WRS: Waiting Room Sampling for Accurate Triangle Counting in Real Graph Streams (Supplementary Document)

Kijung Shin
Carnegie Mellon University, Pittsburgh, PA, USA
Email: kijungs@cs.cmu.edu

***Abstract*—In this supplementary document, we provide a pictorial description, proofs, time and space complexity analyses, descriptions of datasets, and additional experimental results, all of which supplement the main paper [1].**

APPENDIX A
PICTORIAL DESCRIPTION OF THE SAMPLING PROCESS

Figure 5 gives a pictorial representation of the sampling process in WRS.

APPENDIX B
PROOFS

In this section, we present the proofs of Lemma 1, Theorem 1, and Lemma 2 of the main paper.

A. Proof of Lemma 1

Proof. Without loss of generality, we assume $t^{(1)}_{uvw} = t_{vw}$, $t^{(2)}_{uvw} = t_{wu}$, and $t^{(3)}_{uvw} = t_{uv}$. That is, (v, w) arrives earlier than (w, u), and (w, u) arrives earlier than (u, v).

If $type_{uvw} = 1$, (u, v) arrives at time $k+1$ or earlier. When (u, v) arrives, (v, w) and (w, u) are always stored in $\mathcal{S}$. Thus, WRS discovers (u, v, w) with probability 1.

If $type_{uvw} = 2$, we have $t_{uv} - t_{wu} < t_{uv} - t_{vw} \leq k\alpha$. When (u, v) arrives, (v, w) and (w, u) are always stored in $\mathcal{W}$. Thus, WRS discovers (u, v, w) with probability 1.

If $type_{uvw} = 3$, then $t_{uv} - t_{wu} \leq k\alpha$ but $t_{uv} - t_{vw} > k\alpha$. When (u, v) arrives, (w, u) is always stored in $\mathcal{W}$, while (v, w) cannot be in $\mathcal{W}$ but can be in $\mathcal{R}$ with probability $p^{(t_{uv}-1)}$ (see Eq. (1)). For WRS to discover (u, v, w), (v, w) should be in $\mathcal{R}$, thus the probability is $p^{(t_{uv}-1)} = \frac{k(1-\alpha)}{(t_{uv}-1-k\alpha)} = \frac{k(1-\alpha)}{(t^{(3)}_{uvw}-1-k\alpha)}$.

If $type_{uvw} = 4$, we have $t_{uv} - t_{vw} > t_{uv} - t_{wu} > k\alpha$. Thus, when (u, v) arrives, (v, w) and (w, u) cannot be in $\mathcal{W}$. For WRS to discover (u, v, w), both (v, w) and (w, u) should be in $\mathcal{R}$ when (u, v) arrives. The probability of the event is

$$\begin{aligned}&\mathbb{P}[(v, w) \in \mathcal{R} \text{ and } (w, u) \in \mathcal{R}]\\ &= \mathbb{P}[(v, w) \in \mathcal{R}] \times \mathbb{P}[(w, u) \in \mathcal{R} | (v, w) \in \mathcal{R}]\\ &= \frac{k(1-\alpha)}{t_{uv}-1-k\alpha} \times \frac{k(1-\alpha)-1}{t_{uv}-2-k\alpha},\end{aligned}$$

which is equal to the last case of Eq. (2). ■

B. Proof of Theorem 1

Proof. From Eq. (2), if $k(1-\alpha) \geq 2$, we have $\mathbb{E}[x_{uvw}] = 1/p_{uvw} \times p_{uvw} + 0 \times (1 - p_{uvw}) = 1$. Combining this and $c^{(t)} = \sum_{(u,v,w)\in\mathcal{T}^{(t)}} x_{uvw}$ gives

$$\mathbb{E}[c^{(t)}] = \mathbb{E}\left[\sum_{(u,v,w)\in\mathcal{T}^{(t)}} x_{uvw}\right] = \sum_{(u,v,w)\in\mathcal{T}^{(t)}} \mathbb{E}[x_{uvw}] = |\mathcal{T}^{(t)}|,$$

