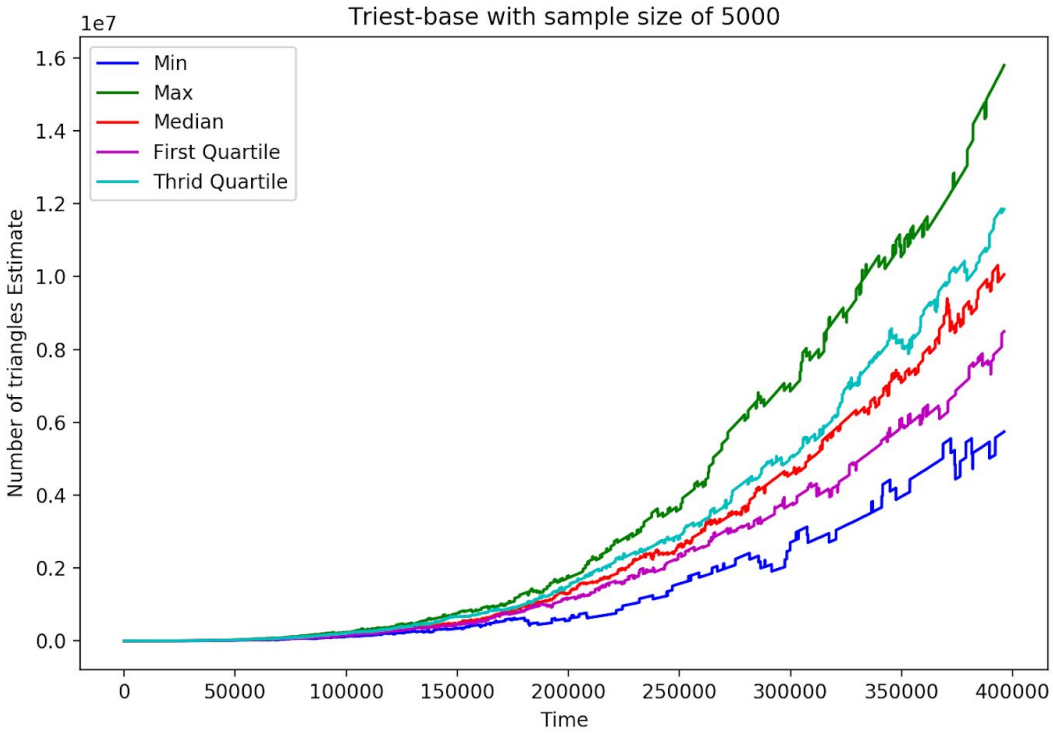
Triangle Estimation Evaluation Report

Data pre-processing: discard “self-edges”(edges from node u to node u) and repeated edges(edges from node u to node v when another edge is already present between u and v)

