Supplementary Material for submission 865

1 Proofs of Lemmas

Lemma 3.2: If applying $p_{uv}^{(s)} = (1-\alpha)^{r_{uv}}$, in any computational round r, all the arrived edges stay in the reservoir with probability $(1-\alpha)^r$.

Proof. Given any edge (u, v), suppose that it arrives in round r_{uv} and $p_{uv}^{(s)} = (1-\alpha)^{r_{uv}}$. In round r $(\geq r_{uv})$, the probability of (u, v) in the reservoir is

$$(1-\alpha)^{r_{uv}} \cdot (1-\alpha)^{r-r_{uv}} = (1-\alpha)^r.$$

□

Lemma 3.3: Let $p_{uv}^{(s)} = k/t_{uv}$, in any computational round r, t be the number of arrived edges, and $p_{uv}^{(r)}$ be the probability that any arrived edge (u, v) staying in the reservoir. If $\alpha \leq 0.7$,

$$\left| p_{uv}^{(r)} - p^* \right| / p^* \leq 1 - \exp(-2\alpha), \quad p^* = k/t. \tag{1}$$

Proof. Suppose that edge t comes in round R_0 $(R_0 \geq 1)$, the number of edges coming following t is Y, and the edge n $(n = t + Y)$ is in round R $(R > R_0)$. Suppose that after edge t, it needs d_0 edges to fulfill the reservoir. Let the number of edges arrive before n in round R is d. The probability of t staying in the reservoir at time n is $p^* = \frac{k}{n}$ in traditional reservoir sampling, and the probability of t staying in the reservoir at time $Y + t$ in GRS using P-II is $\frac{k}{t} \cdot (1-\alpha)^{R-R_0}$.

Firstly, when $R > R_0$, from *Lemma 3.6*, we have

$$\begin{aligned} Y &= d_0 + x_{R_0+1} + x_{R_0+2} \cdots + x_{R-1} + d \\ &= k \cdot (\exp(\alpha) - 1) \cdot (\exp(\alpha \cdot R_0)) \cdot \frac{(\exp(\alpha \cdot (R - R_0 - 1)) - 1}{\exp(\alpha) - 1} + (d + d_0) \\ &= k \cdot (\exp(\alpha \cdot (R-1)) - \exp(\alpha \cdot R_0)) + (d + d_0). \end{aligned} \tag{2}$$

Similarly, we have

$$\begin{aligned} t &= k + x_1 + x_2 \cdots + x_{R_0-1} + x_{R_0} - d \\ &= k \cdot (\exp(\alpha) - 1) \cdot \frac{\exp(\alpha \cdot R_0) - 1}{\exp(\alpha) - 1} + (k - d) \\ &= k \cdot (\exp(\alpha \cdot R_0) - 1) + (k - d) \\ &= k \cdot \exp(\alpha \cdot R_0) - d \\ &\leq k \cdot \exp(\alpha \cdot R_0). \end{aligned} \tag{3}$$

Combine Equations 2 and 3, we have

$$\begin{aligned} n = Y + t &= k \cdot (\exp(\alpha \cdot (R-1)) - \exp(\alpha \cdot R_0)) + (d + d_0) + k \cdot (\exp(\alpha \cdot R_0) - 1) + (k - d) \\ &= k \cdot (\exp(\alpha \cdot (R-1)) - 1) + (d_0 + k) \\ &= k \cdot \exp(\alpha \cdot (R-1)) + d_0 \\ &\geq k \cdot \exp(\alpha \cdot (R-1)). \end{aligned} \tag{4}$$

At last, combine all the analysis above, we have

$$
\begin{aligned}
p_{uv}^{(r)} - p^* &= \frac{k}{n} - \frac{k}{t} \cdot (1-\alpha)^{R-R_0} \\
&\leq \frac{k}{n} - \frac{k}{k \cdot \exp(\alpha \cdot R_0)} \cdot (1-\alpha)^{R-R_0} && (5) \\
&\leq \frac{k}{n} - \exp(-\alpha \cdot R_0) \cdot (\exp(-\alpha(R - R_0 + 1))) && (6) \\
&= \frac{k}{n} - \exp(-\alpha \cdot (R+1)) \\
&= \frac{k}{n} - \exp(-\alpha \cdot (R-1)) \cdot \exp(-2\alpha) \\
&\leq \frac{k}{n} - \frac{k}{n} \cdot \exp(-2\alpha) && (7) \\
&= (1 - \exp(-2\alpha)) \cdot \frac{k}{n}.
\end{aligned}
$$

