**Report**

**Homework 3 Mining Data Stream**

Group member:

Wei Lu (weilu@kth.se)

Tianlong Du ([tianlong@kth.se)](mailto:tianlong@kth.se))

1. Selected Paper

M. Jha, C. Seshadhri, and A. Pinar, [A Space-Efficient Streaming Algorithm for Estimating Transitivity and Triangle Counts Using the Birthday Paradox](https://arxiv.org/pdf/1212.2264.pdf), ACM TKDD, 9-3, 2015.

1. Introduction
2. Implement the reservoir sampling
3. Implement the streaming graph algorithm
4. Result
5. The dataset we use: Facebook (NIPS)
6. Test result by using the dataset

2-1 4-1  
2-1 6-1  
3-1 4-1  
3-1 5-1  
3-1 7-1  
3-1 9-1  
4-3 5-3  
4-1 5-1  
4-1 8-1  
4-3 35346-3  
4-3 140284-3  
4-2 155439-2  
5-1 8-1  
5-4 35346-4  
5-3 64326-3  
5-3 94142-3  
6-4 155439-4  
12-11 13-11  
12-10 14-10  
12-10 16-10  
13-12 14-12  
13-12 17-12  
13-11 17-11  
17-13 19-13  
17-16 19-16  
19-17 276469-17  
29-27 30-27  
30-28 278180-28  
30-29 278181-29  
31-30 278180-30  
53-51 191285-51  
53-51 191286-51  
53-51 191286-51  
53-51 191289-51

triangle estimate 147001404.639  
[5]

1. Bonus
2. What were the challenges you have faced when implementing the algorithm?

The most challenges for use is to understand the whole process of the algorithm.

1. Can the algorithm be easily parallelized? If yes, how? If not, why? Explain.

The algorithm can be parallelized because only minority of the new edges will be inserted to the array of reservoir edges, so we can use parallelized algorithm to filter the edges which can not be inserted to the array of reservoir edges but only one thread can update the array of reservoir at a certain time.

1. Does the algorithm work for unbounded graph streams? Explain.

Yes. Because in this algorithm, we do not need to care about the total length of the graph streams, we just need to keep the sample with the fixed size and the sample contains each element seen so far with probability s/n

1. Does the algorithm support edge deletions? If not, what modification would it need? Explain.

No, our algorithm does not support edge deletions. If the algorithm would support edge deletions, it could record the previous edges which were replaced by the new edges and these previous edges can replace those deleted edges so that the sample can keep the fixed size all the time.

1. Instruction for running
2. Go to command line
3. Run: python lab3.py