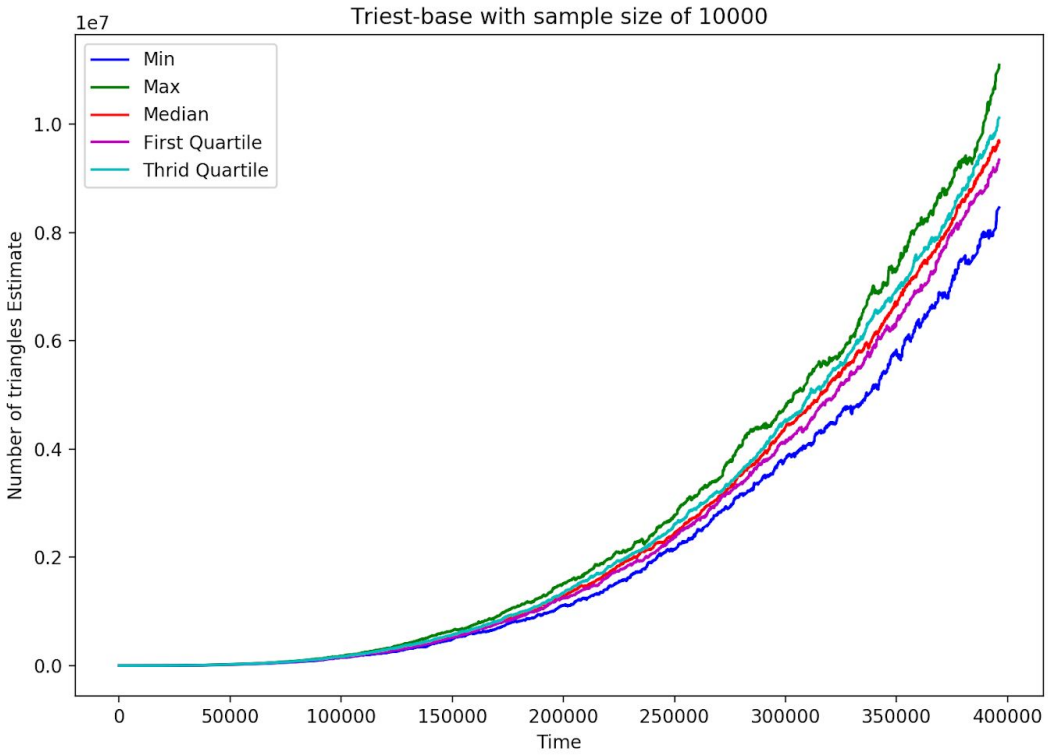
Sample size(M) = 5,000, 10,000, 20,000, 30,000, 40,000

