Reproducibility Report for ACM SIGMOD 2021 Paper: "Parallel Index-Based Structural Graph Clustering and Its Approximation"

HERODOTOS HERODOTOU, Cyprus University of Technology, Cyprus

The reproducibility process was fairly simple as the authors provided three main scripts to execute, even though some errors were raised due to OS and compiler version discrepancies. After contacting the authors for support and manually intervening to resolve the issues, we managed to reproduce the results and compare them against the ones presented in the paper. Even though the reproduced results were not identical (probably due to differences in the hardware), the trends in the results were very similar to the ones in the paper. Overall, we conclude that the authors' findings in the paper match our reproduced results.

1 INTRODUCTION

This reproducibility review concerns the paper "Parallel Index-Based Structural Graph Clustering and Its Approximation" published in ACM SIGMOD 2021 and written by Tom Tseng, Laxman Dhulipala, and Julian Shun, members of the Computer Science & Artificial Intelligence Laboratory of MIT [1]. To reproduce the results as presented in the paper, we needed a high-performance server to execute the experiments, which involved the processing of complex and large-scale datasets. The reproduction process was fairly simple since the authors provided three scripts and a detailed description on how to reproduce the experiments. Their documentation included the steps needed to (i) download and install the required libraries, (ii) download and prepare the involved datasets, and (iii) execute all but one of the experiments that were conducted and generate the results in text format. One experimental comparison was not reproduced because the code was obtained via personal correspondence between the paper's authors and the original code's authors, and is not available publicly.

2 SUBMISSION

The reproducibility process of the paper concerned, consisted of the following steps:

- Clone the GitHub repository https://github.com/tomtseng/sigmod-reproducibility-parallel-scan, which includes a detailed README.md file with hardware information, software prerequisites, and execution instructions.
- Execute the `prepareSoftware.sh` script to download the required packages/libraries used to generate the results from the experiments (CMake, numactl, python3.6, sklearn, and MKL).
- Execute the `prepareData.sh` script to download and prepare the datasets required by the experiments (Orkut, Friendster, brain, WebBase, blood vessel, and cochlea).
- Execute the `runExperiments.sh` script to run the experiments and generate the results.
- The experiment results were generated as text files that have a csv file format. No scripts were provided for generating plots and thus we needed to manually check and compare the reproducibility results against the plots presented in the paper.

3 HARDWARE AND SOFTWARE ENVIRONMENT

Information regarding the environment used in the paper and in the reproducibility process are shown in Table 1, where both hardware and software information is available.

Table 1. Hardware & software environment

	Paper	Repro Review
CPU	Intel(R) Xeon(R) Platinum 8275CL	Intel(R) Xeon(R) Silver 4214Y
Cores	48	48
GHz	3.00GHz	2.2 GHz
RAM	412GB	125GB
Storage	200GB	1.7TB
OS	Ubuntu 18.04, 64-bit, x86	CentOS Linux release 7.6.1810 (Core)
g++	7.5.0	8.3.1
Python	3.6	3.6
Bazel	4	4

4 REPRODUCIBILITY EVALUATION

4.1 Process

The process of reproducing the results through the experiments was fairly simple since the authors did a good job of organizing the setup. However, we have faced some issues that we had to manually resolve, mainly due to operating system discrepancies. The authors generated the results on Ubuntu while the server used to reproduce the experiments is running CentOS. In the script responsible for downloading the required libraries and packages, the authors included the command for downloading Bazel on Ubuntu. This command failed on our server since we have a different operating system and had to resolve this manually by executing the corresponding commands for CentOS.

While trying to download and prepare the datasets, one of the dataset sources was unavailable causing us to delay the reproducibility process since it was the only source for downloading that specific dataset. A few days later, the source became available and we managed to download the dataset. In addition, one of the datasets was available only through `http` instead of `https`, causing the `wget` command in the script provided by the authors to fail. We resolved that manually by adding the `--no-check-certificate` option to the command, which enabled us to download the dataset successfully.

The OS discrepancies raised an issue with the g++ compiler version, since the default on Ubuntu has version v7.5.0. On CentOS, however, the default is v4.8.0 and it was not compatible with the author's code. Hence, we downloaded version v8.3.1, which was the next available version closest to the one the authors used. Then, we also faced a problem with CMake, where we had to upgrade it to version v3.8.2, since the default on our server was version v2.8.12.2, in order to build the source code and run the experiments.

We had contacted the authors of the paper a few times through emails to resolve some of the issues, who always immediately responded back to us to provide technical support or suggestions on resolving the issue raised.

We have managed to reproduce the results successfully, but it was relatively difficult for us to compare the results and evaluate them since the generated results were saved in text files in csv file format. It would have been better if the results were generated graphically in order to better compare them visually and notice the trends more clearly.

4.2 Results

All the major findings of the paper and reproduced results concern 6 figures in total, namely Figures 5–10 in the original paper. While most values in the reproduced results are different compared to the values presented in the paper through the figures, they follow the same trends as discussed in the paper.

According to Figure 5, the authors claim that the parallel GBBSIndexSCAN method with 48 cores achieves a relative speedup of 23–70× compared to the serial GBBSIndexSCAN method with one core. This is applicable in the paper results and can be also verified from our reproduced results.

The trends in Figure 6 match exactly the trends observed in our reproduced results for all the datasets and methods concerned with $\mu = 5$ and varying ϵ.

Similarly, in Figure 7, where ϵ is fixed to 0.6 and μ is varying, the reproduced results match the results as presented in the paper. This applies to all methods and every dataset involved.

In Figure 8, the trends observed are almost identical with minor differences in the order of magnitude. The same points on both the approximate cosine similarity and approximate Jaccard similarity values, fall either below or above the exact cosine similarity of each dataset. Note that the results for the Friendster dataset (one out of the 6 datasets evaluated) are missing from the reproduced results for this graph.

Again, in Figure 9, the results for the Friendster dataset are also missing. In the figure, however, not only are the trends the same, but also the modularity values in the reproduced results are almost exactly the same as in the graphs presented in the paper.

In Figure 10, the reproduced results are sparse, meaning that not all values were generated through the experiments as provided by the authors. Since some values are missing, we cannot determine whether the reproduced results accurately match the experimental results presented in the paper.

5 SUMMARY

Overall, the key claims supported by the paper were faithfully reproduced in our experiments. The overall reproducibility process was fairly easy to process despite the few errors we encountered. The main downside was the lack of a script for generating the plots used in the paper, which would have facilitated our comparison.

REFERENCES

[1] Tom Tseng, Laxman Dhulipala, and Julian Shun. 2021. Parallel Index-based Structural Graph Clustering and its Approximation. In *Proceedings of the 2021 International Conference on Management of Data (SIGMOD)*. ACM, 1851–1864.