

# Reproducibility Report for ACM SIGMOD 2021 Paper: “REDS: Rule Extraction for Discovering Scenarios”

PHILIPPE CUDRÉ-MAUROUX, U. of Fribourg, Switzerland

The code and data provided allow to reproduce all results and plots from the original paper [1]. The code is relatively easy to run on a Windows machine, but more difficult to run on a different platform due to a few scripts and the limited availability of some tools (e.g., *RTools*, which provides a convenient tool-chain for building the exact R packages that the authors used, but that is only available on Windows platforms). As the scripts used to generate the plots include tens of thousands of individual experiments, running all the scripts can take more than one week on a modest machine.

## 1 INTRODUCTION

In the following, I briefly summarize my work on reproducing the results from “REDS: Rule Extraction for Discovering Scenarios”. The paper was accepted at SIGMOD 2021 and was authored by Vadim Arzamasov and Klemens Böhm from Karlsruhe Institute of Technology, whom I thank for their availability during the reproducibility process.

## 2 SUBMISSION

All the information required to reproduce the results of the paper are available on a dedicated GitHub repository<sup>1</sup>. The GitHub repository contains the following items:

- a README file describing the structure of the repository, the software and hardware setup the author used to run their experiments, as well as the process to run their code
- a package (“reds\_0.14.tar.gz”) containing the implementation of the various algorithms discussed in the paper, as well as the quality metrics and auxiliary functions used by the authors to produce their results. This package corresponds to commit 12da8bd1b535169975aa686d4-bdf4015d3a46563 from the REDS repository<sup>2</sup>
- a folder (“install\_requirements”) with an R script to install the various R packages used by the other scripts (see below) using the RTools toolchain<sup>3</sup>
- a folder (“third\_party”) with a Windows batch file (“get\_data.bat”) to download third-party data as well as several R scripts to generate the corresponding results (Table 5 and Figure 13 in the original paper)
- a folder (“main”) containing the main data used in the paper, a batch file to retrieve a third-party R script, as well as a series of R scripts to generate, analyze, and plot all the main results from the paper (i.e., Figures 6-12, & 14, as well as Tables 3-4)
- a GNU GPL license.

## 3 HARDWARE AND SOFTWARE ENVIRONMENT

The authors report that they used a virtual machine with 32 cores running at 2GHz (AMD EPYC 7551) and 128GB of memory to obtain their results. They used R version 3.6.3 and a number of specific R packages installed using Rtools<sup>35</sup>.

I started to reproduce the results of the paper using various Mac machines. After a few issues, however (see below), I decided to switch to a Windows environment. I used, in the end, a Lenovo

<sup>1</sup>[https://github.com/Arzik1987/REDS\\_experiments](https://github.com/Arzik1987/REDS_experiments)

<sup>2</sup><https://github.com/Arzik1987/reds>

<sup>3</sup><https://cran.r-project.org/bin/windows/Rtools/>

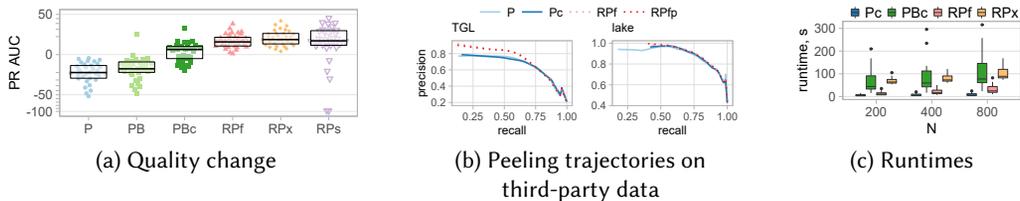


Fig. 1. Some performance results from the reproducibility runs

X1 Yoga Gen 4 with an Intel Core i7-8565U CPU (4 cores at 1.80 GHz) and 16 GB of RAM. I installed the exact same versions of R and the exact same packages as used by the authors.

## 4 REPRODUCIBILITY EVALUATION

### 4.1 Process

The process used to reproduce the results from the original paper is fairly straightforward:

- download the GitHub repository
- install R version 3.6.3 and RTools35
- run the installation script to install the various R packages using RTools
- run the batch files and the R scripts to generate the third-party results and graphs
- run the batch file and the R scripts to generate all the remaining results and graphs.

I started by following those different steps on various Mac machines. I could successfully run the Windows batch files manually (I contacted the authors to clarify a few points in this context), obtain the third-party data and scripts, and run most of the experiments. However, I was faced with many minor compatibility issues when running the last experiments and analyzing and plotting the results, due to a slight mismatch between the R libraries I used and the ones used by the authors (RTools being not readily available for Mac, I resorted to installing some of the libraries manually).

After some struggle, I decided to use an older laptop to run the scripts on a Windows platform. All the scripts ran pretty smoothly after this decision, and I could reproduce all results (modulo some negligible differences) without any major issues. I encountered a few problems when running the batch files, and a few errors and crashes when running the rest of the scripts, but nothing prevented me from generating all the results and plots in the end. As the scripts run tens of thousands of individual experiments to generate all results, it can take several days (or even weeks on older machines) to go through the entire experimental process described above.

### 4.2 Results

I could successfully reproduce all the results from the original paper. I give below a few examples of the results I obtained. Figure 1 (a) gives the PR AUC relative scores for the various REDS variants (corresponding to the first subplot of Figure 7 in the original paper). Figures 1 (b) depicts the peeling trajectories using third-party data (corresponding to Figure 13 in the original paper). Figure 1 (c), finally, is the only plot that is substantially different from the original paper, as it focuses on runtimes, and as I used, in the end, a machine that was significantly less powerful than the one used by the authors (as a result, the values reported in the figure are roughly twice higher than the ones reported in the first subplot of Figure 9 in the original paper).

## 5 SUMMARY

The material provided by the authors makes the reproducibility of their results relatively easy. In conclusion, the results I obtained fully support the claims of the original paper.

## REFERENCES

- [1] Vadim Arzamasov and Klemens Böhm. 2021. REDS: Rule Extraction for Discovering Scenarios. In *SIGMOD '21: International Conference on Management of Data, Virtual Event, China, June 20-25, 2021*. ACM, 115–128. <https://doi.org/10.1145/3448016.3457301>