Triest-base

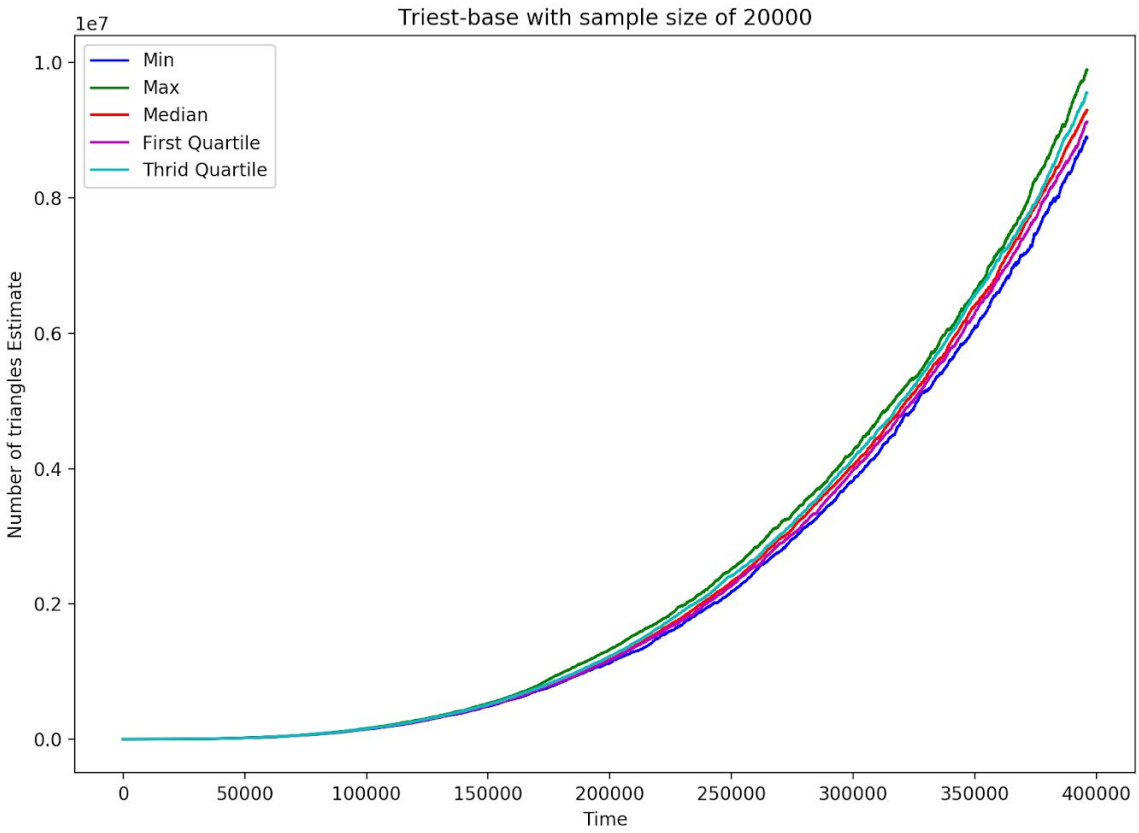
M = 5,000



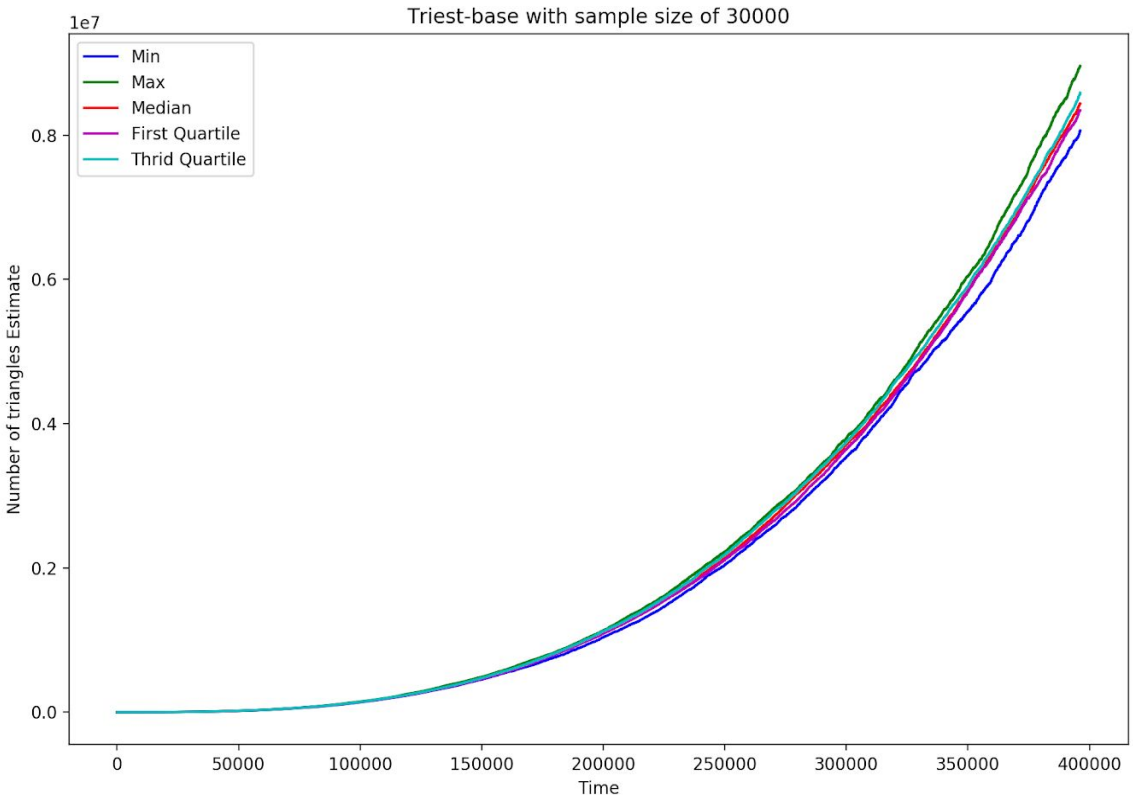
M = 10,000


Triest-base with sample size of 10000
1e7
Min
Max
Median
First Quartile
Thrid Quartile
Number of triangles Estimate
1.0
0.8
0.6
0.4
0.2
0.0
0
50000
100000
150000
200000
250000
300000
350000
400000
Time

