TRIÈST: Counting Local and Global Triangles in Fully-Dynamic Streams with Fixed Memory Size

Lorenzo De Stefani
Brown University
Providence, RI, USA
lorenzo@cs.brown.edu

Matteo Riondato*
Two Sigma Investments
New York, NY, USA
matteo@twosigma.com

Alessandro Epasto*
Google
New York, NY, USA
aepasto@google.com

Eli Upfal
Brown University
Providence, RI, USA
eli@cs.brown.edu

“Ogni lassada xe persa”[1] – Proverb from Trieste, Italy.

ABSTRACT

We present TRIÈST, a suite of one-pass streaming algorithms to compute unbiased, low-variance, high-quality approximations of the global and local (i.e., incident to each vertex) number of triangles in a fully-dynamic graph represented as an adversarial stream of edge insertions and deletions.

Our algorithms use reservoir sampling and its variants to exploit the user-specified memory space at all times. This is in contrast with previous approaches, which require hard-to-choose parameters (e.g., a fixed sampling probability) and offer no guarantees on the amount of memory they use. We analyze the variance of the estimations and show novel concentration bounds for these quantities.

Our experimental results on very large graphs demonstrate that TRIÈST outperforms state-of-the-art approaches in accuracy and exhibits a small update time.

1. INTRODUCTION

Exact computation of characteristic quantities of Web-scale networks is often impractical or even infeasible due to the humongous size of these graphs. It is natural in these cases to resort to efficient-to-compute approximations of these quantities, which, when of sufficiently high quality, can be used as proxies for the exact values.

In addition to being huge, many interesting networks are fully-dynamic and can be represented as a stream whose elements are edges/nodes insertions and deletions occurring in an arbitrary (even adversarial) order. Characteristic quantities in these graphs are intrinsically volatile, hence there is limited added value in maintaining them exactly. The goal is rather to keep track, at all times, of a high-quality approximation of these quantities. For efficiency, the algorithms should aim at exploiting the available memory space as much as possible and they should require only one pass over the stream.

We introduce TRIÈST, a suite of sampling-based, one-pass algorithms for adversarial fully-dynamic streams to approximate the global number of triangles and the local number of triangles incident to each vertex. Mining local and global triangles is a fundamental primitive with many applications (e.g., community detection [1], topic mining [10], spam/anomaly detection [3, 27], ego-networks mining [12] and protein interaction networks analysis [29].)

Many previous works on triangle estimation in streams also employ sampling (see Sect. 3), but they usually require the user to specify in advance an edge sampling probability p that is fixed for the entire stream. This approach presents several significant drawbacks. First, choosing a p that allows to obtain the desired approximation quality requires to know or guess a number of properties of the input (e.g., the size of the stream). Second, a fixed p implies that the sample size grows with the size of the stream, which is problematic when the stream size is not known in advance: if the user specifies a large p, the algorithm may run out of memory, while for a smaller p it will provide a suboptimal estimation. Third, even assuming to be able to compute a p that ensures (in expectation) full use of the available space, the memory would be fully utilized only at the end of the stream, and the estimations computed throughout the execution would be suboptimal.

Contributions. We address all the above issues by taking a significant departure from the fixed-probability, independent edge sampling approach taken even by state-of-the-art methods [27]. Specifically:

- We introduce TRIÈST (TRIangle Estimation from STreams), a suite of one-pass streaming algorithms to approximate, at each time instant, the global and local number of triangles in a fully-dynamic graph stream (i.e., a sequence of edges additions and deletions in arbitrary order) using a fixed amount of memory. This is the first contribution that enjoys all these properties. TRIÈST only requires the user to specify the amount of available memory, an interpretable parameter that is definitively known to the user.
- Our algorithms maintain a sample of edges: they use the reservoir sampling [27] and random pairing [14] sampling...
schemes to exploit the available memory as much as possible. To the best of our knowledge, ours is the first application of these techniques to subgraph counting in fully-dynamic, arbitrarily long, adversarially ordered streams. We present an analysis of the unbiasedness and of the variance of our estimators, and establish strong concentration results for them. The use of reservoir sampling and random pairing requires additional sophistication in the analysis, as the presence of an edge in the sample is not independent from the concurrent presence of another edge. Hence, in our proofs we must consider the complex dependencies in events involving sets of edges. The gain is worth the effort: we prove that the variance of our algorithms is smaller than that of state-of-the-art methods 22, and this is confirmed by our experiments.

- We conduct an extensive experimental evaluation of TRIEST on very large graphs, some with billions of edges, comparing the performances of our algorithms to those of existing state-of-the-art contributions 18 27 32. Our algorithms significantly and consistently reduce the average estimation error by up to 90% w.r.t. the state of the art, both in the global and local estimation problems, while using the same amount of memory. Our algorithms are also extremely scalable, showing update times in the order of hundreds of microseconds for graphs with billions of edges. Due to space constraints, the proofs and additional experimental results can be found in the extended version 9.

3. RELATED WORK

The literature on triangle counting is extremely rich, including exact algorithms, graph sparsifiers 25 36, complex-valued sketches 29 28, and MapReduce algorithms 30 31 33. Here we restrict the discussion to the works most related to ours, i.e., to those presenting algorithms for counting or approximating the number of triangles from data streams. We refer to the survey by Latapy 25 for an in-depth discussion of other works.

Many previous contributions presented algorithms for more restricted (i.e., less generic) settings than ours, or for which the constraints on the computation are more lax 2 6 19 22. For example, Becchetti et al. 8 and Kolountzakis et al. 21 present algorithms for approximate triangle counting from static graphs by performing multiple passes over the input. Pavan et al. 32 and Jha et al. 18 propose algorithms for approximating only the global number of triangles from edge-insertion-only streams. Kutzkov and Pagh 23 present a one-pass algorithm for fully-dynamic graphs, but the triangle count estimation is (expensively) computed only at the end of the stream and the algorithm requires, in the worst case, more memory than what is needed to store the entire graph. Ahmed et al. 1 apply the sampling-and-hold approach to insertion-only graph stream mining to obtain, only at the end of the stream and using non-constant space, an estimation of many network measures including triangles.

