Reproducibility Report for ACM SIGMOD 2020 Paper: "Theoretically-Efficient and Practical Parallel DBSCAN"

HUANCHEN ZHANG, Tsinghua University, China

The main results of the paper can be reproduced and verified fairly easily based on the source code and scripts provided by the authors.

1 INTRODUCTION

The paper titled "Theoretically-Efficient and Practical Parallel DBSCAN"[1] was written by Yiqiu Wang, Yan Gu, and Julian Shun from MIT and UC Riverside. We are able to use the materials (i.e., the source code, experiment scripts, and datasets) provided to run the experiments on our local server. The reproduced results verifies the main figures in the paper.

2 SUBMISSION

The reproducibility submission includes the following:

- Source code Repository: `https://github.com/wangyiqiu/dbscan/tree/master/reproducibility`
- data generators: `https://www.dropbox.com/s/q9kn3wt4mv4nek1/uniformGen.py?dl=0`
- data sources: `https://www.dropbox.com/sh/ehhv9thpuvb36jq/AADQowvv9FfQ8ZYdAPL9qJs1a?dl=0`

The authors also provided detailed guidelines on how to run the scripts, including environmental dependencies. The scripts run without issues. A few datasets and data generators, however, are missing in order to reproduce all the figures in the paper (see the "Results" section)

3 HARDWARE AND SOFTWARE ENVIRONMENT

Table 1. Hardware & Software environment

	Paper	Repro Review
CPU	Intel Xeon Platinum 8124M	Intel(R) Xeon(R) Gold 6254
cores	36	18
GHz	3	1.2
RAM	144 GB	745 GB

4 REPRODUCIBILITY EVALUATION

4.1 Process

We got the scripts to run fairly easily following the the tutorials provided by the authors. The implementation uses CilkPlus which is no longer supported by the latest version of g++. We noticed that the code might also run with OpenMP, but there is no tutorial on that from the authors.

4.2 Results

We reran the experiments on the four main algorithms in the paper: EXACT, EXACT-qt, APPROX, APPROX-qt, all with bucketing. We were able to reproduce the main results. For Table 2, Figure 10(c), and Figure 10(g) in the paper, we couldn't verify the results because of missing datasets. Also we did not find test scripts for the comparison algorithms (hpdbscan, pdsdbscan) in the submission. We include some of the reproduced results in fig. 1 using the provided dataset "3D_VisualVar_10M".

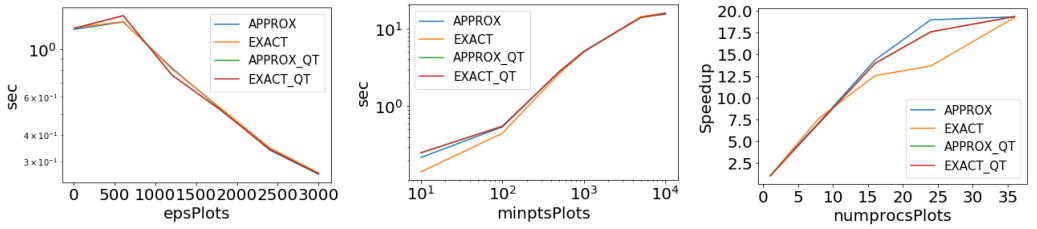

Fig. 1. Reproduced Results – Corresponding to Figure 5(b), 6(b), and 7(b) in the original paper

5 SUMMARY

The overall quality of the submission if good. The experiments are easy to run, and the main results can be reproduced.

REFERENCES

[1] Yiqiu Wang, Yan Gu, and Julian Shun. 2020. Theoretically-Efficient and Practical Parallel DBSCAN. In Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data (SIGMOD '20). Association for Computing Machinery, New York, NY, USA, 2555–2571.