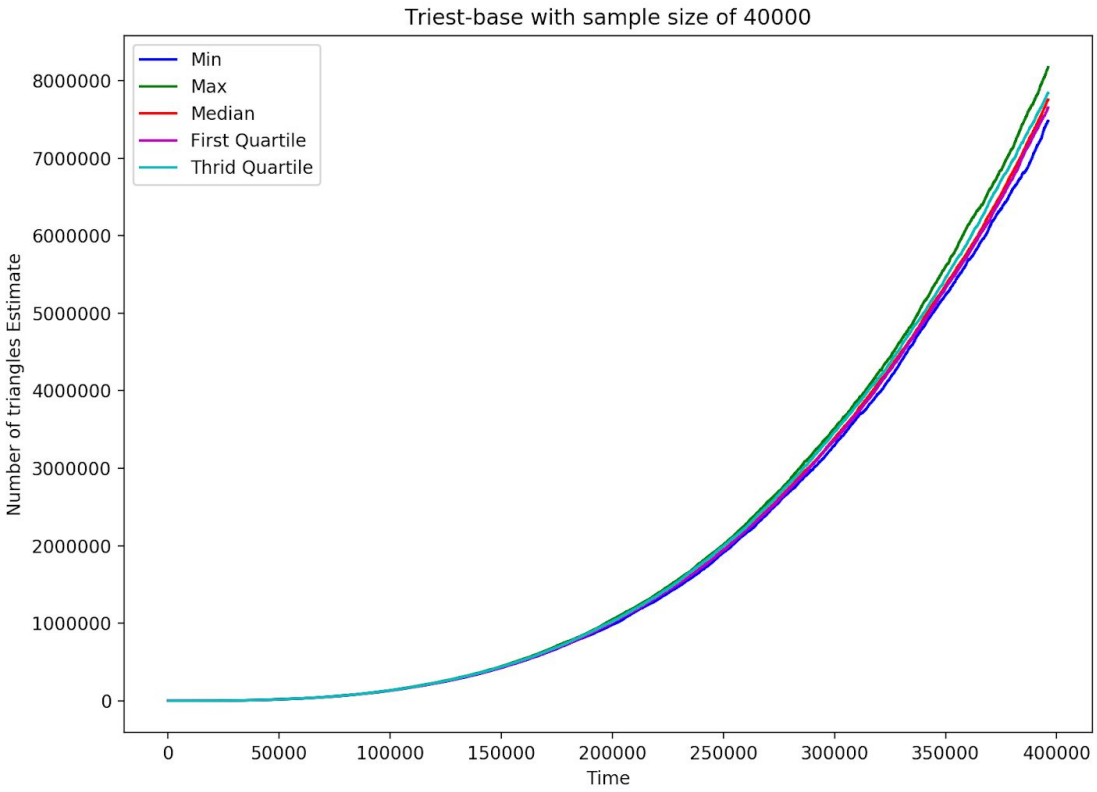
M = 20,000


Triest-base with sample size of 20000
1e7
Min
Max
Median
First Quartile
Thrid Quartile
Number of triangles Estimate
Time
0.0
0.2
0.4
0.6
0.8
1.0
0
50000
100000
150000
200000
250000
300000
350000
400000

M = 30,000


Triest-base with sample size of 30000
1e7
Min
Max
Median
First Quartile
Thrid Quartile
Number of triangles Estimate
0.0
0.2
0.4
0.6
0.8
0
50000
100000
150000
200000
250000
300000
350000
400000
Time