None of these works has all the features offered by TRIEST: performs a single pass over the data, handles fully-dynamic streams, uses a fixed amount of memory space, requires a single interpretable parameter, and returns an estimation at each time instant. Furthermore, our experimental results show that we outperform the algorithms from 18 32 on insertion-only streams.

Lim and Kang 27 present an algorithm for insertion-only streams that is based on independent edge sampling with a fixed probability. Since the memory is not fully utilized during most of the stream, the variance of the estimate is large. Our approach handles fully-dynamic streams and makes better use of the available memory space at each time instant, resulting in a better estimation, as shown by our analytical and experimental results.

Vitter 37 presents a detailed analysis of the reservoir sampling scheme and discusses methods to speed up the algorithm by reducing the number of calls to the random number generator. Random Pairing 14 is an extension of reservoir sampling to handle fully-dynamic streams with insertions and deletions. Cohen et al. 8 generalize and extend the Random Pairing approach to the case where the elements on the stream are key-value pairs, where the value may be negative (and less than −1). In our settings, where the value is not less than −1 (for an edge deletion), these generalizations do not apply and the algorithm presented by Cohen et al. 8 reduces essentially to Random Pairing.

4. ALGORITHMS

We present TRIEST, a suite of three novel algorithms for approximate global and local triangle counting from edge streams. The first two work on insertion-only streams, while the third can handle fully-dynamic streams where edge deletions are allowed.

Parameters. Our algorithms keep an edge sample S of up to M edges from the stream, where M is a positive integer

2. PRELIMINARIES

We study the problem of counting global and local triangles in a fully-dynamic undirected graph as an arbitrary (adversarial) stream of edge insertions and deletions.

Formally, for any (discrete) time instant \( t \geq 0 \), let \( G^{(t)} = (V^{(t)}, E^{(t)}) \) be the graph observed up to and including time \( t \). At time \( t = 0 \) we have \( V^{(0)} = E^{(0)} = \emptyset \). For any \( t > 0 \), at time \( t + 1 \) we receive an element \( e_{t+1} = (u, v, \epsilon) \) from a stream, where \( \epsilon \in \{+, -\} \) and \( u, v \) are two distinct vertices. The graph \( G^{(t+1)} = (V^{(t+1)}, E^{(t+1)}) \) is obtained by inserting a new edge or deleting an existing edge as follows:

\[
E^{(t+1)} = \begin{cases} 
E^{(t)} \cup \{(u,v)\} & \text{if } \epsilon = "+" \\
E^{(t)} \setminus \{(u,v)\} & \text{if } \epsilon = "-" 
\end{cases}
\]

If \( u \) or \( v \) do not belong to \( V^{(t)} \), they are added to \( V^{(t+1)} \). Nodes are deleted from \( V^{(t)} \) when they have degree zero. Edges can be added and deleted in the graph in an arbitrary adversarial order, i.e., as to cause the worst outcome for the algorithm, but we assume that the adversary has no access to the random bits used by the algorithm. We assume that all operations have effect: if \( e \in E^{(t)} \) (resp. \( e \notin E^{(t)} \)), \((+,\epsilon)\) (resp. \((-,-)\)) can not be on the stream at time \( t + 1 \).

Given a graph \( G^{(t)} = (V^{(t)}, E^{(t)}) \), a triangle in \( G^{(t)} \) is a set of three vertices \( \{u, v, w\} \subseteq V^{(t)} \) such that \( \{(u,v),(v,w),(w,u)\} \subseteq E^{(t)} \). We refer to the vertices forming a triangle as its corners. We denote with \( \Delta^{(t)} \) the set of all triangles in \( G^{(t)} \), and, for any vertex \( u \in V^{(t)} \), with \( \Delta_u^{(t)} \) the subset of \( \Delta^{(t)} \) containing all and only the triangles that have \( u \) as a corner.

Problem definition. We study the Global (resp. Local) Triangle Counting Problem in Fully-Dynamic Streams, which requires to compute, at each time \( t \geq 0 \) an estimation of \( |\Delta^{(t)}| \) (resp. for each \( u \in V \) an estimation of \( |\Delta_u^{(t)}| \)).
parameter. For ease of presentation, we realistically assume $M \geq 6$. In Sect. 1 we motivated the design choice of only requiring $M$ as a parameter and remarked on its advantages over using a fixed sampling probability $p$. Our algorithms are designed to use the available space as much as possible.

**Counters.** TRIÉST algorithms keep counters to compute the estimations of the global and local number of triangles. They always keep one global counter $\tau$ for the estimation of the global number of triangles. Only the global counter is needed to estimate the total triangle count. To estimate the local triangle counts, the algorithms keep a set of local counters $\tau_u$ for a subset of the nodes $u \in V^t$. The local counters are created on the fly as needed, and always destroyed as soon as they have a value of 0. Hence our algorithms use $O(M)$ space (with one exception, see Sect. 4.2).

**Notation.** For any $t \geq 0$, let $G^S = (V^S, E^S)$ be the subgraph of $G^{t+1}$ containing all and only the edges in the current sample $S$. We denote with $N^S_u$ the neighborhood of $u$ in $G^S$; $N^S_u = \{ v \in V^t : (u,v) \in S \}$ and with $N^S_{u,v} = N^S_u \cap N^S_v$ the shared neighborhood of $u$ and $v$ in $G^S$.