Equation 5 is hold because of Equation 3. Equation 7 is hold because of Equation 4. Equation 6 is hold only when $\alpha < 0.7$, we have

$$\exp(-\alpha(R - R_0 + 1)) \leq (1-\alpha)^{R-R_0}, \tag{8}$$

which we prove as follows. In order to investigate the difference between $\exp(-\alpha(R - R_0 + 1))$ and $(1-\alpha)^{R-R_0}$, we define $Y = R - R_0$ and a function

$$f(Y) = \exp(-\alpha \cdot (Y+1)) - (1-\alpha)^Y. \tag{9}$$

Then, we calculate the derivation of function $f(Y)$

$$f'_Y = -\alpha \cdot \exp(-\alpha \cdot (Y+1)) - \ln(1-\alpha) \cdot (1-\alpha)^Y.$$

Because $Y \geq 1$, $f'_Y < 0$, which means that $f(Y)$ is a decreasing function. When $Y = 1$, we have the maximum of $f(Y)$ is $f_{max} = \exp(-2\alpha) - 1 + \alpha$. Define a function

$$g(\alpha) = \exp(-2\alpha) - 1 + \alpha,\ \alpha \in \left[\frac{1}{k}, 1\right].$$

Then, we calculate the derivation of $g(\alpha)$ function

$$g'(\alpha) = -2\exp(-2\alpha) + 1,$$

which is an increasing function with zero point, so that $g(\alpha)$ is a function that first decreases and then increases when $\alpha = -\frac{1}{2}\ln\frac{1}{2}$, $g_\alpha = g_{min}$, $g_{min} < 0$. Therefore, $g(\alpha)$ has zero points. It is easy to know $g(0) = 0$. Then, using Bisection method, we have $g(0.7) \cdot g(0.8) < 0$. Therefore, $g(\alpha) < 0$ when $0 < \alpha \leq 0.7$. Moreover, $f_{max} < 0$, when $0 < \alpha \leq 0.7$, so that $(\exp(-\alpha \cdot (Y+1))) < (1-\alpha)^Y$. Therefore, we have

$$p_{uv}^{(r)} - p^* \leq (1 - \exp(-2\alpha)) \cdot \frac{k}{n} \Longrightarrow \left|p_{uv}^{(r)} - p^*\right| / p^* \leq 1 - \exp(-2\alpha).$$

□

Lemma 3.4: In algorithm GREAT^I, the expected number of edges arrived in computational round r is $x_r^I = \alpha \cdot k/(1-\alpha)^r$.

Proof. The sampling probability in round r is $p_r = (1-\alpha)^r$, then we have

$$x_r \cdot (1-\alpha)^r = k \cdot \alpha \Longrightarrow x_r^I = \alpha \cdot k/(1-\alpha)^r.$$

□

Lemma 3.5: In algorithm GREAT^I, at the beginning of computational round r, suppose that there are Q free slots in the reservoir where each slot can store one edge. The expected number of edges arrived to put one sampled edge in the i^{th} slot is $y_{r,i}^I = 1/(1-\alpha)^r$.

Proof. The sampling probability in round r is $p_r = (1-\alpha)^r$, then we have

$$y_{r,i} \cdot (1-\alpha)^r = 1 \Longrightarrow y^I_{r,i} = 1/(1-\alpha)^r.$$

$\square$

Lemma 3.6: In algorithm GREAT^{II}, the expected number of edges arrived in computational round r is $x^{II}_r = (\exp(\alpha)-1)\cdot\exp((r-1)\cdot\alpha)\cdot k$.