M = 40,000

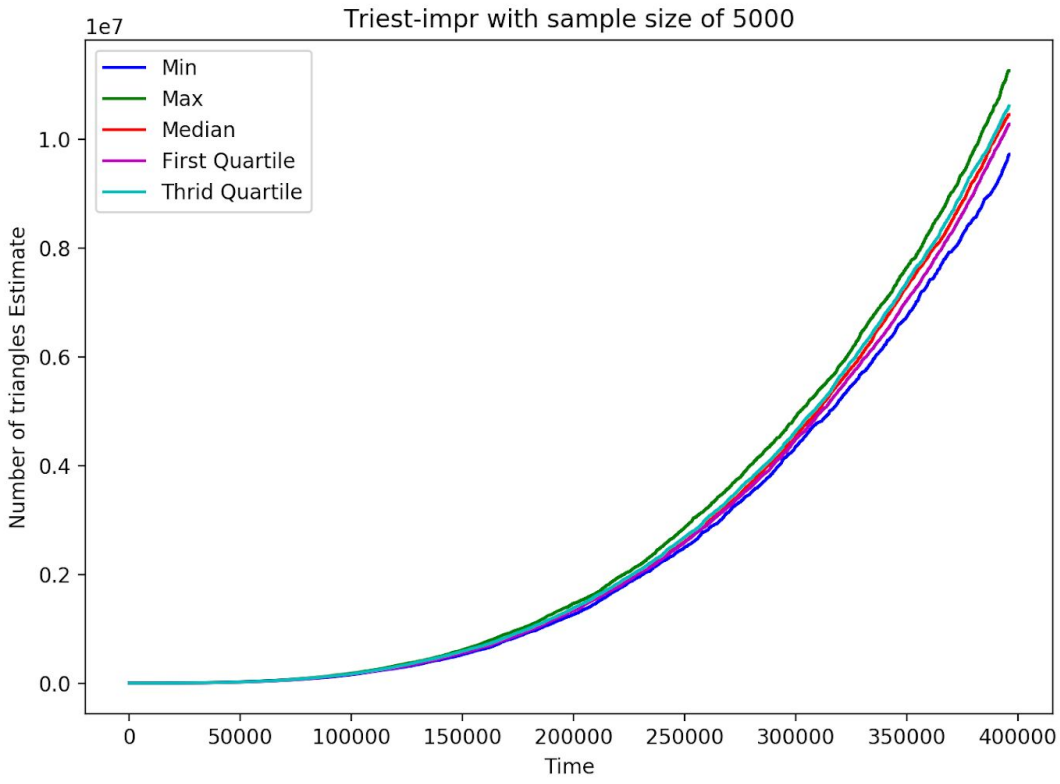


For Triest-base algorithm, as sample size increases, max, median, and the third quartiles estimations of triangle count all tend to decrease. The max estimation decreases the most, from 1.6e7 at time near 400000 in the first graph of sample size 5000 to 8000000 at the same time in the last graph of sample size 40000. The min and the first quartiles estimations tend to increase at first and then decrease as sample size increases. The larger the sample size is, the more “stable” the estimations are in general. As shown in the graph of sample size 5000, the difference between max estimations and min estimations from 20 runs at some time ts is large compared with the difference