which proves Eq. (3) for every $t \in \{1, 2, ...\}$. Likewise, $\mathbb{E}[x_{uvw}] = 1$ and $c^{(t)}_u = \sum_{(u,v,w)\in\mathcal{T}^{(t)}_u} x_{uvw}$ imply

$$\mathbb{E}[c^{(t)}_u] = \mathbb{E}\left[\sum_{(u,v,w)\in\mathcal{T}^{(t)}_u} x_{uvw}\right] = \sum_{(u,v,w)\in\mathcal{T}^{(t)}_u} \mathbb{E}[x_{uvw}] = |\mathcal{T}^{(t)}_u|,$$

which proves Eq. (4) for every $u \in \mathcal{V}^{(t)}$ and $t \in \{1, 2, ...\}$. ■

C. Proof of Lemma 2

Proof. By the definition of $type_{uvw}$, $type_{uvw} \geq 2$ implies

$$t^{(3)}_{uvw} > k + 1. \tag{9}$$

First, we show the case when $type_{uvw} = 2$. Eq. (7) and Eq. (9) give

$$\frac{t^{(3)}_{uvw}-1}{k} \times \frac{t^{(3)}_{uvw}-2}{k-1} - 1 > 0 = \mathrm{Var}[x_{uvw}].$$

Second, we show the case when $type_{uvw} = 3$ and $t^{(3)}_{uvw} > 1 + \frac{\alpha}{1-\alpha}k$. From $t^{(3)}_{uvw} > 1 + \frac{\alpha}{1-\alpha}k$, $\left(1 + \frac{k}{t^{(3)}_{uvw}-1}\right)\alpha < 1$ holds. This and Eq. (9) imply $\left(1 + \frac{k}{t^{(3)}_{uvw}-1}\right)\left(1 - \frac{k-1}{t^{(3)}_{uvw}-2}\right)\alpha < 1 - \frac{k-1}{t^{(3)}_{uvw}-2}$. Again, this and Eq. (9) give $\left(1 - \frac{k(k-1)}{(t^{(3)}_{uvw}-1)(t^{(3)}_{uvw}-2)}\right)\alpha < \left(1 + \frac{k}{t^{(3)}_{uvw}-1} - \frac{k-1}{t^{(3)}_{uvw}-2} - \frac{k(k-1)}{(t^{(3)}_{uvw}-1)(t^{(3)}_{uvw}-2)}\right)\alpha < 1 - \frac{k-1}{t^{(3)}_{uvw}-2}$. This is equivalent to $(k-1)(t^{(3)}_{uvw} - 1 - k\alpha) < (t^{(3)}_{uvw} - 1)(t^{(3)}_{uvw} - 2)(1-\alpha)$, which is again equivalent to $\frac{t^{(3)}_{uvw}-1-k\alpha}{k(1-\alpha)} - 1 < \frac{t^{(3)}_{uvw}-1}{k} \times \frac{t^{(3)}_{uvw}-2}{k-1} - 1$. Combining this and Eq. (7) gives

$$\mathrm{Var}[x_{uvw}] = \frac{t^{(3)}_{uvw}-1-k\alpha}{k(1-\alpha)} - 1 < \frac{t^{(3)}_{uvw}-1}{k} \times \frac{t^{(3)}_{uvw}-2}{k-1} - 1.$$

Lastly, the same conclusion holds when $type_{uvw} = 3$ and $\alpha < 0.5$. Eq. (9) and $\alpha < 0.5$ imply $t^{(3)}_{uvw} > 1 + \frac{\alpha}{1-\alpha}k$. This and $type_{uvw} = 3$ give the second case, which is proven above.

