# Towards a Protocol for Benchmark Selection in IPC

**Mauro Vallati**
PARK Research group
University of Huddersfield
m.vallati@hud.ac.uk

**Tiago Vaquero**
California Institute of Technology
Massachussets Institute of Technology
tvaquero@caltech.edu

## Abstract

The planning competition has traditionally played an important role in motivating research and advances in Planning & Scheduling techniques. Despite its pivotal role in the planning community, some aspects of the competition have not been engineered yet. This is the case for the protocol for selecting benchmark instances. Benchmarks are of critical importance, since they can significantly affect competition results.

In this paper we describe desirable properties of a selection protocol, discuss methods exploited in past SAT and planning competitions, and identify challenges that organisers of future competitions have to address in order to improve reliability and usefulness of the insights gained by looking at competitions' results.

## Introduction

Competitions are important events to improve a particular research area. Some examples are the International SATisfability Competition (SAT), the Conference on Automated Deduction ATP System Competition (CASC), the Trading Agent Competition (TAC) and many others. This strategy has been used by the AI Planning & Scheduling (P&S) community to develop innovative planning techniques since 1998 through the *International Planning Competition* (IPC) and also to promote the development of knowledge engineering tools and systems since 2005 through the *International Competition on Knowledge Engineering for Planning and Scheduling* (ICKEPS). Both competitions attract researchers in the AI community, especially IPC due to its motivational aspect of developing better and more powerful planning engines to address increasingly large (and hopefully complex) problems. Techniques tested in competitions are then available to be used in real-world applications.

In the IPC, participating planning engines are tested against several benchmark problems, a few of them are inspired by real-world problems. The selection and design of these benchmark domains and problems instances have become one of the main challenges encountered during the organisation of this competition. Given a set of target domains, it is well known that benchmark instances selection can directly impact results (Howe and Dahlman 2002). Moreover, the very small number of theoretical studies on complexity of instances and transition phase (which changes between domains) (Helmert 2006; Rintanen 2004), and the growing number of participants exploiting different planning techniques, make the selection and decision of IPC problem instances a significant challenge for organisers. It is worth noting that the selection of benchmarks is only one of the many difficulties faced by competition organisers. Organising a competition requires a significant amount of work. In fact, organisers of past competitions had to face a lot of pressure for selecting and generating new domains and problems by themselves. For addressing the selection issues, organisers have been using different selection strategies and criteria throughout the years, which have been the target of several post-competition discussions and criticisms. Given the importance of competitions for the community, the responsibility of generating benchmarks should be shared between all the members, rather than delegated to organisers only. Also for this reason, a protocol for selecting benchmarks is highly desirable.

In this paper, we emphasise and highlight the need for a protocol for selecting IPC benchmark problems that: is transparent; reproducible; avoids bias to any particular planning technique; adapts to the state-of-the-art and the existing participating planners; allows and motivates new benchmark domains to be added (e.g., challenging domains from ICKEPS); supports the realisation and exclusion of outdated and uninteresting problems for the participating planners; aims to evaluate and understand the technological progress in the long-term and, possibly, fosters the evaluation of planning techniques in new potential real-world applications.

## Desirable Properties

In this section we provide two lists of desirable properties: one for the selection protocol itself, and one for the selected benchmark instances. Although properties of the instances are induced by a proper selection protocol, thus are somehow implicitly guaranteed by the protocol properties, we prefer to divide the lists and make their properties clear for the sake of readability. Desirable properties of a selection protocol are:

- **Transparency**. Others can follow the method and, while considering the same "environmental" conditions, produce the same (sort of) problems.
- **Generality**. It can be applied to any set of planners, on

any target domain.

- **Unbiased**. It does not favour a system against another.
- **History-aware**. This avoids tailored algorithms. It limits the impact of approaches based on learning, which exploits problems and domains from previous competitions.
- **Progress-driven**. It motivates technological progress in new domains and problems.