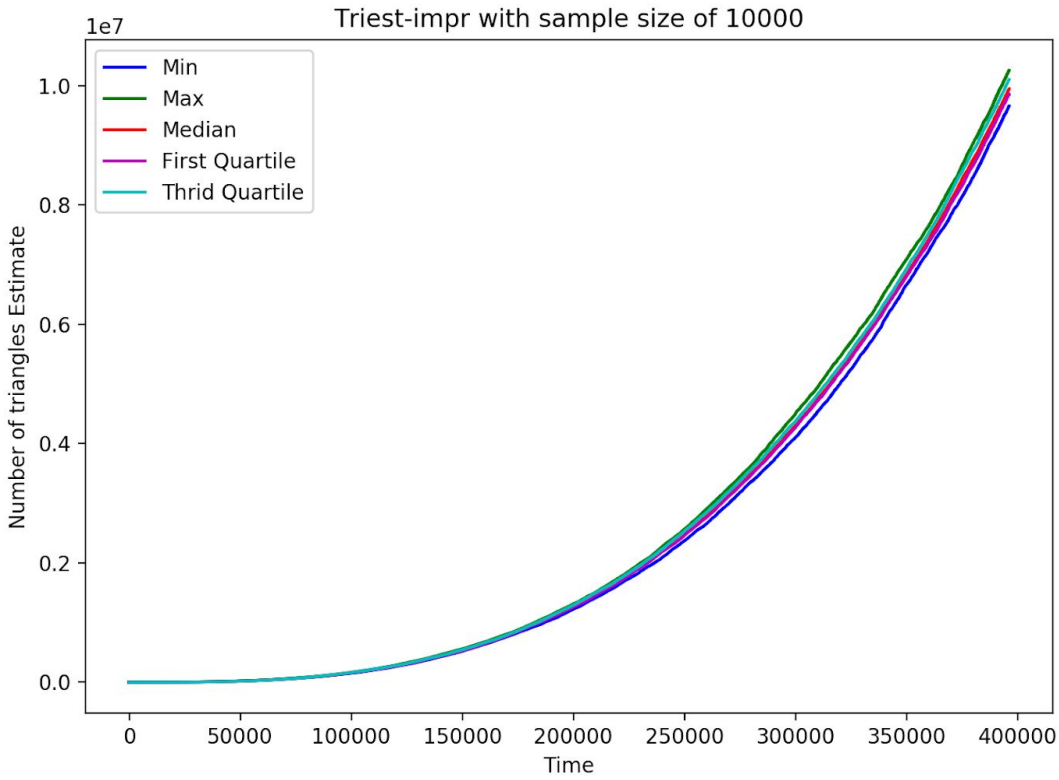
between max and min in graph of other sample sizes. Though triest-base estimate is an unbiased estimate, the variance/standard deviation of it may be large(the results from different runs vary much) when the sample size is small since the lines of min, max, median, the first and the third quartiles are relatively separable. As sample size increases, the lines of min, max, median, the first and the third quartiles become more and more close to each other.