Hence, Eq. (8) holds under any of the given conditions. ■

APPENDIX C
TIME AND SPACE COMPLEXITY ANALYSES

In this section, we prove the time and space complexities of WRS. Especially, we show that WRS has the same time and space complexities as the state-of-the-art algorithms [2], [3]. We assume that sampled edges are stored in the adjacency list format in memory, as in our implementation used for our experiments. However, storing them sequentially, as in Figure 5, does not change the results below.

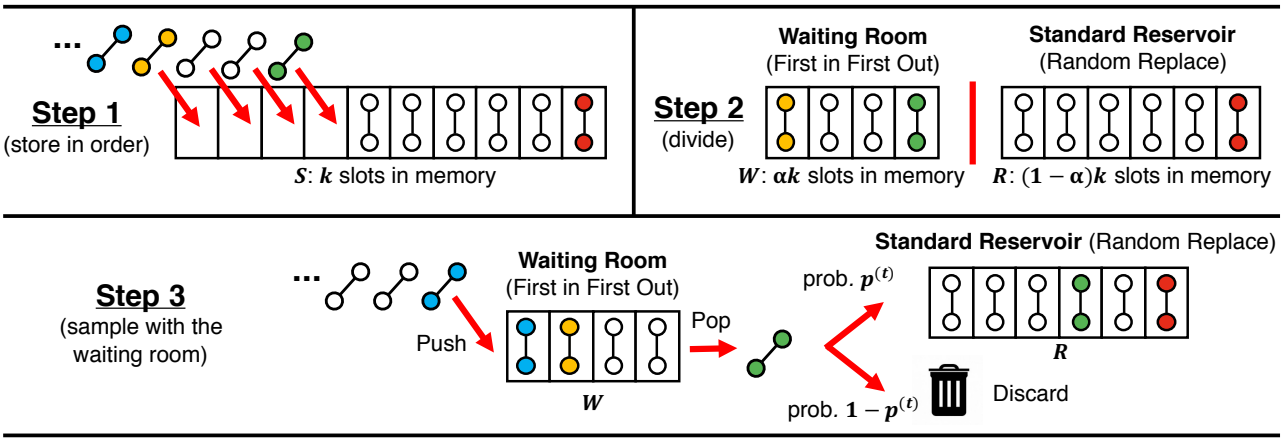


Fig. 5: Pictorial description of the sampling process in WRS. Once the given memory space is full (by Step 1), the memory space is divided into the waiting room and the reservoir (by Step 2). In Step 3, the latest αk edges are stored in the waiting room, while the remaining older edges are uniformly sampled in the reservoir.

The worst-case time complexity of WRS is linear in the memory budget and in the number of edges in the input stream, as formalized in Theorem 2.

Theorem 2 (Worst-Case Time Complexity of WRS)**.** *Processing an incoming edge in Algorithm 1 takes $O(k)$, and thus processing t edges in the input stream takes $O(kt)$.*

Proof. The most expensive step in processing an incoming edge (u, v) in Algorithm 1 is to find their common neighbors $\hat{\mathcal{N}}_u \cap \hat{\mathcal{N}}_v$ in line 2. Computing $\hat{\mathcal{N}}_u \cap \hat{\mathcal{N}}_v$ requires accessing $|\hat{\mathcal{N}}_u| + |\hat{\mathcal{N}}_v| = O(k)$ edges. ■

However, this analysis assuming the worst-case graph stream is too pessimistic for real graph streams, where $|\hat{\mathcal{N}}_u| + |\hat{\mathcal{N}}_v|$ is usually much smaller than k.

Theorem 3 gives the space complexity of WRS. Note that, except the space for outputs (specifically local triangle counts), WRS only requires $O(k)$ space.

