Measurability and Reproducibility in University Timetabling Research: Discussion and Proposals

Andrea Schaerf and Luca Di Gaspero

Dipartimento di Ingegneria Elettrica, Gestionale e Meccanica
Università degli Studi di Udine
via delle Scienze 208, I-33100, Udine, Italy
{schaerf,l.digaspero}@uniud.it

Abstract. In this paper, we first discuss the level of compliance for timetabling research to two important research qualities, namely measurability and reproducibility, analyzing what we believe are the most important contributions in the literature. Secondly, we discuss some practices that, in our opinion, could contribute to the improvement on the two aforementioned qualities for future papers in timetabling research.

For the sake of brevity, we restrict our scope to university timetabling problems (exams, courses, or events), and thus we leave out other equally important timetabling problems, such as high-school, employee, and transportation timetabling.

1 Introduction

Thanks mainly to the PATAT conference series, researchers on timetabling problems have recently started to meet regularly to share experiences and results. This situation has the positive effect of generating both a common language and a common spirit that is the base ground for cross-fertilization of research groups in the timetabling community.

However, according to what we have seen at the recent PATAT conferences, the road for timetabling to become a well-established research community is still long. The main issue, in our opinion, is that most timetabling papers tend to describe the authors' specific problem and *ad hoc* solution algorithm without taking enough care of either the *measurability* or the *reproducibility* of the results. The reader is thus 'left alone' to judge the quality of the paper, and to understand what can be learned from it.

This issue is, to some extent, common to all the experimental areas of computer science and operations research, as clearly explained by Johnson in his seminal paper [17]. Nevertheless, we believe that this is particularly true in timetabling research, probably because of its shorter standing as a scientific community.

Regarding measurability (or comparability), we believe that several 'research infrastructures' are necessary in order to create the ground for truly measurable results. Specifically, they range from common formulations, to benchmark instances, to instance generators, to solution validators, and others. Related to it,

E.K. Burke and H. Rudová (Eds.): PATAT 2006, LNCS 3867, pp. 40–49, 2007.

but somewhat complementary, is the issue of reproducibility. To this aim, beside the features just mentioned, it would be also necessary to create the conditions for sharing code and/or executables among researchers.

In this paper, we try to describe the main contributions with respect to these crucial qualities of experimental research in timetabling, and we also present some personal opinions on how to proceed to improve on them. For the sake of brevity, we restrict our scope to university timetabling problems (exams, courses, or events), and we leave out other equally important timetabling problems, such as high-school, employee and transportation timetabling. Nevertheless, to some extent, the proposed guidelines can have a broader application to all timetabling domains.

In detail, we first survey what, in our opinion, are the most important steps that have been pursued so far in timetabling research in terms of either measurability or reproducibility of results (Section 2). Secondly, we propose our personal 'best practices' for improving these two qualities in the timetabling research (Section 3). Our aim is to encourage both the authors to write research papers of high level in these important aspects and the reviewers to demand it when judging a paper.

2 Significant Contributions

In this section, we review the most significant contributions to the aim of creating the ground for the development of high quality measurable and reproducible research in timetabling. We first discuss the 'standard' problem formulations, the benchmark instances (datasets), and the related file formats adopted. Next, we move to the comparison methods proposed, such as competitions between algorithms and statistical tools. Finally, we discuss the issue of the objective validation of the proposed results.

2.1 Problem Formulations and Benchmark Instances

It is well known that timetabling problems vary not only from country to country, but also from university to university, and even in different departments of the same university the problem is not quite the same (see, e.g., [27]).

Nevertheless, throughout the years it has been possible to define common underlying formulations that could be used for the comparison of algorithms. In fact, a few basic formulations have become standards *de facto*, as they have been used by many researchers. Needless to say, standard formulations allow the researchers to compare their results and to co-operate for the solution. Furthermore, in some cases algorithms developed for more complex *ad hoc* formulations can be adapted to the basic standard ones so as to assess their objective quality.

For the Examination Timetabling problem (ETTP), Carter et al. [7] propose a set of formulations which differ from each other based on some components of the objective function. Carter also makes available a set of benchmark instances [6] extracted from real data, which represent a large variety of different situations. Formulations and benchmarks by Carter have stimulated a large body