### 4.1 A first algorithm — TRIÉST-BASE

We first present TRIÉST-BASE, which works on insertion-only streams and uses standard reservoir sampling [37] to maintain the edge sample $S$:

- If $t \leq M$, then the edge $e_t = (u,v)$ on the stream at time $t$ is deterministically inserted in $S$.
- If $t > M$, TRIÉST-BASE flips a biased coin with heads probability $M/t$. If the outcome is heads, it chooses an edge $(w,z) \in S$ uniformly at random, removes $(w,z)$ from $S$, and inserts $(u,v)$ in $S$. Otherwise, $S$ is not modified. After each insertion (resp. removal) of an edge $(u,v)$ from $S$, TRIÉST-BASE calls the procedure UpdateCounters that increments (resp. decrements) $\tau_u$, $\tau_v$ and $\tau_{u,v}$ by $|N^S_{u,v}|$ and $\tau_{u,v}$ by one, for each $c \in N^S_{u,v}$.

The pseudocode for TRIÉST-BASE is presented in Alg. 1.

#### Algorithm 1 TRIÉST-BASE

**Input:** Insertion-only edge stream $\Sigma$, integer $M \geq 6$

1: $S \leftarrow \emptyset$, $t \leftarrow 0$, $\tau \leftarrow 0$
2: for each element $(+, (u,v))$ from $\Sigma$ do
3: $t \leftarrow t + 1$
4: if SAMPLEEdge($u,v,t$) then
5: $S \leftarrow S \cup \{(u,v)\}$
6: UpdateCounters$(+, (u,v))$
7: else if FLIPBASEDcoin$(\frac{M}{t})$ = heads then
8: $(u',v') \leftarrow$ random edge from $S$
9: $S \leftarrow S \setminus \{(u',v')\}$
10: UpdateCounters$(-, (u',v'))$
11: return True
12: return False

13: function SAMPLEEdge($u,v,t$)
14: if $t \leq M$ then
15: $\tau \leftarrow + 1$
16: return True
17: return False

18: function UpdateCounters($\bullet, (u,v))$
19: $N^S_{u,v} \leftarrow N^S_u \cap N^S_v$
20: for all $c \in N^S_{u,v}$ do
21: $\tau \leftarrow \tau + 1$
22: $\tau_u \leftarrow \tau_u + 1$
23: $\tau_v \leftarrow \tau_v + 1$
24: $\tau_{u,v} \leftarrow \tau_{u,v} + 1$

---

**Estimation.** For any $t \geq 0$, let $\xi(t) = \max\{1, \frac{t(t-1)(t-2)}{M(M-1)(M-2)}\}$. Let $\tau(t)$ (resp. $\tau_u(t)$) be the value of the counter $\tau$ (at the end of time step $t$ (i.e., after the edge on the stream at time $t$ has been processed by TRIÉST-BASE) (resp. the value of the counter $\tau_u$ at the end of time step $t$ if there is such a counter, 0 otherwise). When queried at the end of time $t$, TRIÉST-BASE returns $\xi(t)^t \tau(t)$ (resp. $\xi(t)^t \tau_u(t)$) as the estimation for the global (resp. local for $u \in V^t$) triangle count.

**Analysis.**

**Theorem 1.** We have

\[
\xi(t)^t \tau(t) = |\Delta(t)| \text{ if } t \leq M, \\
\mathbb{E}[\xi(t)^t \tau(t)] = |\Delta(t)| \text{ if } t > M.
\]

The TRIÉST-BASE estimations are not only unbiased in all cases, but actually exact for $t \leq M$, i.e., for $t \leq M$, they are the true global/local number of triangles in $G^t$.

We now analyze the variance of the estimation returned by TRIÉST-BASE for $t > M$ (the variance is 0 for $t \leq M$). Let $t \geq 0$. For any $u \in V^t$, let $r_u(t)$ be the number of unordered pairs of distinct triangles from $\Delta_u$ sharing an edge. Similarly, let $r_u(t) = \frac{1}{2} \sum_{w \in V^t} r_{u,w}(t)$ be the total number of unordered pairs of distinct triangles sharing an edge. We also define $w_u(t) = (|\Delta_u| - \tau_u(t))$ as the number of unordered pairs of distinct triangles that do not share any edge, and analogously for $w_u(t)$.

**Theorem 2.** For any $t > M$, let $f(t) = \xi(t)^t - 1$, $g(t) = \xi(t)^t (M - 3)(M - 4) + \tau_u(t)$, and $h(t) = \xi(t)^t (M - 3)(M - 4) + \tau_u(t) + w_u(t)$. We have:

\[
\var{\xi(t)^t \tau(t)} = |\Delta(t)| f(t) + r_u(t) g(t) + w_u(t) h(t). \tag{1}
\]

In our proofs, we carefully account for the fact that, as we use reservoir sampling [37], the presence of an edge $a$ in $S$ is not independent from the concurrent presence of another edge $b$ in $S$. This is not the case for samples built using fixed-probability independent edge sampling. When computing the variance, we must consider not only pairs of triangles that share an edge (as for independent edge sampling approaches), but also pairs of triangles sharing no edge, as their respective presences in the sample are not independent events. The gain is worth the additional sophistication needed in the analysis, because the contribution to the variance by triangles not sharing edges is non-negative ($h(t) \leq 0$), i.e., it reduces the variance. A comparison of the variance of our estimator with that obtained with a fixed-probability independent edge sampling approach, is discussed below.

Let $h(t)$ denote the maximum number of triangles sharing a single edge in $G^t$. The following concentration theorem relies on 1. a result by Hajnal and Szemerédi [15] on graph coloring, 2. a novel concentration result for fixed-probability independent edge sampling, and 3. a Poisson-approximation-like result on probabilities of general events under reservoir sampling w.r.t. their probabilities under independent edge sampling. These ingredients are then combined to obtain the following result. The details can be found in our extended online version [9].

