SSumM: Sparse Summarization of Massive Graphs (Software User Guide)

Kyuhan Lee (kyuhan.lee@kaist.ac.kr)

May 29, 2020

1 General Information

- Version: 1.0

2 Introduction

SSumM (**S**parse **Sum**marization of **M**assive Graphs) is a scalable and effective graph summarization algorithm that yields a sparse summary graph. Compared to its state-of-the-art competitors, SSumM has the following advantages:

- *Concise*: yields up to 11.2× smaller summary graphs with similar reconstruction error.

- *Accurate*: achieves up to 4.2× smaller reconstruction error with similarly concise outputs.

- *Scalable*: summarizes 26× larger graphs while exhibiting linear scalability.

Detailed information about SSumM is explained in the following paper:

- Kyuhan Lee*, Hyeonsoo Jo*, Jihoon Ko, Sungsu Lim, and Kijung Shin, "SSumM: Sparse Summarization of Massive Graphs", KDD 2020

3 Installation

- In order to compile all the tools, it requires OpenJDK 12 or later be installed in the system.

- For compilation (optional), type *./compile.sh*.

- For packaging (optional), type *./package.sh*.

- For demo (optional), type *make*.

4 Input File Format of SSumM

SSumM assumes that the input graph $G = (V, E)$ is undirected without self-loops. The format of an input file is as follows. Each line represents a single edge. Each edge $\{u, v\} \in E$ joins two distinct nodes $u \neq v \in V$, seperated by a tab. Each node $v \in V$ is assigned to an unique integer id. The sample file ‘toygraph.txt’ contains 5 nodes and 6 edges as follows:

E.g. toygraph.txt

```
1	3
1	5
2	4
3	4
3	5
4	5
```

5 Output File Format of SSumM

The output file contains information about subnodes (nodes in G) belonging to each supernode $s \in S$ of the output summary graph $\overline{G} = (S, P, \omega)$ and information about each superedge $p \in P$. The first integer on each line following the line “<Subnode of each supernode>” represents the id of the supernode, and the following integers separated by tabs represent the ids of the subnodes belonging to that supernode. Each line following the line “<Superedge info>” represents a single superedge. The three integers separated by tabs represent the id of the source supernode, the id of the destination supernode, and the weight of the superedge (i.e., the number of subedges belonging to the superedge). A sample summary graph with 2 supernodes and 1 superedges is stored in the output file ‘summary_toygraph.txt’ as follows:

E.g. summary_toygraph.txt

```
<Subnode of each supernode>
3	2	5	3
4	4	1
<Superedge info>
3	4	5
```

6 Running SSumM

6.1 How to Execute

```
./run.sh  input_path  compression_ratio  reconstruction_error
```

6.2 Parameters

- *input_path*: Path to the input text file in the format described above.
- *compression_ratio*: The desired size of a summary graph compared relative to the input graph size in bits. This parameter should be a real number between 0 and 1.
- *reconstruction_error*: The reconstruction error measure to be used. This parameter should be either 1 (RE_1) or 2 (RE_2).

7 Running SSumM on a Test Example

An example for obtaining a summary graph $\overline{G}$ whose size in bits is 20% of the input graph 'ego_facebook.txt'. RE_1 is used as the reconstruction error measure.

1. Run SSumM with compression_ratio = 0.20, optimizing quality with RE_1

```
./run.sh  ego_facebook.txt  0.20  1
```

2. The output after running SSumM

```
------------------------------------------
Data Read Start: 2020-02-05 06:39:13
|V|: 4039
|E|: 88234
Finished reading the input graph
Start Time: 2020-02-05 06:39:14
iter: 1
iter: 2
iter: 3
iter: 4
iter: 5
------------------------------------------
Elapsed Time: 1364.0 ms
Original size: 2114048.25 bits
Summary size: 422345.02 bits (19.978022%)
L1 error: 5.44e − 03
```