Theorem 3 (Space Complexity of WRS)**.** *Let $\mathcal{V}^{(t)}$ be the set of nodes in the graph consisting of the first t edges in the input stream. Processing t edges in the input stream by Algorithm 1 requires $O(k)$ space in case of global triangle counting and $O(k + |\mathcal{V}^{(t)}|)$ space in case of local triangle counting.*

Proof. Algorithm 1 uses $O(k)$ space for sampling edges, and it uses $O(|\mathcal{V}^{(t)}|)$ space for maintaining local triangle counts, which need not be maintained in case of global triangle counting. ■

APPENDIX D
DESCRIPTION OF REAL GRAPH STREAMS

In this section, we describe the real dynamic graph streams used in our experiments.

- **ArXiv** [4]: A citation network between papers in ArXiv's High Energy Physics. Each edge (u, v) represents that paper u cited paper v. We used the submission time of u as the creation time of (u, v).
- **Facebook** [5]: A friendship network between users of Facebook. Each edge (u, v) represents that user v appeared in the friend list of user u. Edges whose creation times are unknown were ignored.
- **Email** [6]: An email network from Enron Corporation. Each edge (u, v) represents that employee u sent to or received from person v (who may be a non-employee) at least one email. We used the creation time of the first email between u and v as the creation time of (u, v).
- **Youtube** [7]: A friend network between users of Youtube. Each edge (u, v) represents that user u and user v became friends. Edges created before 12/10/2006 were ignored since their exact creation times are unknown.
- **Patent** [8]: A citation network between patents. Each edge (u, v) indicates that patent u cited patent v. We used the time when u was granted as the creation time of (u, v).

The self loops, the duplicated edges, and the directions of the edges were ignored in all the graph streams.

APPENDIX E
ADDITIONAL EXPERIMENTS

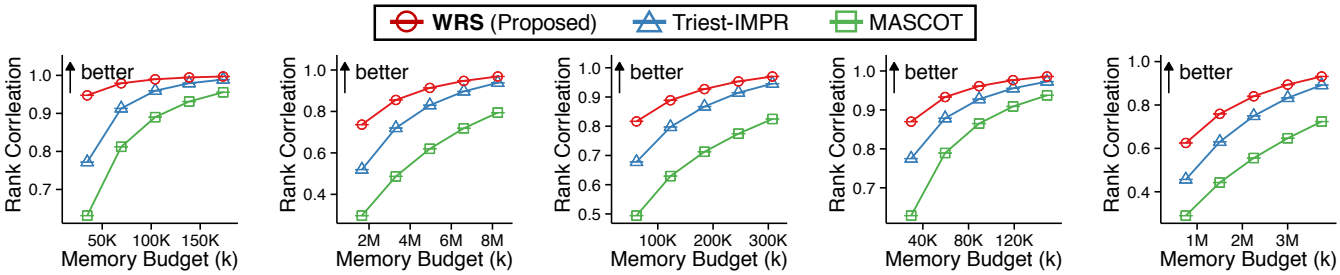
In this section, we present the results of additional experiments to answer the following questions:

- **Q4. Accuracy in terms of Rank Correlation**: Is WRS more accurate than its competitors especially when we compare their accuracies in terms of Rank Correlation?
- **Q5. Effects of α**: How does the relative size α of the waiting room affect the accuracy of WRS? What is the optimal value of α?

The detailed experimental settings were the same with those in the main paper.

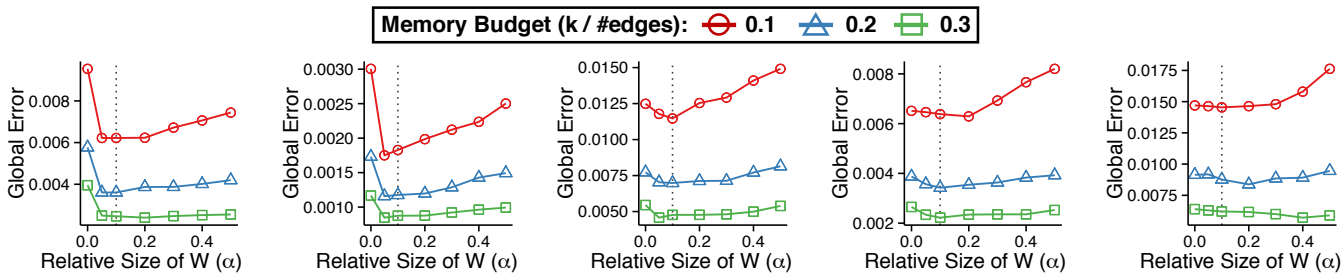
A. Q4. Accuracy in terms of Rank Correlation

