Reproducibility Report for ACM SIGMOD 2022 Paper: "HypeR: Hypothetical Reasoning With What-If and How-To Queries Using a Probabilistic Causal Approach"

YUXIN TANG, Rice University, USA

This paper proposes a framework called HypeR that enables "what-if" and "how-to" analyses for databases. These analyses help users explore hypothetical scenarios and devise strategies without actually changing the database. Current methods examine the impact of potential updates on a specific query, but real-world updates could influence other parts of the database due to hidden dependencies. HypeR accommodates these dependencies using a probabilistic causal model. It extends SQL syntax to express these hypothetical scenarios, provides efficient algorithms and optimizations to calculate their outcomes using causal and probabilistic database concepts, and validates the approach's effectiveness through experimentation.

1 INTRODUCTION

This reproducibility report concerns the paper HypeR: Hypothetical Reasoning With What-If and How-To Queries Using a Probabilistic Causal Approach [3], which is a joint work between the University of Chicago, Duke University and University of California, San Diego. The paper's experiments are based on five different datasets, including **Adult**, **German**, **Amazon**, **German-Syn**, and **Student-Syn**. The authors also evaluate the quality of solutions on What-if and How-to queries generated by HypeR with respect to the ground truth and baselines over synthetic datasets. To that end, the authors analyze HypeR's runtime on factors like query complexity and dataset properties, including the number of tuples and the structure of the causal graph.

2 SUBMISSION

The reproducibility submission [1] contained all source code and scripts necessary to run the result. With an extra explanation included [2], source code can be used for the following purposes:

- Download all the necessary code and data (adult and german)
- Install all the dependencies (mip, seaborn, pandas, numpy, sklearn, scikit-learn)
- Generate all the necessary plots (Figure 6, 8a, 8b, 9, 10a, 10b, 11a, 11b, 12a, 12b)
- Reproduce individual plots

3 HARDWARE AND SOFTWARE ENVIRONMENT

The following table shows the environment used in the paper and in the reproducibility. The authors use a MacOS laptop with 16GB RAM and a 2.3 GHz Dual-Core Intel Core i5 processor. Authors use random forest regressor to estimate conditional probabilities.

Table 1. Hardware & Software environment

	Paper	Repro Review
CPU	Intel Core i5 processor	Intel E5-2650
cores	2	48
GHz	2.3	2.2
RAM	16GB	256GB
OS	MacOS	Ubuntu 18.04

4 REPRODUCIBILITY EVALUATION

4.1 Process

The `reproduce.sh` and `run_all.sh` scripts were entirely sufficient for the task at hand. The only change I have made is changing the version of scikit-learn to 0.23.2. The set-up and installation process took a few minutes, while the experiments ran overnight for approximately 36 hours. Upon completion, output directories contained a series of plots. Despite occasional warnings, the process was error-free and proceeded without any issues.

4.2 Results

During the evaluation process, all results outlined in the original paper were meticulously cross-verified. Some of the experiments have some speedup that could be attributed to enhanced CPU speeds during the evaluation. Any minor deviations observed did not significantly alter the paper's primary findings and can be ascribed to basic hardware differences, such as superior cache and faster CPUs in the evaluation hardware. Notable, though minor deviations include:
Figure 6a - **Figure 1(a)** The query output drops to 0.60 for the case with a sample size of around 1K.
Figure 9a - **Figure 2(b)** Hyper's output quality is better than the ground truth - Opt-discrete under all the cases except when # of Buckets equals to 10.
Figure 10a - **Figure 2(c)** The grey bar (Hyper) shown on Figure 10a disappears on the reproduced plot.
Figure 11b - **Figure 2(d)** The running time for Opt-HowTo is reduced from the range [1000,4000] to [200,1200] during the reproducing process.

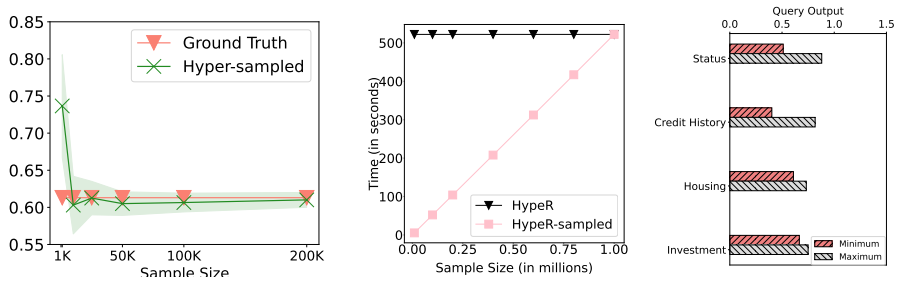

Fig. 1. Reproduced Figures 6a, 6b, and 8a.

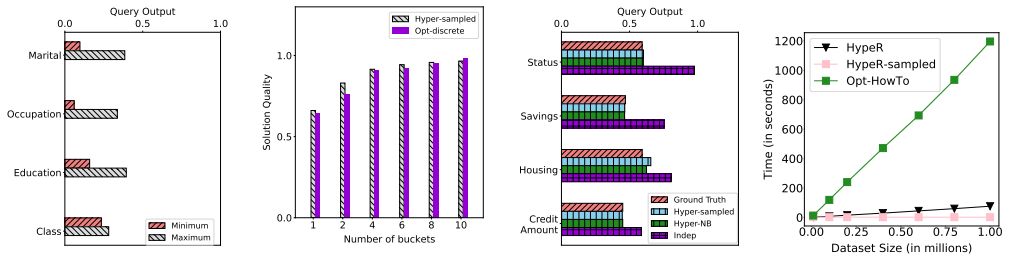

Fig. 2. Reproduced Figures 8b, 9a, 10a, and 11b.

5 SUMMARY

Most claims in the paper have been successfully reproduced with minimal effort, making this an excellent contribution to SIGMOD reproducibility. It sets a high benchmark for papers presented at SIGMOD.

REFERENCES

[1] 2023. https://github.com/sainyam/Hyper-Code. [Online].
[2] 2023. https://github.com/sainyam/Hyper-Code/blob/main/reproducibility/readme.pdf. [Online].
[3] Sainyam Galhotra, Amir Gilad, Sudeepa Roy, and Babak Salimi. 2022. HypeR: Hypothetical Reasoning With What-If and How-To Queries Using a Probabilistic Causal Approach. In *Proceedings of the 2022 International Conference on Management of Data* (Philadelphia, PA, USA) *(SIGMOD '22)*. Association for Computing Machinery, New York, NY, USA, 1598–1611. https://doi.org/10.1145/3514221.3526149