In order to be useful, a set of benchmark instances must include problems that are:

- **Challenging**. Problems must not be trivially solvable or unsolvable. They must provide information about the performance of participants.
- **Interesting**. They investigate possible exploitation of planning in real-world scenario, or test innovative features.
- **Diverse**. They do not refer to the same kind of problems or models.
- **Finite**. The selected instances must be in a finite number. Moreover, the smaller the set of benchmarks, the easier is to re-run the competition and reproduce results.

It should be noted that the properties we introduce consider also the importance of planning competitions for the community. IPC is a major event of the planning community, therefore it should provide also some guidance about applications, limitations and strengths of the existing solvers, as well as identifying future avenues of research while situating the technological progress.

## Existing Protocols

In this section we describe the existing techniques that have been used for selecting benchmark instances in the International SAT competition, and in the deterministic track of International Planning Competitions.

### SAT competition

We observed that a very similar selection protocols have been used in SAT competitions since 2012 (Balint et al. 2012; 2013; Belov et al. 2014a). Hereinafter, we will focus on the policies used in the latest edition.

In the 2014 SAT competition, two main sets of benchmarks are considered: (i) uniform random and (ii) application and hard combinatorial. The way in which corresponding instances are selected is different. For the first set, two different sizes – medium and huge – of uniform random formulae are generated by using existing generators (Belov et al. 2014c). The huge benchmarks have millions of clauses and a clause-to-variable ratio ranges from far from the phase-transition ratio to relatively close. On the other hand, medium benchmarks are smaller, but have clause-to-variable ratio equal to the phase-transition ratio. Remarkably, given the theoretical knowledge about complexity of random SAT instances (Rossi, Van Beek, and Walsh 2006), the uniform random benchmark selection does not need to consider the performance of actual solvers; complexity is theoretically assessed.

A different protocol is used for selecting application and hard combinatorial benchmarks (Belov et al. 2014b). In such tracks, it is important to consider the performance of solvers. Firstly, benchmarks collected by previous competitions (either used or unused) and newly submitted benchmarks have been divided into buckets. The assignment to a specific bucket is guided by the combinatorial problem the benchmark originates from, and the submitter. This partition is done in order to limit possible biases deriving from the use of large number of instances that refers to the same problem, or that have been used in previous competitions.

The empirical hardness of benchmarks is evaluated by using five well-performing solvers, per track, from the 2012 SAT challenge. To consider differences in performance due to environment / technological improvements, the CPU runtimes have been scaled. According to solvers performance, benchmarks have been rated as follows:

**Easy** Benchmarks solved by all the considered solvers in less then 1/10-th of the competition's timeout.

**Medium** Benchmarks solved by all the solvers within the competition's timeout.

**Hard** Benchmarks solved by at least one solver within the double of the timeout, and not solved by at least one solver within the competition's timeout.

**Too-hard** Benchmarks unsolved by any solver within two times the considered cutoff time.

For each track, 300 benchmarks are selected from the medium and easy classes. The selection process, that must provide a 50-50 ratio between satisfiable and unsatisfiable formulae, is controlled by the following constraints:

1. no more than 10% of the instances should come from the same bucket;
2. the percentage of new benchmarks should be as high as possible;
3. the ratio of Medium to Hard benchmarks should be as close to 50-50 as possible. However, this constraint has been relaxed by selecting 20% of the benchmarks from Medium, Hard and Too-hard classes. This for reducing the influence of selected solvers.
4. the performance of the solvers used for the evaluation of the benchmarks should be as uniform as possible, to avoid bias due to a specific technique.

### International Planning Competition

In the following we describe the protocols used by the organisers of the deterministic track of the IPC 2011 and 2014. We focus on deterministic track since it is the largest one, in terms of participants, and thus requires clearly defined strategies for benchmark selection.

Before describing the protocols, we would remark that over time, IPC organisers have put more and more effort in studying suitable techniques for selecting benchmark instances. This is due to a number or reasons: firstly, the growing number of participants; secondly, the wide range of problems and domains that can be modelled in PDDL; thirdly, the importance of guaranteeing an unbiased set of instances; and finally, since the IPC has usually been held every 2-3 years, the difficulty of estimating the progress of the state of the art.

