Reproducibility Report for ACM SIGMOD 2021 Paper: "Efficient Graph Summarization using Weighted LSH at Billion-Scale"

HARIDIMOS KONDYLAKIS, FORTH-ICS, Greece

The reproducibility run generated data and results corresponding to Figures 2–4 in the SIGMOD paper, which show similar behavior to the figures in the paper. This partially confirms the experimental results of the paper.

1 INTRODUCTION

This reproducibility report is for the following paper:

- Quinton Yong, Mahdi Hajiabadi, Venkatesh Srinivasan, Alex Thomo. Efficient Graph Summarization using Weighted LSH at Billion-Scale. SIGMOD Conference 2021: 2357-2365 [1].

The experiments in the paper consist of three main parts: a) LDME v.s SWeG, b) Results of tuning k, c) LDME vs. Mosso and VoG, d) Distributed Environment Experiments, and e) Stochastic Block Model Experiments.

- The reproducibility run recreated figures similar to Figures 2-4 for a) and b)
- The reproducibility run did not generate Figure 5 required for confirming the results for c) d) and e).

The recreated Figures 2-4 confirm that indeed the method proposed in the paper outperforms SWeG in a single machine experiment, however, it was not able to verify whether this claim holds also for Mosso and VOG, for distributed environments, and for the Stochastic Block Model.

2 SUBMISSION

The reproducibility submission contains information on source code, datasets, and experimentation in a one-page description. We describe each item in the following:

- Github repository with code and scripts: https://github.com/QuintonYong/LDME
- Detailed readme file: https://github.com/QuintonYong/LDME/blob/master/README.md
- Datasets: Datasets are available through the following link http://law.di.unimi.it/datasets.php which is reported in the readme file. The ones to be downloaded should be selected based on the paper and also a preprocessing step is required to make them usable. The steps required for preprocessing a dataset are also reported in the readme file.
- Hardware: The code is implemented using Java and can be used by any Unix or Windows environment with the necessary libraries. The authors omit the details of the hardware on which they run the single node experiments, they only report the hardware configuration for the distributed environment – however they offer no guidelines or code for the distributed environment.

3 HARDWARE AND SOFTWARE ENVIRONMENT

PAs already mentioned the authors omit the details of the hardware they used for their single machine experiment. Nevertheless, for the reproducibility review, the details of the hardware used are reported in Table 1.

Table 1. Hardware & Software environment

	Repro Review
CPU	Intel(R) Xeon(R) CPU E5-2630
cores	18
GHz	2.30
RAM	64GB
Storage	5x2TB SATA 7.2 RPM hard drives in RAID 0

4 REPRODUCIBILITY EVALUATION

4.1 Process

We followed the reproducibility submission to set-up and run the experiments. Unfortunately, the main process was not automated by any scripts and as such:

(1) We had to manually download the datasets and apply the preprocessing steps required.
(2) Then we had to generate the experiment scripts in order to reproduce the experiments
(3) Then we had to identify the correct parameters from the results and to collect and visualize them.

Three key lessons that should be learned for future reproducibility submissions are that

(1) The reproducibility submission should specify which figures in the paper are expected to be generated and why those are enough to support the key results?
(2) The reproducibility submission should include specifics on the hardware of the experiments in the paper run.
(3) The reproducibility submission should be accompanied by the necessary scripts automating dataset acquisition/preprocessing, the batch of experiments as well as for visualizing the results (or at least collecting the numbers automatically)

4.2 Results

Reproduced results: The reproducibility run recreated figures similar to Figures 2-4 for a) and b) confirming that indeed LDME outperforms SWeG in single machine experiments.

Not reproduced results: The reproducibility run did not generate Figure 5 required for confirming that indeed LDME outperforms Mosso and VoG, that it outperforms competitors in a distributed environment as well as in the stochastic block model experiments.

REFERENCES

[1] Quinton Yong, Mahdi Hajiabadi, Venkatesh Srinivasan, and Alex Thomo. 2021. Efficient Graph Summarization using Weighted LSH at Billion-Scale. In *SIGMOD '21: International Conference on Management of Data, Virtual Event, China, June 20-25, 2021*. ACM, 2357–2365. https://doi.org/10.1145/3448016.3457331