Triest-impr

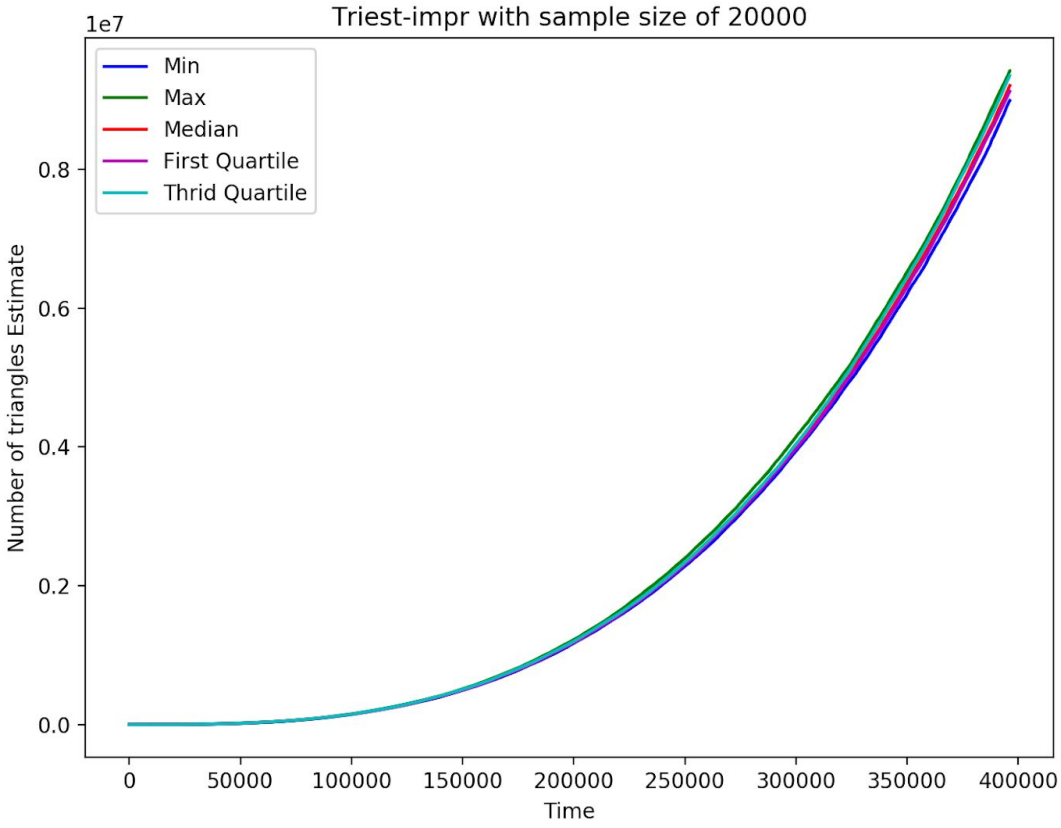
M = 5,000



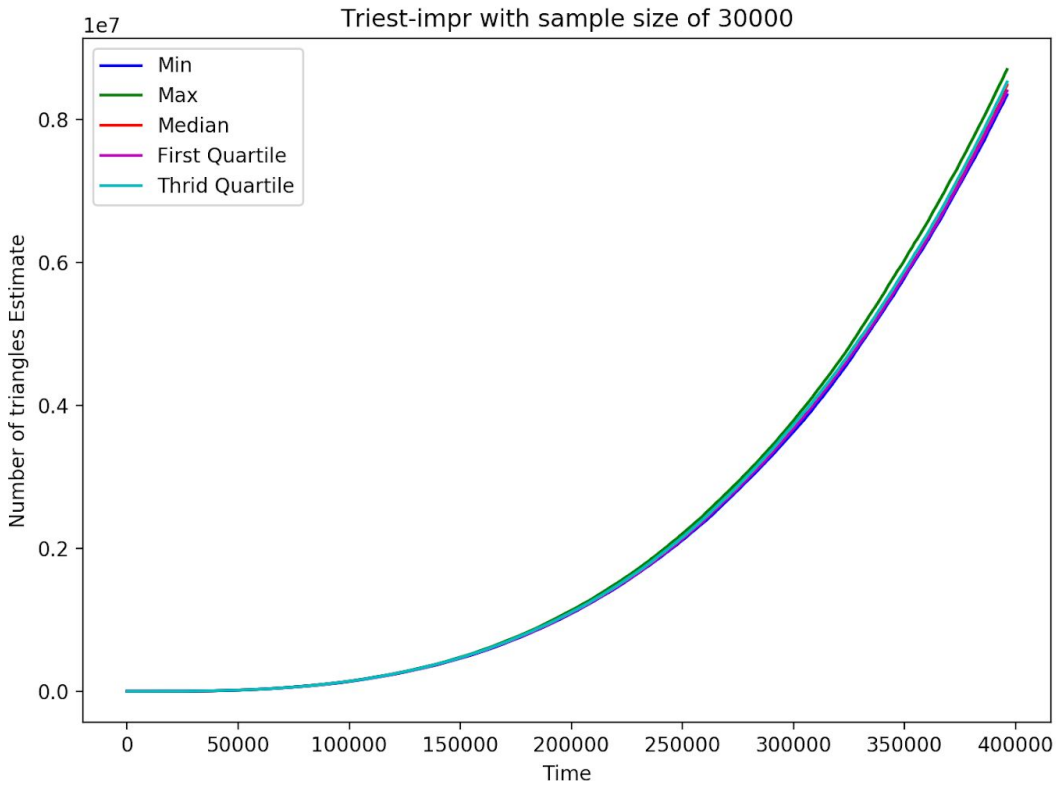
M = 10,000


Triest-impr with sample size of 10000
1e7
Min
Max
Median
First Quartile
Thrid Quartile
Number of triangles Estimate
1.0
0.8
0.6
0.4
0.2
0.0
0
50000
100000
150000
200000
250000
300000
350000
400000
Time