---

1Two distinct triangles can share at most one edge.
Theorem 3. Let \( t \geq 0 \) and assume \( |\Delta(t)| > 0 \)\(^2\) For any \( \epsilon, \delta \in (0, 1) \), let
\[
\Phi = \sqrt{8\epsilon^{-2} 3h(t) + 1} \ln \left( \frac{(3h(t) + 1)t}{\delta} \right).
\]
If
\[
M \geq \max \left\{ t \Phi \left( 1 + \frac{1}{2} \ln^{2/3} (t \Phi) \right), 12 \epsilon^{-1} + \epsilon^2, 25 \right\},
\]
then \( |\xi(t) - \Delta(t)| < \epsilon |\Delta(t)| \) with probability \( > 1 - \delta \).

Results analogous to those in Thms. 1, 2, and 3 hold for
\[\lim_{t \to \infty} \frac{|\Delta(t)|}{\epsilon, \delta} \in\] sufficiently smaller \( p \), the variance of this estimator is
\[\epsilon, \delta \in \] sufficiently smaller \( \epsilon \), and \( \delta \). We may show the following result.

Lemma 1. Let \( 0 < \alpha < 1 \) be a constant. For any constant \( \alpha M \geq 8 \) and any \( \alpha T \), we have \( \xi(t) < \xi(t) \).

For example, if we set \( \alpha = 0.99 \) and run trièst-base with \( M \geq 400 \) and mas-\c c-c in \( P/\alpha \), we have that
\[\text{trièst-base has strictly smaller variance than mas-}\c c-c for \] 99% of the stream.

What about \( t = T \)? The difference between the definitions of \( \xi(t) \) and \( \xi(t) \) is in the presence of \( f(M/T) \) instead of \( f(t) \) (resp. \( g(M/T) \) instead of \( g(t) \)) as well as the additional term \( u(1 - h(M, 0) \leq 0 \) in our \( \xi(t) \). Let \( M(T) \) be an arbitrary increasingly increasing function of \( T \). For \( T \to \infty \) we can show that
\[\lim_{T \to \infty} \frac{\int f(M(T)/T) \, dt}{g} = \lim_{T \to \infty} \frac{\int g(M(T)/T) \, dt}{g} = 1, \] hence, informally, \( \xi(t) \to \xi(t) \) for \( T \to \infty \).

A similar discussion also holds for the method we present in Sect. 2, and explains the results of our experimental evaluations, which show that our algorithms have strictly lower (empirical) variance than fixed probability approaches for most of the stream.

Update time. The time to process an element of the stream is dominated by the computation of the shared neighbor-191 hood \( N_u \) in \( \text{Update} \). A MERCORD-based algorithm for the intersection requires \( O(\deg(u) + \deg(v)) \)

\[\text{time, where the degrees are w.r.t. the graph } G_S. \] By storing the neighborhood of each vertex in a Hash Table (resp. an AVL-tree), the update time can be reduced to \( O(\min(\deg(v), \deg(u))) \) (resp. amortized time \( O(\min(\deg(v), \deg(u)) + \log \max \{\deg(v), \deg(u)\}) \)).

4.2 Improved insertion algorithm – trièst-impr

TRIÈST-IMPR is a variant of TRIÈST-BASE with small modifications that result in higher-quality (i.e., lower variance) estimations. The changes are:

1. \( \text{UpdateCounters} \) is called unconditionally for each element on the stream, before the algorithm decides whether or not to insert the edge into \( S \). W.r.t. the pseudocode in Alg. 1, this change corresponds to moving the call to \( \text{UpdateCounters} \) on line 6 to before the \( \text{if} \) block. Mascot \[\text{uses a similar idea, but } \text{TRIÈST-IMPR} \] significantly different from \( \text{TRIÈST-IMPR} \) allows edges to be removed from the sample, since it uses reservoir sampling.

2. \( \text{TRIÈST-IMPR never decrements} \) the counters when an edge is removed from \( S \). W.r.t. the pseudocode in Alg. 1, we remove the call to \( \text{UpdateCounters} \) on line 13.

3. \( \text{UpdateCounters} \) performs a weighted increase of the counters using \( \eta(t) = \max \{1, (t - 1)(t - 2)/(S(M - 1))\} \) as weight. W.r.t. the pseudocode in Alg. 1, we replace line 19 with line \( \eta(t) \) on lines 23–25 (given change 2 above, all the calls to \( \text{UpdateCounters} \) have \( \epsilon = + \)).

Counters. If we are interested only in estimating the global number of triangles in \( G(t) \), \( \text{TRIÈST-IMPR} \) needs to maintain only the counter \( \tau \) and the edge sample \( S \) of size \( M \), so it still requires space \( O(\lambda) \). If instead we are interested in estimating the local triangle counts, at any time \( t \) it maintains (non-zero) local counters only for the nodes \( u \) such that at least one triangle with a corner \( u \) has been detected by the algorithm up until time \( t \). The number of such nodes may be greater than \( O(\lambda) \), but this is the price to pay to obtain estimations with lower variance (Thm. 5).

Estimation. When queried for an estimation, \( \text{TRIÈST-IMPR} \) returns the value of the corresponding counter, unmodified.

Analysis.

Theorem 4. We have \( \tau(t) = |\Delta(t)| \) if \( t \leq M \) and \( \mathbb{E} \left[ \tau(t) \right] = |\Delta(t)| \) if \( t > M \).

As in TRIÈST-BASE, the estimations by TRIÈST-IMPR are exact at time \( t \leq M \) and unbiased for \( t > M \).

We now show an upper bound to the variance of the TRIÈST-IMPR estimations for \( t > M \). The proof relies on a very careful analysis of the covariance of two triangles which depends on the order of arrival of the edges in the stream (which we assume to be adversarial). Let \( \lambda(t) \) be the number of unordered pairs \( (\lambda, \gamma) \) of distinct triangles in \( G(t) \) that share an edge \( g \) and are such that \( g \) is neither the last edge of \( \lambda \) on the stream nor the last edge of \( \gamma \) on the stream. For any node \( u \in V(t) \), let \( \lambda(u) \) be similarly defined, but considering only the triangles incident to \( u \).