**2011 edition** In the deterministic track of the IPC 2011, organisers adopted two different methods for selecting benchmarks, according to the fact that data on their difficulty were available or not (López, Celorrio, and Olaya 2015).

For newly introduced domains, state-of-the-art planners are used for evaluating the difficulty of reduced test sets of problems, which are generated using randomised generators, with some specific parameters. A cutoff of 300 seconds has been considered for these tests. The easiest problems are those solved in tens of seconds, the most difficult problems are those which are unsolvable by considered planners, in a 300 seconds cutoff. By following a trial-and-error procedure, a suitable set of parameters is found, and can be used for generating 20 benchmark instances.

For domains used in previous competitions, the publicly available data is used for ranking planning tasks according to their expected difficulty, measured by using the Glicko score, and for selecting them.

**2014 edition** In the deterministic track of the IPC 2014, organisers provided a protocol for selecting, within a specific domain, a set of suitable instances, tailored for the participating planners.[1] They defined as "trivial" instances in which almost all the planners performed very similar – in terms of quality of plans for satisficing subtracks, and runtime for the Agile subtrack – and "too complex" those instances which are not solved by any planner. For each target domain:

1. identify size;
2. given the sizes, generate between 30 and 50 instances per domain, using available generators;
3. anonymise planners;
4. run all the planners on the generated instances;
5. collect results, in terms of solved problems and quality of solutions;
6. order problems by number of planners which solved them;
7. selection of 20 benchmarks.

In the first step, if the domain has been already used in previous IPCs, then the sizes of larger benchmark problems (top half), and also extended them, following the "trend" used by organisers, are taken. Otherwise, some well-known planners, either from literature or from IPC 2011, are used. If no generator is available, all the available instances have been considered.

In step 7, if between (circa) 10 and 20 instances have been solved by some considered planners, then select the top 20 instances accordingly to the order in step 6. If most of the instances are either trivial or too complex, according to planners' performance, then the process is started again from step 1. Otherwise, trivial and too complex instances are removed, in order to obtain a final set of 20 problems.

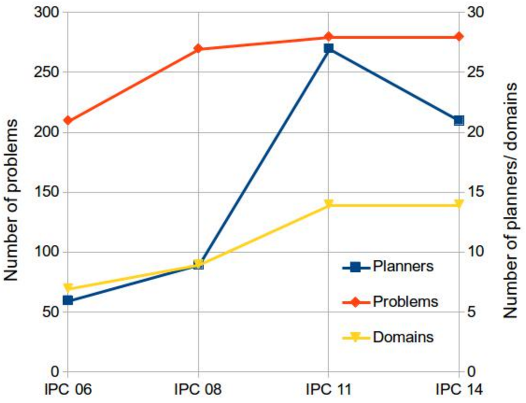

Figure 1: Number of participants, instances and domains considered in IPCs since 2006.

## Challenges

Having provided the desirable properties of benchmarks selection protocols, and having introduced protocols used in two major competitions within the artificial intelligence area, in this section we discuss the challenges that should be faced, in order to furtherly improve the importance and significance of competitions.

As a first remark, we observe that the notion of *quality* of benchmarks is missing. In the protocols introduced in the previous section, organisers do not explicitly mention this aspect of benchmarks. Although some properties of useful benchmarks have been identified – and we introduced a few of them – having a formal definition of quality would be extremely helpful, and would also allow the exploitation of knowledge engineering approaches for defining sound selection protocols. Remarkably, a first effort in this direction was done by the IPC 2011 organisers. They tried to measure quality of planning tasks as fitness to a normal distribution (López, Celorrio, and Olaya 2015). In the competitions context, quality of benchmarks is not "static", but it depends on the current state of the art, as well as on the potential applications of the evaluated solvers. In planning, good quality instances should test planning techniques in real-world applications. As a matter of fact, the IPC should investigate pioneering uses of planning.