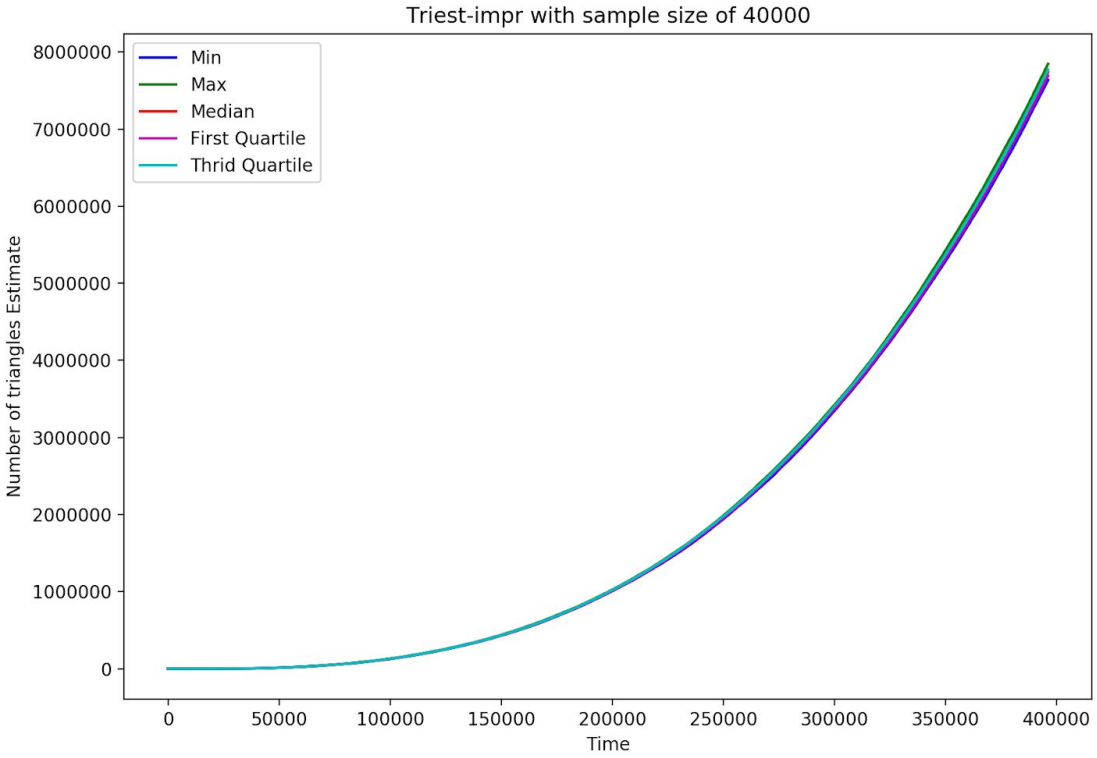
M = 20,000


Triest-impr with sample size of 20000
1e7
Min
Max
Median
First Quartile
Thrid Quartile
Number of triangles Estimate
0.0
0.2
0.4
0.6
0.8
0
50000
100000
150000
200000
250000
300000
350000
400000
Time

M = 30,000


Triest-impr with sample size of 30000
1e7
Min
Max
Median
First Quartile
Thrid Quartile
Number of triangles Estimate
0.0
0.2
0.4
0.6
0.8
0
50000
100000
150000
200000
250000
300000
350000
400000
Time

M = 40,000



For Triest-impr algorithm, as sample size increases, max, min, median, the first and the third quartiles estimations all tend to decrease(tend to converge to the actual number of triangles in the dataset).

Difference between Triest-base and Triest-impr

First of all, In Triest-base algorithm graphs, it's possible that for some t'>= t, the estimations on t' are smaller than estimations on t(a decrease on estimations may happen as time increases) which the actual number of triangles on t' >= the actual number of triangles on t. However, Triest-impr algorithm never decreases the estimate and only uses G_{S} to identify new triangles.

Second, Triest-base only counts a triangle if all three edges are in S(memory). But if two edges are in S, and the third one is on the stream right now, we may infer that the triangle exists, so we should count it. Triest-impr algorithms solves this problem by first incrementing the counter D and then deciding whether to insert the edge into S.
Lastly, the Triest-impr algorithm seems to be more reliable than Triest-base algorithm since the variability of estimates in Triest-impr is smaller, as suggested by the graphs and mentioned in previous paragraphs.Triest-impr outperformed Triest-base especially when sample size is very small such as 5000 in this example since different runs in Triest-base tend to generate estimates that differ much.