Theorem 5. Then, for any time \( t > M \), we have
\[
\mathbb{V} \left[ \tau(t) \right] \leq |\Delta(t)| \frac{(\eta(t) - 1) + z(t)}{M}.
\]

For the sake of clarity, in Thm. 5, we chose not to present a stricter but more complex bound involving triangles that do not share any edge, which, as in Thm. 2 would add a non-positive term to the variance (i.e., reduce the variance).
Theorem 6. Let $t \geq 0$ and assume $|\Delta(t)| > 0$. For any $\varepsilon, \delta \in (0, 1)$, if
\[ M > \max \left\{ \frac{2\varepsilon(t)(t-2)}{\delta^2|\Delta(t)| + 2} + 1, \frac{2\varepsilon(t)(t-1)}{\delta^2|\Delta(t)|^2 + 2\varepsilon(t)} \right\} \]
then $|\tau(\cdot)| - |\Delta(t)| < \varepsilon|\Delta(t)|$ with probability $1 - \delta$.

In Thms. 4 and 6 it is possible to replace the value $\varepsilon(\cdot)$ with the more interpretable $\tau(\cdot)$, which is agnostic to the order of the edges on the stream but gives a looser upper bound to the variance.

Results analogous to those in Thms. 4, 5 and 6 hold for the local triangle count for any $u \in V(t)$, replacing the global quantities with the corresponding local ones.

4.3 Fully-dynamic algorithm – TRIEST-FD

TRIEST-FD computes unbiased estimates of the global and local triangle counts in a fully-dynamic stream where edges are inserted/deleted in any arbitrary, adversarial order. It is based on random pairing (RP) [13], a sampling scheme that extends reservoir sampling and can handle deletions. The idea behind the RP scheme is that edge deletions seen on the stream will be “compensated” by future edge insertions. Following RP, TRIEST-FD keeps a counter $d_i$ (resp. $d_o$) to keep track of the number of uncompensated edge deletions involving an edge $e$ that was (resp. was not) in $S$ at the time the deletion for $e$ was on the stream.

When an edge deletion for an edge $e \in E(t-1)$ is on the stream at the beginning of time step $t$, then, if $e \in S$ at this time, TRIEST-FD removes $e$ from $S$ (effectively decreasing the number of edges stored in the sample by one) and increases $d_i$ by one. Otherwise, it just increases $d_o$ by one. When an edge insertion for an edge $e \not\in E(t-1)$ is on the stream at the beginning of time step $t$, if $d_i + d_o = 0$, then TRIEST-FD follows the standard reservoir sampling scheme. If $|S| < M$, then $e$ is deterministically inserted in $S$ without removing any edge from $S$ already in $S$, otherwise it is inserted in $S$ with probability $M/t$, replacing an uniformly-chosen edge already in $S$. If instead $d_i + d_o > 0$, then $e$ is inserted in $S$ with probability $d_i/(d_i + d_o)$; since it must be $d_i > 0$, then it must be $|S| < M$ and no edge already in $S$ needs to be removed. In any case, after having handled the eventual insertion of $e$ into $S$, the algorithm decreases $d_i$ by 1 if $e$ was inserted in $S$, otherwise it decreases $d_o$ by 1. TRIEST-FD also keeps track of $\omega(\cdot) = |E(\cdot)|$ by appropriately incrementing or decrementing a counter by 1 depending on whether the element on the stream is an edge insertion or deletion. The pseudocode for TRIEST-FD is presented in Alg. 2 where the UpdateCounters procedure is the one from Alg. 1.

Estimation. We denote $M(t)$ the size of $S$ at the end of time $t$ (we always have $M(t) \leq M$). For any time $t$, let $d_i(t)$ and $d_o(t)$ be the value of the counters $d_i$ and $d_o$ at the end of time $t$ respectively, and let $\omega(t) = \min\{M(t), s(t) + d_i(t) + d_o(t)\}$. Define
\[ \kappa(t) = 1 - \sum_{j=0}^{M(t)} \left( \frac{\omega(t)}{j} \right) \left( \frac{d_i(t) + d_o(t)}{(\omega(t) - j)} \right) \left( \frac{\omega(t) + d_i(t) + d_o(t)}{(\omega(t)} \right). \]

Algorithm 2 TRIEST-FD

Input: Fully Dynamic edge stream $\Sigma$, integer $M \geq 6$
1: $S \leftarrow \emptyset, d_i \leftarrow 0, d_o \leftarrow 0, t \leftarrow 0, s \leftarrow 0$
2: for each element $(u, v) \in \Sigma$ do
3: $t \leftarrow t + 1$
4: $s \leftarrow s + 1$
5: if $s = \infty$ then
6: if SampleEdge $(u, v)$ then
7: UpdateCounters $(t, (u, v))$
8: else if $(u, v) \in S$ then
9: UpdateCounters $(-, (u, v))$
10: $S \leftarrow S \setminus \{(u, v)\}$
11: $d_i \leftarrow d_i + 1$
12: else $d_o \leftarrow d_o + 1$
13: function SampleEdge $(u, v)$
14: if $d_i + d_o = 0$ then
15: if $|S| < M$ then
16: $S \leftarrow S \cup \{(u, v)\}$
17: return True
18: else if FlipBasedCoin $(\frac{d_i}{\omega(t)})$ is heads then
19: Select $(z, w)$ uniformly at random from $S$
20: UpdateCounters $(-, (z, w))$
21: $S \leftarrow S \setminus \{(z, w)\}$
22: return True
23: else if FlipBasedCoin $(\frac{d_o}{\omega(t)})$ is heads then
24: $S \leftarrow S \cup \{(u, v)\}$
25: $d_i \leftarrow d_i - 1$
26: return True
27: else
28: $d_o \leftarrow d_o - 1$
29: return False

For any three positive integers $a, b, c$ s.t. $a \leq b \leq c$, define
\[ \psi_{a,b,c} = \frac{b^a - b}{a} \]

When queried at the end of time $t$, for an estimation of the global number of triangles, TRIEST-FD returns
\[ \rho(t) = \begin{cases} 0 & \text{if } M(t) < 3 \\ \frac{\tau(t)}{\kappa(t)} \psi_{3,M(t),s(t)} = \frac{\tau(t)}{\kappa(t)} M(t)(M(t) - 1)(M(t) - 2) & \text{otherwise} \end{cases} \]

When estimating $|\Delta_u(t)|$ for $u \in V(t)$, the definition for $\rho_u(t)$ uses $\tau_u(t)$ and has the additional condition that $\rho_u(t) = 0$ if there is no counter $\tau_u(t)$. TRIEST-FD can keep track of $\kappa(t)$ during the execution, each update of $\kappa(t)$ taking time $O(1)$. Hence the time to return the estimations is still $O(1)$.

