CoCoS: Fast and Accurate Distributed Triangle Counting in Graph Streams

1 General Information

- Version: 1.0
- Date: Oct-10-2019
- Authors: Kijung Shin (kijungs@kaist.ac.kr) and Jinoh Oh (jinoh.postech@gmail.com)

2 Introduction

CoCoS is a distributed streaming algorithm for global and local triangle counting in graph streams. **CoCoS** processes and samples edges across multiple machines to reduce redundancy in computation and storage.

CoCoS has the following advantages:

- *Accurate*: CoCoS produces up to 30X smaller estimation error than its competitors with similar speeds.
- *Fast*: CoCoS runs in linear time up to 10.4X faster while giving more accurate estimates.
- *Theoretically Sound*: CoCoS gives unbiased estimates.

Detailed information about the algorithm is explained in the following paper

- "*CoCoS: Fast and Accurate Distributed Triangle Counting in Graph Streams"*
  Kijung Shin, Euiwoong Lee, Jinoh Oh, Mohammad Hammoud, and Christos Faloutsos
  ACM Transactions on Knowledge Discovery from Data (TKDD)

3 Installation

- This package requires that *c++ 0x* (or higher) be installed in the system and set in PATH.
- This package requires that *MPICH 3.1* (or higher) be installed in the system and set in PATH.
- For compilation (optional), type *'make'*
- For demo (optional), type '*make demo'*

4 Input File Format

The input file lists edges in a graph. Each line corresponds to an edge and consists of the source node id and the destination node id, which are separated by a tab. Additionally, we assume the followings:

- No parallel edges. For example, both edge (1,2) and edge (2,1) cannot be in the input file at the same time.
- Node ids are integers in range [0, #nodes -1]

example_graph.txt is an example of the input file.

5 Output Files Format

Two output files are created for each trial.

- *global(trial#).txt*: this file has the estimated number of global triangles.
- *local(trial#).txt*: this file lists the estimated number of local triangles of each node. Each line consists of the node id and the number of its local triangle count, which are separated by a tab.

example_output directory contains the examples of the output files.

6 Running CoCoS (Simple)

6.1 Simulating CoCoS (simple) in a Single Machine

```
mpirun -n [#processes] ./bin/mpi --method simple --trial [#trials] --budget [budget] [input_graph] [output_directory]
```

- *#processes*: number of processes that will be used to run the algorithm. One of the processes runs the master and the aggregator. This parameter should be an integer greater than or equal to two.
- *#trials*: number of times that CoCoS (Simple) will be executed. This parameter should be an integer greater than or equal to one.
- *budget*: maximum number of edges that can be stored in each worker. This parameter should be an integer greater than two.
- *input_graph*: the path of an input graph file. See 4 for the detailed format of the input file.
- *output_directory:* the path of the directory where output files will be stored. See 5 for the detailed formats of the output files.

6.2 Running CoCoS (simple) in a Distributed Setting

```
mpiexec -n [#processes] –f [machinefile] ./bin/mpi --method simple --trial [#trials] --budget [budget] [input_graph] [output_directory]
```

- *#processes*: number of processes that will be used to run the algorithm. One of the processes runs the master and the aggregator. This parameter should be an integer greater than or equal to two.
- *machinefile*: the path of a machinefile. The host listed first runs the master and the aggregator; and the remaining hosts run workers. See 10 of mpich manual for the details of a machinefile.
- *#trials*: number of times that CoCoS (Simple) will be executed. This parameter should be an integer greater than or equal to one.
- *budget*: maximum number of edges that can be stored in each worker. This parameter should be an integer greater than two.
- *input_graph*: the path of an input graph file in the host machine of the master. See 4 for the detailed format of the input file.
- *output_directory:* the path of the directory where output files will be stored. The directory is located in the host machine of the master. See 5 for the detailed formats of the output files.

7 Running CoCoS (Opt)

7.1 Simulating CoCoS (opt) in a Single Machine

```
mpirun -n [#processes] ./bin/mpi --trial [#trials] --tolerance [tolerance] --budget [budget] [input_graph] [output_directory]
```

- *#processes*: number of processes that will be used to run the algorithm. One of the processes runs the master and the aggregator. This parameter should be an integer greater than or equal to two.
- *#trials*: number of times that CoCoS (Opt) will be executed. This parameter should be an integer greater than or equal to one.
- *budget*: maximum number of edges that can be stored in each worker. This parameter should be an integer greater than two.
- *tolerance*: tolerance for load difference. This parameter should be a real number greater than or equal to one. The default value is 0.2.
- *input_graph*: the path of an input graph file. See 4 for the detailed format of the input file.
- *output_directory:* the path of the directory where output files will be stored. See 5 for the detailed formats of the output files.

7.2 Running CoCoS (opt) in a Distributed Setting

```
mpiexec -n [#processes] –f [machinefile] ./bin/mpi --trial [#trials] --budget [budget] --tolerance [tolerance] [input_graph] [output_directory]
```

- *#processes*: number of processes that will be used to run the algorithm. One of the processes runs the master and the aggregator. This parameter should be an integer greater than or equal to two.
- *machinefile*: the path of a machinefile. The host listed first runs the master and the aggregator; and the remaining hosts run workers. See 10 of mpich manual for the details of a machinefile.
- *#trials*: number of times that CoCoS (Opt) will be executed. This parameter should be an integer greater than or equal to one.
- *budget*: maximum number of edges that can be stored in each worker. This parameter should be an integer greater than two.
- *tolerance*: tolerance for load difference. This parameter should be a real number greater than or equal to one. The default value is 0.2.
- *input_graph*: the path of an input graph file in the host machine of the master. See 4 for the detailed format of the input file.
- *output_directory:* the path of the directory where output files will be stored. The directory is located in the host machine of the master. See 5 for the detailed formats of the output files.

8 Running Tri-Fly

8.1 Simulating Tri-Fly in a Single Machine

```
mpirun -n [#processes] ./bin/mpi --method naive --trial [#trials] --budget [budget] [input_graph] [output_directory]
```

- *#processes*: number of processes that will be used to run the algorithm. One of the processes runs the master and the aggregator. This parameter should be an integer greater than or equal to two.
- *#trials*: number of times that Tri-Fly will be executed. This parameter should be an integer greater than or equal to one.
- *budget*: maximum number of edges that can be stored in each worker. This parameter should be an integer greater than two.
- *input_graph*: the path of an input graph file. See 4 for the detailed format of the input file.
- *output_directory:* the path of the directory where output files will be stored. See 5 for the detailed formats of the output files.

8.2 Running TRI-FLY in a Distributed Setting

```
mpiexec -n [#processes] –f [machinefile] ./bin/mpi --method naive --trial [#trials] --budget [budget] [input_graph] [output_directory]
```

- *#processes*: number of processes that will be used to run the algorithm. One of the processes runs the master and the aggregator. This parameter should be an integer greater than or equal to two.
- *machinefile*: the path of a machinefile. The host listed first runs the master and the aggregator; and the remaining hosts run workers. See 10 of mpich manual for the details of a machinefile.
- *#trials*: number of times that TRI-FLY will be executed. This parameter should be an integer greater than or equal to one.
- *budget*: maximum number of edges that can be stored in each worker. This parameter should be an integer greater than two.
- *input_graph*: the path of an input graph file in the host machine of the master. See 4 for the detailed format of the input file.
- *output_directory:* the path of the directory where output files will be stored. The directory is located in the host machine of the master. See 5 for the detailed formats of the output files.