source of Hardness. TOCT **9**(3), 15:1–15:20 (2017)
8. Giunchiglia, E., Narizzano, M., Tacchella, A.: Clause/Term Resolution and Learning in the Evaluation of Quantified Boolean Formulas. JAIR **26**, 371–416 (2006)
9. Hooker, J.N.: Testing heuristics: We have it all wrong. J. Heuristics **1**(1), 33–42 (1995)
10. Janota, M., Klieber, W., Marques-Silva, J., Clarke, E.: Solving QBF with Counterexample Guided Refinement. Artif. Intell. **234**, 1–25 (2016)
11. Kleine Büning, H., Bubeck, U.: Theory of Quantified Boolean Formulas. In: Handbook of Satisfiability, FAIA, vol. 185, pp. 735–760. IOS Press (2009)
12. Letz, R.: Lemma and Model Caching in Decision Procedures for Quantified Boolean Formulas. In: TABLEAUX. LNCS, vol. 2381, pp. 160–175. Springer (2002)
13. Lonsing, F., Egly, U.: Incremental QBF Solving. In: CP. LNCS, vol. 8656, pp. 514–530. Springer (2014)
14. Lonsing, F., Egly, U.: Evaluating QBF Solvers: Quantifier Alternations Matter. In: CP. LNCS, vol. 11008, pp. 276–294. Springer (2018)
15. Lonsing, F., Egly, U.: Evaluating QBF Solvers: Quantifier Alternations Matter. CoRR **abs/1701.06612** (2018), `http://arxiv.org/abs/1701.06612`, CP 2018 proceedings version with appendix
16. Lonsing, F., Egly, U.: QRAT^+: Generalizing QRAT by a More Powerful QBF Redundancy Property. In: IJCAR. LNCS, vol. 10900, pp. 161–177. Springer (2018)
17. Silva, J.P.M., Lynce, I., Malik, S.: Conflict-Driven Clause Learning SAT Solvers. In: Handbook of Satisfiability, FAIA, vol. 185, pp. 131–153. IOS Press (2009)
18. Zhang, L., Malik, S.: Conflict Driven Learning in a Quantified Boolean Satisfiability Solver. In: ICCAD. pp. 442–449. ACM / IEEE Computer Society (2002)