Analysis. Let $t^*$ be the first $t \geq M + 1$ such that $|E(t)| = M + 1$, if such a time step exists (otherwise $t^* = +\infty$).

Theorem 7. We have $\rho(t) = |\Delta(t)|$ for all $t < t^*$, and $\mathbb{E} [\rho(t)] = |\Delta(t)|$ for all $t \geq t^*$.

The proof relies on properties of RP and on the definitions of $\kappa(t)$ and $\rho(t)$.

Theorem 8. Let $t > t^*$ s.t. $|\Delta(t)| > 0$ and $s(t) \geq M$. Suppose we have $d(t) = d_i(t) + d_o(t) \leq \alpha(t)$ total unpaired deletions at time $t$, with $0 \leq \alpha < 1$. If $M \geq \frac{1}{2\sqrt{\alpha' - \alpha}} 7\ln(s(t))$ for some $\alpha < \alpha' < 1$, we have:
\[ \var{\rho(t)} \leq (\kappa(t))^{-2} |\Delta(t)| \left( \psi_{3,M(1-\alpha'}, s(t) - 1 \right) + (\kappa(t))^{-2} \left( \psi_{2,5,M(1-\alpha'), s(t)} \psi_{5,5,M(1-\alpha'), s(t)} - 1 \right) \]

The following result relies on Chebyshev’s inequality and on Thm. 8.

For any three positive integers $a, b, c$ s.t. $a \leq b \leq c$, define
\[ \psi_{a,b,c} = \frac{b^a - b}{a} \]
Theorem 9. Let \( t \geq t^* \) s.t. \( |\Delta^{(t)}| > 0 \) and \( s^{(t)} \geq M \). Let \( d_i^{(t)} = d_i^{(t')} + d_i^{(t)} \leq \alpha s^{(t)} \) for some \( 0 \leq \alpha < 1 \). For any \( \epsilon, \delta \in (0, 1) \), if for some \( \alpha < \alpha' < 1 \)

\[
M > \max \left\{ \frac{1}{\sqrt{d^{(t)} - \alpha}}, \frac{1}{\sqrt{d^{(t)} - \alpha'}} \right\},
\]

\[
(1 - \alpha')^{-1} \left\{ \frac{1}{\sqrt{d^{(t)} - \alpha'}} \right\},
\]

\[
(1 - \alpha')^{-1} \left\{ \frac{1}{\sqrt{d^{(t)} - \alpha'}} \right\}
\]

then \( |\delta^{(t)} - |\Delta^{(t)}|| < \epsilon |\Delta^{(t)}| \) with probability \( 1 - \delta \).

Results analogous to those in Thms. 7, 8 and 9 hold for the local triangle count for any \( u \in V^{(t)} \), replacing the global quantities with the corresponding local ones.

5. EXPERIMENTAL EVALUATION

We evaluated TRIÈST on several real-world graphs with up to a billion edges. The algorithms were implemented in C++
 and ran on the Brown University CS department cluster\(^4\). Each run employed a single core and used at most 4 GB of RAM. We report here only a subset of the results. Additional details are available in the extended online version. The code is available from \url{http://bigdata.cs.brown.edu/triangles.html}.

Datasets. We created the streams from the following publicly available graphs (properties in Table 1).

- Patent (Co-Aut.) and Patent (Cit.). The Patent (Co-Aut.) and Patent (Cit.) graphs are obtained from a dataset of \( \approx 2 \) million U.S. patents granted between ‘75 and ‘99 \(^{15}\). In Patent (Co-Aut.), the nodes represent inventors and there is an edge with timestamp \( t \) between two co-inventors of a patent if the patent was granted in year \( t \). In Patent (Cit.), nodes are patents and there is an edge \( (a, b) \) with timestamp \( t \) if patent \( a \) cites \( b \) and \( a \) was granted in year \( t \).

- LastFm. The LastFm graph is based on a dataset \(^7\)\(^{24}\) of \( \approx 20 \) million last.fm song listenings, \( \approx 1 \) million songs and \( \approx 1000 \) users. There is a node for each song and an edge between two songs if \( \geq 3 \) users listened to both on day \( t \).

- Yahoo! Answers. The Yahoo! Answers graph is obtained from a sample of \( \approx 160 \) million answers to \( \approx 25 \) million questions posted on Yahoo! Answers \(^{33}\). An edge connects two users at time \( \max(t_1, t_2) \) if they both answered the same question at times \( t_1, t_2 \) respectively. We removed 6 outliers questions with more than 5000 answers.

- Twitter. This is a snapshot \(^5\)\(^{24}\) of the Twitter followers/following network with \( \approx 41 \) million nodes and \( \approx 1.5 \) billions edges. We do not have time information for the edges, hence we assign a random timestamp to the edges (of which we ignore the direction).