Proof. In algorithm GREAT^{II}, in round 0, the reservoir is empty and k edges are sampled with probability 1. When the reservoir is full, edges are randomly removed with probability α. There will be $Q(Q = k\cdot\alpha)$ empty slots. In round 1, edges are sampled with probability $\frac{k}{t}$ and t starts at $k+1$. Assume that filling these Q empty slots requires x_1 edges, then we have

$$\begin{aligned}\frac{k}{k+1}+\frac{k}{k+2}\cdots+\frac{k}{k+x_1} &= Q\\ \frac{1}{k+1}+\frac{1}{k+2}\cdots+\frac{1}{k+x_1} &= \frac{Q}{k}=\alpha.\end{aligned}$$

If we have

$$\frac{1}{k+1}+\frac{1}{k+2}+\ldots+\frac{1}{k+x_1}\approx\int_k^{k+x_1}\frac{1}{u}du, \tag{10}$$

then we can solve

$$\begin{aligned}&\int_k^{k+x_1}\frac{1}{u}du=\alpha\\ \Longrightarrow&\ln\left(\frac{k+x_1}{k}\right)=\alpha\\ \Longrightarrow&x_1=k\cdot\exp(\alpha)-k\\ &x_1=k\cdot(\exp(\alpha)-1).\end{aligned}$$

The approximation of Equation 10 will be proved later.

Similarly, in round 2, t starts at $k+x_1+1$, and it needs x_2 edges to fulfill these s empty slots, so we have

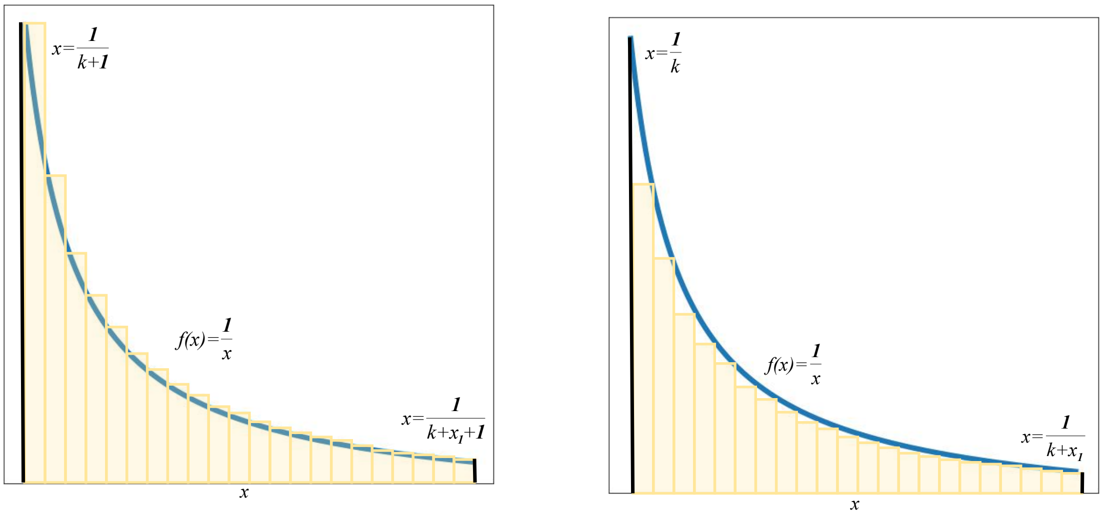
$$\begin{aligned}&\frac{k}{k+x_1+1}+\frac{k}{k+x_1+2}+\ldots+\frac{k}{k+x_1+x_2}=s\\ \Longrightarrow&\int_{k+x_1}^{k+x_1+x_2}\frac{1}{u}du=\alpha\\ \Longrightarrow&\ x_2=(\exp(\alpha)-1)\cdot(k+x_1).\end{aligned}$$

Repeat the above process, and after r rounds, we have

$$\begin{aligned}x_0&=k\\ x_1&=(\exp(\alpha)-1)\cdot k\\ x_2&=(\exp(\alpha)-1)\cdot(k+x_1)\\ x_3&=(\exp(\alpha)-1)\cdot(k+x_1+x_2)\\ &\ldots\\ x_r&=(\exp(\alpha)-1)\cdot(k+x_1+x_2+\ldots+x_{r-1}).\end{aligned}$$

Then,

$$\begin{aligned}x_{r-1}&=(\exp(\alpha)-1)\cdot(k+x_1+x_2+\ldots+x_{r-2})\\ x_r&=(\exp(\alpha)-1)\cdot(k+x_1+x_2+\ldots+x_{r-2}+x_{r-1})\\ x_r-x_{r-1}&=(\exp(\alpha)-1)\cdot x_{r-1}\\ x_r&=\exp(\alpha)\cdot x_{r-1}.\end{aligned}$$



(a) Max.