It is still unclear if there exists a *right* number of benchmarks. Figure 1 shows the variation of number of planners, instances and domains in sequential satisficing subtracks of the IPC over time. It should be noted that last IPCs organisers had a large set of domains to choose and start from, while in earlier IPCs benchmarks had to be developed. Commonly, it is believed that the larger the set of benchmarks, the more accurate the overall evaluation of participants. This leads to computationally expensive competitions, which require significant amount of CPU and human hours to be run (although most of the work can be done automatically through existing IPC software (Linares López, Celorrio, and Helmert 2013)). Moreover, large set of benchmarks can possibly include low quality instances, which introduce noise

[1] The protocol can be found at: http://helios.hud.ac.uk/scommv/IPC-14/selection.html

in the evaluation. It would be interesting to identify a formal way for estimating the number of required benchmarks for assessing the performance of a given number of planning systems, for facilitating planners exploitation outside the community.

It is worthy to note that in the IPC, differently for most of the other contests in AI, domains are explicitly given and are of primary importance. Domains strongly affect the structure of problems, thus having a significant impact on the planners' performance. Therefore, in the IPC both domains and instances have to be wisely selected. Although techniques have been described and exploited for selecting problems, the domain selection process is still some sort of obscure task, which has not been deeply investigated yet. Interestingly, in IPCs, the trend is to increase the number of considered domains, and reducing the number of problems per domain. Also, given the strong impact that different models of the same domain or different model refinements can have on the performance of planners (Riddle, Holte, and Barley 2011; Vaquero et al. 2010), it might be interesting to consider, within a competition, more than one specific PDDL model. Such different models do not have to exploit different PDDL-features; this has been done in previous IPCs. Sadly, state-of-the-art planners support a very limited subset of them. Here we suggest to test different ways of encoding the same domain, with the same set of PDDL features. For instance, using models of blocksworld using 3 or 4 operators, and evaluate planners by considering their average performance. The generation of different models of domains could be for example the scope of future ICKEPS competitions. Moreover, the analysis on the different performance of planners on different models can give useful insights on knowledge engineering aspects of domain modelling (e.g., given a particular model it might be possible to map the planners that would have the better performance). Such evaluation, if put in practice, will significantly increase the already high pressure on organisers. Once again, we would like to emphasise that generation and selection of benchmark should be done collectively by the community, and fostered by a shared protocol. For instance, models can be generated by exploiting crowdsourcing (Zhuo 2015).

All the competitions are using existing techniques for identifying a large set of promising benchmarks. Such techniques should be as various as possible, i.e. exploiting very different planning approaches, in order to avoid biases and identify challenging sets. The actual benchmarks are then usually selected by considering the performance of participants. Even though this reduces the number of useless instances –e.g., trivial or too complex– this can possibly introduce some bias. Specifically, benchmarks might be too focused on the competitors, thus ignoring the larger situation of the state of the art.

Current benchmarks selection protocols are mainly focused on the CPU time needed by a system for solving the given instances. This helps to discriminate between trivial, challenging and too complex instances. In tracks where the runtime is not considered in the metric, like the satisficing subtrack of the deterministic IPC, this approach can be improved by considering also aspects that are accounted for the metric. For instance, the presence of multiple solutions, with different costs, can be useful when evaluating planners that should return high quality plans.

It has been shown that different configurations of hardware and software can differently affect the performance of domain-independent planners (Howe and Dahlman 2002). Given that, it would be interesting to assess the reliability of a competition results, which are collected on a single system, with regards to their generalisation on different infrastructures. Potentially, running the competition few times on different systems and merging results would provide a more accurate evaluation, but of course, will be extremely costly.

Finally, a competition should also provide a clear picture about the progress of the state of the art, mainly with regards to the previous competition. This evaluation is twofold. Firstly, we are interested in evaluating the progress in terms of planning performance; i.e., new planners have to be faster, solve more and more problems, and/or find better plans. Secondly, the progress also involves the languages used for representing knowledge. In particular, are the new languages able to model more problems? Do they positively affect the performance of solvers? While the evaluation of the planners' performance progress seems to be mostly related to the size –as an indicator of complexity– of problems that can be solved, the evaluation of languages is mostly connected with knowledge engineering aspects. On this matter, a cooperation between the IPC and the ICKEPS is strongly suggested.