Ground truth. To evaluate the accuracy of our algorithms, we computed the ground truth for our smaller graphs (i.e., the exact number of global and local triangles for each time step), using an exact algorithm. The entire current graph is stored in memory and when an edge \( u, v \) is inserted (or deleted) we update the current count of local and global triangles by checking how many triangles are completed (or broken). As exact algorithms are not scalable, computing the exact triangle count is feasible only for small graphs such as Patent (Co-Aut.), Patent (Cit.) and LastFm. Table 1 reports the exact total number of triangles at the end of the stream for those graphs (and an estimate for the larger ones using TRIÈST-IMPR with \( M = 1000000 \)).

Table 1: Properties of the dynamic graph streams. \( |V|, |E|, |E_u|, |\Delta| \) refer respectively to the number of nodes appearing in the graph, the number of edge addition events, the number of distinct edges additions, and the maximum number of triangles in the graph (for Yahoo! Answers and Twitter estimated with TRIÈST-IMPR \( M = 1000000 \)), otherwise computed exactly with the naïve algorithm.

![Diagram](image_url)
In the original work \cite{27}, this variant had no suffix and was simply called Mascot. We add the -I suffix to avoid confusion. The variant Mascot-A can be forced to store the entire graph with probability 1 (using an adversarial edge order) so we do not consider it here.

\cite{18} and \cite{32}. Mascot is a suite of local triangle counting methods (but provides also a global estimation). The other two are global triangle counting approaches. None of these can handle fully-dynamic streams, in contrast with \textsc{trièst}-\textsc{fd}. We first compare the three methods to \textsc{trièst} for the global triangle counting estimation. Mascot comes in two memory efficient variants: the basic Mascot-C variant and an improved Mascot-I variant. Both variants sample edges with fixed probability \( p \), so there is no guarantee on the amount of memory used during the execution. To ensure fairness of comparison, we devised the following experiment. First, we run both Mascot-C and Mascot-I for \( t = 10 \) times with a fixed \( p \) using the same random bits for the two algorithms run-by-run (i.e. the same coin tosses used to select the edges) measuring each time the number of edges \( M_i' \) stored in the sample at the end of the stream (by construction this is the same for the two variants run-by-run). Then, we run our algorithms using \( M = M_i' \) for \( i \in [\ell] \). We do the same to fix the size of the edge memory for \cite{18} and \cite{32}. This way, all algorithms use the same amount of memory for storing edges (run-by-run).

We use the MAPE (Mean Average Percentage Error) to assess the accuracy of the global triangle estimators over time. The MAPE measures the average percentage of the prediction error with respect to the ground truth, and is widely used in the prediction literature \cite{17}. For \( t = 1, \ldots, T \), let \( \tilde{\Delta}(t) \) be the estimator of the number of triangles at time \( t \), the MAPE is defined as 
\[
\text{MAPE} = \frac{1}{T} \sum_{t=1}^{T} \frac{|\tilde{\Delta}(t) - \Delta(t)|}{\Delta(t)}.
\]

In Fig. 2(a) we compare the average MAPE of \textsc{trièst}-\textsc{base} and \textsc{trièst}-\textsc{impr} as well as the two Mascot variants and the other two streaming algorithms for the Patent (Co-Aut.) graph, fixing \( p = 0.01 \). \textsc{trièst}-\textsc{impr} has the lowest error. Both \textsc{Mascot-I} and \textsc{Mascot-A} have very high update times compared to our method and the two Mascot variants.

LastFm
<table>
<thead>
<tr>
<th>Graph</th>
<th>Impr.</th>
<th>Max. MAPE</th>
<th>Avg. MAPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patent (Cit.)</td>
<td>Y</td>
<td>0.01 0.1907 0.0363</td>
<td>0.1149 0.0213 -81.4%</td>
</tr>
<tr>
<td>Y</td>
<td>0.1 0.0839 0.0124</td>
<td>0.0605 0.0070 -88.5%</td>
<td></td>
</tr>
<tr>
<td>Y</td>
<td>0.1 0.0317 0.0037</td>
<td>0.0245 0.0022 -91.1%</td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Comparison of the quality of the local triangle estimations in LastFM between \textsc{trièst-imp} and \textsc{Mascot-I}). We outperform \textsc{Mascot-I} using the same amount of memory.

<table>
<thead>
<tr>
<th>p</th>
<th>Mascot-I</th>
<th>\textsc{trièst}</th>
<th>Change</th>
<th>Mascot-I</th>
<th>\textsc{trièst}</th>
<th>Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.99</td>
<td>1.00</td>
<td>+1.18%</td>
<td>0.79</td>
<td>0.30</td>
<td>-62.02%</td>
</tr>
<tr>
<td>0.05</td>
<td>0.97</td>
<td>1.00</td>
<td>+2.48%</td>
<td>0.99</td>
<td>0.47</td>
<td>-52.79%</td>
</tr>
<tr>
<td>0.01</td>
<td>0.85</td>
<td>0.98</td>
<td>+14.28%</td>
<td>1.35</td>
<td>0.89</td>
<td>-34.24%</td>
</tr>
</tbody>
</table>

Figure 2: Average MAPE and average update time of the various methods on the Patent (Co-Aut.) graph with \( p = 0.01 \) – insertion only. \textsc{trièst-imp} has the lowest error. Both \textsc{Pavan et al.} and \textsc{Jha et al.} have very high update times compared to our method and the two Mascot variants.

\cite{27} are up to \( \approx 3 \) orders of magnitude slower than the Mascot variants and \textsc{trièst}. This is expected as both algorithms have an update complexity of \( O(M) \) (they have to go through the entire reservoir graph at each step), while both Mascot algorithms and \textsc{trièst} need only to access the neighborhood of the nodes involved in the edge addition.7 This allows both algorithms to efficiently exploit larger memory sizes. We can use efficiently \( M \) up to 1 million edges in our experiments,

\footnote{The MAPE is not defined for \( t \) s.t. \( \Delta(t) = 0 \) so we compute it only for \( t \) s.t. \( |\Delta(t)| > 0 \). All algorithms we consider are guaranteed to output the correct answer for \( t \) s.t. \( |\Delta(t)| = 0 \).}
which only requires few megabytes of RAM. Mascot is one order of magnitude faster than TRIEST (which runs in $\approx 28$ micros/op), because it does not have to handle edge removal from the sample, as it offers no guarantees on the used memory. As we will show, TRIEST has much higher precision and scales well on billion-edges graphs.