(b) Max.

Figure 1: The approximation of definite integral.

Therefore, $\{x_r^{II}\}_{r=1}$ is a geometric progression. The common ratio is $\exp(\alpha)$ and the first term is $x_1^{II} = k(\exp(\alpha) - 1)$, and the general term is

$$x_r^{II} = (\exp(\alpha) - 1) \cdot \exp(\alpha \cdot (r-1)) \cdot k.$$

Then, we give the bound of approximated Equation 10. Because $\frac{1}{k+1} + \frac{1}{k+2} + \ldots + \frac{1}{k+x_1}$ is an approximation of the area of the trapezoid with curved edge, which is constructed by the line of function $f(x) = \frac{1}{x}$, line $x = \frac{1}{k+1}$, line $x = \frac{1}{k+x_1+1}$, and the x-axis. This curved-edge trapezoid can be partitioned into x_1 small rectangles of width 1. The area of the curved-edge trapezoid is approximated by the sum of the rectangular areas and choose the left intersect point of the rectangles and the curved-edge trapezoid as the height, as shown in Figure 1a. Since $f(x) = \frac{1}{x}$ is a decreasing function, the sum of the rectangular areas is greater than the actual area of the curved edge trapezoid. Similarly, $\frac{1}{k+1} + \frac{1}{k+2} + \ldots + \frac{1}{k+1+x_1}$ is also the approximation of the curved-edge trapezium surrounded by the line of function $f(x) = \frac{1}{x}$, line $x = 1$, line $x = \frac{1}{k+x_1}$, and the x-axis. The difference is to choose the right intersect point of the rectangles and the curved-edge trapezoid as the height, as shown in Figure 1b, and the approximate area of the rectangles is less than the actual area of the curved-edge trapezoid. Therefore, we have

$$\int_{k+1}^{k+x_1+1} \frac{1}{u} du \leq \frac{1}{k+1} + \frac{1}{k+2} + \ldots + \frac{1}{k+x_1+1} \leq \int_{k}^{k+x_1} \frac{1}{u} du.$$

Moreover,

$$\left| \left[\frac{1}{k+1} + \frac{1}{k+2} + \ldots + \frac{1}{k+x_1+1} \right] - \int_{k}^{k+x_1} \frac{1}{u} du \right| \leq \left| \int_{k}^{k+x_1} \frac{1}{u} du - \int_{k+1}^{k+x_1+1} \frac{1}{u} du \right| = \ln \left(\frac{(k+1)(k+x_1)}{k(k+x_1+1)} \right).$$

It is the same for round r,

$$\left| \left[\frac{1}{k + \sum_i^{r-1} x_i + 1} + \frac{1}{k + \sum_i^{r-1} x_i + 2} + \ldots + \frac{1}{k + \sum_i^{r} x_i} \right] - \int_{k+\sum_i^{r-1} x_i}^{k+\sum_i^r x_i} \frac{1}{u} du \right| < \ln \left(\frac{(k+1)(k + \sum_i^r x_i)}{k(k + \sum_i^r x_i + 1)} \right).$$

Therefore, Equation 10 holds and the lemma is proved.

□

Lemma 3.7: In algorithm GREAT^{II}, at the beginning of computational round r, suppose that there are Q free slots in the reservoir where each slot can store one edge. The expected number of edges arrived to put one sampled edge in the i^{th} slot is

$$y_{r,i}^{II} = k \cdot \exp(\alpha \cdot (r-1)) \cdot \left(\exp\left(\frac{1}{k}\right) - 1 \right) \cdot \exp\left(\frac{1}{k} \cdot (i-1)\right).$$

Proof. The prove is similar with *Lemma 3.6*. In round 0, the reservoir is empty, and k edges are sampled with probability 1. When reservoir is full, edges are randomly removed with probability α. There will be $Q (Q = k \cdot \alpha)$ empty slots. In round r, edges are sampled with probability $\frac{k}{t}$ and t starts at $k + \sum_{j=1}^{r-1} x_j + 1$. Assume that filling the first empty slots requires $y_{r,1}$ edges, then we have