We compared the accuracies of the considered methods in local triangle counting using Spearman's rank correlation coefficient [9]. Specifically, we used it to measure the similarity of (a) the ranking of the nodes in terms of the true local triangle counts and (b) their ranking in terms of the estimated local triangle counts, at the end of each input stream. The coefficient



(a) ArXiv (b) Patent (c) Email (d) Facebook (e) Youtube

Fig. 6: WRS is accurate. M: million, K: thousand. In all the datasets, WRS shows the highest accuracy in global and local triangle counting regardless of memory budget k. The relative size α of the waiting room is fixed to 0.1.



(a) ArXiv (b) Patent (c) Email (d) Facebook (e) Youtube

Fig. 7: Effects of α on the accuracy of WRS . Using about 10% of given memory space for the waiting room ($\alpha = 0.1$) gives higher accuracy than using no space for the waiting room ($\alpha = 0$) or using half the space for the waiting room ($\alpha = 0.5$).

has a value between -1 and 1, and a higher value indicates higher accuracy in local triangle counting.

Figure 6 shows the results in the real graph streams with different memory budgets k. In the figure, the average values over $1,000$ runs were reported with error bars indicating the estimated standard errors. In all the datasets, WRS was most accurate, giving highest rank correlation coefficients regardless of memory budgets.

B. Q5. Effects of the Size of the Waiting Room of WRS

We measured how the accuracy of WRS changes depending on α, the relative size of the waiting room. Figure 7 shows the results with different memory budgets. Here, we used global error as the accuracy metric, and the average values over $1,000$ runs were reported. In all the datasets and regardless of memory budgets, using proper amount of memory space for the waiting room gave better accuracy than using no space for the waiting room ($\alpha = 0$) and using half the space for the waiting room ($\alpha = 0.5$). Although proper α values depended on datasets and memory budgets, the accuracy was maximized when α was about 0.1 in most of the cases.

REFERENCES

[1] K. Shin, "Wrs: Waiting room sampling for accurate triangle counting in real graph streams," in *ICDM*, 2017.

[2] Y. Lim and U. Kang, "Mascot: Memory-efficient and accurate sampling for counting local triangles in graph streams," in *KDD*, 2015.

[3] L. De Stefani, A. Epasto, M. Riondato, and E. Upfal, "Triest: Counting local and global triangles in fully-dynamic streams with fixed memory size," in *KDD*, 2016.

[4] J. Gehrke, P. Ginsparg, and J. Kleinberg, "Overview of the 2003 kdd cup," *ACM SIGKDD Explorations Newsletter*, vol. 5, no. 2, pp. 149–151, 2003.

[5] B. Viswanath, A. Mislove, M. Cha, and K. P. Gummadi, "On the evolution of user interaction in facebook," in *WOSN*, 2009.

[6] B. Klimt and Y. Yang, "Introducing the enron corpus." in *CEAS*, 2004.

[7] A. Mislove, "Online social networks: Measurement, analysis, and applications to distributed information systems," Ph.D. dissertation, Rice University, 2009.

[8] B. H. Hall, A. B. Jaffe, and M. Trajtenberg, "The nber patent citation data file: Lessons, insights and methodological tools," National Bureau of Economic Research, Tech. Rep., 2001.

[9] C. Spearman, "The proof and measurement of association between two things," *The American journal of psychology*, vol. 15, no. 1, pp. 72–101, 1904.