## Conclusion

Selecting benchmark instances is a critical task that every AI competition has to face. It has a dramatic impact on the final results and, given the pivotal role of competitions within AI communities, the selection of benchmarks strongly affects also the future development of the specific area. Given the importance of benchmark selection, and the high pressure organisers have to face on this regards, it would be desirable that the whole community supports organisers in this difficult task. In particular, this can be done also by exploiting a protocol. A proper protocol will lead to more reliable and informative competition results. Even though its central role, desirable characteristics and properties of a selection protocol have not been thoroughly discussed yet.

In this paper, we provide a list of desirable properties of both the selection protocol and the selected instances. We discuss methods used in SAT and Planning competitions for selecting benchmarks in order to gain useful insights on the limitations and strengths of the existing exploited approaches. Such gained knowledge is then used for highlighting challenges and, possibly, providing avenues for improving selection protocols in future planning competitions. In particular, we observe that: (i) a formal technique for selecting domain models is missing; (ii) it is unclear what should be the "right" number of benchmarks; (iii) it might be useful to consider the evaluation metric – used in the competition for evaluating planning systems – also in the selection protocol. Finally, we would remark the importance of the competition for assessing the progress of the state of the art, and for pioneering innovative applications of automated planning.

# References

Balint, A.; Belov, A.; Diepold, D.; Gerber, S.; Järvisalo, M.; and Sinz, C. 2012. Sat challenge 2012.

Balint, A.; Belov, A.; Heule, M. J.; and Järvisalo, M. 2013. Sat competition 2013.

Belov, A.; Diepold, D.; Heule, M. J.; and Järvisalo, M. 2014a. Sat competition 2014.

Belov, A.; Heule, M. J.; Diepold, D.; and Järvisalo, M. 2014b. The application and the hard combinatorial benchmarks in sat competition 2014. In *Proceedings of SAT competition 2014*, 81–82.

Belov, A.; Heule, M. J.; Diepold, D.; and Järvisalo, M. 2014c. Generating the uniform random benchmarks. In *Proceedings of SAT competition 2014*, 80.

Helmert, M. 2006. New complexity results for classical planning benchmarks. In *In International Conference on Automated Planning and Scheduling ICAPS*, 52–61.

Howe, A. E., and Dahlman, E. 2002. A critical assessment of benchmark comparison in planning. *Journal of Artificial Intelligence Research* 17(1):1–33.

Linares López, C.; Celorrio, S. J.; and Helmert, M. 2013. Automating the evaluation of planning systems. *AI Commun.* 26(4):331–354.

López, C. L.; Celorrio, S. J.; and Olaya, A. G. 2015. The deterministic part of the seventh international planning competition. *Artificial Intelligence*.

Riddle, P. J.; Holte, R. C.; and Barley, M. W. 2011. Does representation matter in the planning competition? In *Proceedings of the Ninth Symposium on Abstraction, Reformulation, and Approximation, SARA*.

Rintanen, J. 2004. Phase transitions in classical planning: An experimental study. In *In International Conference on Automated Planning and Scheduling ICAPS*, volume 2004, 101–110.

Rossi, F.; Van Beek, P.; and Walsh, T. 2006. *Handbook of constraint programming*. Elsevier.

Vaquero, T. S.; Silva, J. R.; Beck, J. C.; et al. 2010. Improving Planning Performance Through Post-Design Analysis. In *Proceedings of the ICAPS'10 Workshop on Knowledge Engineering for Planning and Scheduling. Toronto, Canada*, 45–52.

Zhuo, H. H. 2015. Crowdsourced action-model acquisition for planning. In *Proceedings of the AAAI Conference on Artificial Intelligence*.