$$
\begin{aligned}
&\frac{k}{k + \sum_{j=1}^{r-1} x_j + 1} + \frac{k}{k + \sum_{j=1}^{r-1} x_j + 2} \cdots + \frac{k}{k + \sum_{j=1}^{r-1} x_j + y_{r,1}} = 1 \\
&\frac{1}{k + \sum_{j=1}^{r-1} x_j + 1} + \frac{1}{k + \sum_{j=1}^{r-1} x_j + 2} \cdots + \frac{1}{k + \sum_{j=1}^{r-1} x_j + y_{r,1}} = \frac{1}{k} \\
&\Longrightarrow \int_{k+\sum_{j=1}^{r-1} x_j}^{k+\sum_{j=1}^{r-1} x_j + y_{r,1}} \frac{1}{u} du = \frac{1}{k} \\
&\Longrightarrow \ln \left(\frac{k + \sum_{j=1}^{r-1} x_j + y_{r,1}}{k + \sum_{j=1}^{r-1} x_j} \right) = \frac{1}{k} \\
&\Longrightarrow y_{r,1} = \left(k + \sum_{j=1}^{r-1} x_j \right) \cdot \left(\exp\left(\frac{1}{k}\right) - 1 \right).
\end{aligned}
$$

Similarly for empty slots 2 to Q, repeat the above process, and reveals that after r rounds, we have

$$
\begin{aligned}
y_{r,1} &= \left(k + \sum_{j=1}^{r-1} x_j \right) \cdot \left(\exp\left(\frac{1}{k}\right) - 1 \right) \\
y_{r,2} &= \left(k + \sum_{j=1}^{r-1} x_j + y_{r,1} \right) \cdot \left(\exp\left(\frac{1}{k}\right) - 1 \right) \\
&\cdots \\
y_{r,Q} &= \left(k + \sum_{j=1}^{r-1} x_j + y_{r,1} + y_{r,2} \cdots + y_{r,Q-2} + y_{r,Q-1} \right) \cdot \left(\exp\left(\frac{1}{k}\right) - 1 \right).
\end{aligned}
$$

Then, we have

$$
\begin{aligned}
y_{r,Q} &= \left(k + \sum_{j=1}^{r-1} x_j + y_{r,1} + y_{r,2} \cdots + y_{r,Q-2} + y_{r,Q-1} \right) \cdot \left(\exp\left(\frac{1}{k}\right) - 1 \right) \\
y_{r,Q-1} &= \left(k + \sum_{j=1}^{r-1} x_j + y_{r,1} + y_{r,2} \cdots + y_{r,Q-2} \right) \cdot \left(\exp\left(\frac{1}{k}\right) - 1 \right) \\
y_{r,Q} - y_{r,Q-1} &= y_{r,Q-1} \cdot \left(\exp\left(\frac{1}{k}\right) - 1 \right) \\
y_{r,Q} &= y_{r,Q-1} \cdot \exp\left(\frac{1}{k}\right).
\end{aligned}
$$

Therefore, $\{y_{r,i}\}_{i=1}$ is a geometric progression. The common ratio is $\exp\left(\frac{1}{k}\right)$ and first term is $\left(k + \sum_{i=1}^{r-1} x_i\right) \cdot \left(\exp\left(\frac{1}{k}\right) - 1\right)$, and the general term is

$$
y_{r,i}^{II} = \left(k + \sum_{j=1}^{r-1} x_j \right) \cdot \left(\exp\left(\frac{1}{k}\right) - 1 \right) \cdot \exp\left(\frac{1}{k}(i-1) \right).
$$

From *Lemma 3.6* we have

$$
\sum_{j=1}^{r-1} x_j = k \cdot (\exp(\alpha) - 1) \cdot \frac{\exp(\alpha \cdot (r-1)) - 1}{\exp(\alpha) - 1} = k \cdot (\exp(\alpha \cdot (r-1)) - 1).
$$

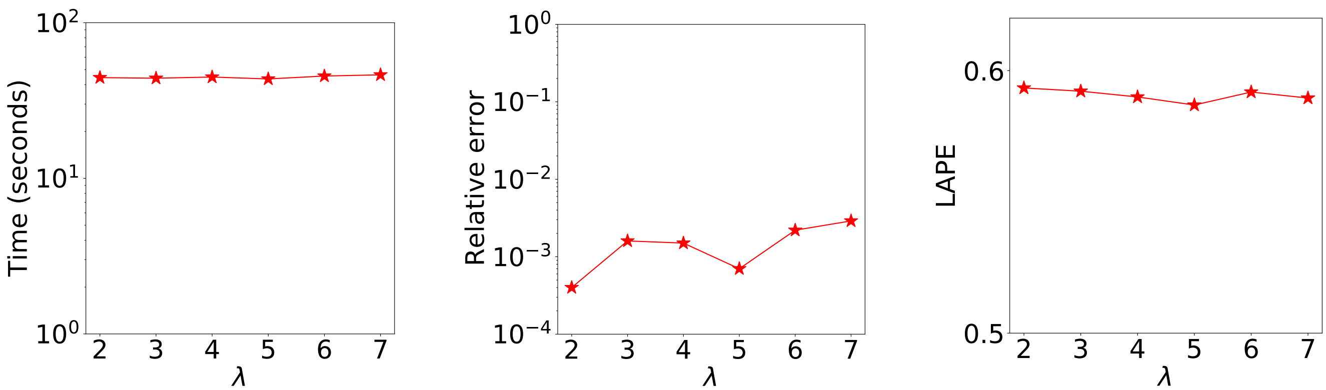
Therefore,

$$y^{II}_{r,i} = \left(k + \sum_{j=1}^{r-1} x_j\right) \cdot \left(\exp\left(\frac{1}{k}\right) - 1\right) \cdot \exp\left(\frac{1}{k} \cdot (i-1)\right) = k \cdot \exp\left(\alpha \cdot (r-1)\right) \cdot \left(\exp\left(\frac{1}{k}\right) - 1\right) \exp\left(\frac{1}{k}(i-1)\right).$$

□

2 Experiment

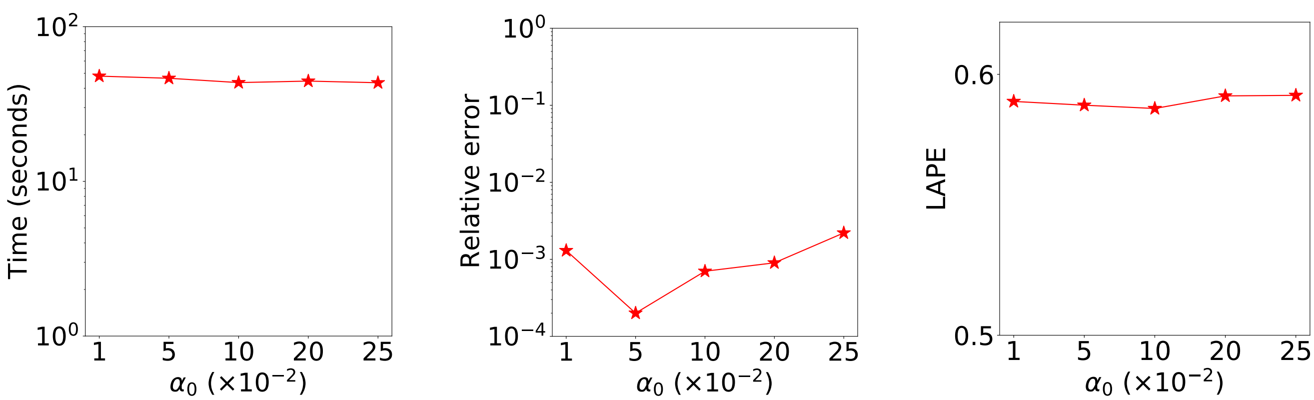
Effect of λ. λ is the number of rounds using the fixed initial value α_0 in GREAT$^+$. Figure 2 shows the performance of algorithm GREAT$^+$ when varying λ on dataset StackOverflow. We observe that the running time, the relative error and the LAPE are insensitive to λ.



(a) Time. (b) Relative error. (c) LAPE.

Figure 2: Varying λ in GREAT$^+$ on StackOverflow.

Effect of α_0. Parameter α_0 is the initial value of α in algorithm GREAT$^+$. Figure 3 shows the performance of algorithm GREAT$^+$ when varying α_0 on dataset StackOverflow. We observe that the running time and the LAPE are insensitive to α_0, while the relative error has a slightly increasing trend. This is because the larger the α_0, the lower the accuracy.



(a) Time. (b) Relative error. (c) LAPE.

Figure 3: Varying α_0 in GREAT$^+$ on StackOverflow.