Given the slow execution of the other algorithms on the larger datasets we compare in details TRIEST only with Mascot. Table 2 shows the average MAPE of the two approaches. The results confirm the pattern observed in Figure 2(a) TRIEST-BASE and TRIEST-IMPR both have an average error significantly smaller than that of the basic Mascot-C and improved Mascot variant respectively. We achieve up to a 91% (i.e., 9-fold) reduction in the MAPE while using the same amount of memory. This experiment confirms the theory: reservoir sampling has overall lower or equal variance in all steps for the same expected total number of sampled edges. To further validate this observation we run TRIEST-IMPR and of the improved Mascot variant using the same (expected memory) $M = 100000$. Figure 3 shows the max-min estimation over 10 runs. TRIEST-IMPR shows significantly lower variance over the evolution: the max-min estimation lines are closer to the ground truth virtually all time. This confirms our theoretical observations in the previous sections. Even with very low $M$ (about 2/10000 of the size of the graph) TRIEST gives a good estimation.

**Local triangle counting.** We compare the precision in local triangle count estimation of TRIEST with that of Mascot [27] using the same approach of the previous experiment. We can not compare with JHA ET AL. and PAVAN ET AL. algorithms as they provide only global estimation. As in [27], we measure the Pearson coefficient and the average $\varepsilon$ error (see [27] for definitions). In Table 3 we report the Pearson coefficient and average $\varepsilon$ error over all timestamps for the smaller graphs. TRIEST (significantly) improves (i.e., has higher correlation and lower error) over the state-of-the-art Mascot, using the same amount of memory.

**Memory vs accuracy trade-offs.** We study the tradeoff between the sample size $M$ vs the running time and accuracy of the estimators. Figure 4(a) shows the tradeoffs between the accuracy of the estimation and the size $M$ for the smaller graphs for which the ground truth number of triangles can be computed exactly using the naive algorithm. Even with small $M$ TRIEST-IMPR achieves very low MAPE value. As expected, larger $M$ corresponds to higher accuracy and for the same $M$ TRIEST-IMPR outperforms TRIEST-BASE. Figure 4(b) shows the average time per update in microseconds ($\mu$s) for TRIEST-IMPR as function of $M$. Larger $M$ requires longer update times (a larger sample implies a larger graph on which to count triangles). On average a few hundreds of microseconds are sufficient for handling any update even in very large graphs with billions of edges. Our algorithms can handle hundreds of thousands of edge updates per second with very small error (Fig. 4(a)), and therefore can be used efficiently and effectively in high-velocity contexts.

**Alternative edge orders.** In all previous experiments the edges are added in their natural order (i.e., in order of their appearance) [24]. While the natural order is the most important use case, we have assessed the impact of other ordering on the accuracy of the algorithms. We experiment with both the uniform-at-random (u.a.r.) order of the edges and the random BFS order: until all the graph is explored a BFS is started from a u.a.r. unvisited node edges are added in order of their visit (neighbors are explored in u.a.r. order). The results for the random BFS order (Fig. 5) and for the

---

**Figure 3:** Variance of TRIEST-IMPR with $M = 10000$ and of Mascot with same expected memory, on LastFM. TRIEST-IMPR has a smaller variance: the max/min estimation lines are closer to the ground truth. (Average estimations are qualitatively similar and not shown).

**Figure 4:** Trade-off between $M$ and MAPE or avg. update time ($\mu$s) – edge insertion only. Higher $M$ implies lower errors but higher update times.

---

The experiments in [18] use $M$ in the order of $10^3$, and in [32] large $M$ values require large batches for efficiency.

We attempted to run the other two algorithms but they did not complete after 12 hours for the larger datasets in Table 2 with the prescribed $p$ parameter setting.

For efficiency, in this test we evaluate the local number of triangles of all nodes every 1000 edge updates.

Excluding Twitter for which we used the random order, given the lack of timestamps.
Figure 5: Average MAPE on Patent (Co-Aut.), with $p = 0.01$ – insertion only in Random BFS order. TRIEST-IMPR has the lowest error.

Figure 6: Evolution of the global number of triangles – fully dynamic case.

Figure 7: Trade-offs between the avg. update time ($\mu s$) and $M$ for TRIEST-FD.

Table 4: Estimation errors for TRIEST-FD.

5.2 Fully-dynamic case

We evaluate TRIEST-FD on fully-dynamic streams. We cannot compare TRIEST-FD with the algorithms previously used [18, 27, 22] as they only handle insertion-only streams.

In the first set of experiments we model deletions using the widely used sliding window model, where a sliding window of the most recent edges defines the current graph. The sliding window model is of practical interest as it allows to observe recent trends in the stream. For Patent (Co-Aut.) & (Cit.) we keep in the sliding window the edges generated in the last 5 years, while for LastFm we keep the edges generated in the last 30 days. For Yahoo! Answers we keep the last 100 millions edges in the window.

Figure 6 shows the evolution of the global number of triangles in the sliding window model using TRIEST-FD using $M = 200,000$ ($M = 1,000,000$ for Yahoo! Answers). The sliding window scenario is significantly more challenging than the addition-only case (very often the entire sample of edges is flushed away) but TRIEST-FD maintains good variance and scalability even when, as for LastFm and Yahoo! Answers, the global number of triangles varies quickly.

Continuous monitoring of triangle counts with TRIEST-FD allows to detect patterns that would otherwise be difficult to notice. For LastFm (Fig. 6(c)) we observe a sudden spike of several order of magnitudes. The dataset is anonymized so we cannot establish which songs are responsible for this spike. In Yahoo! Answers (Fig. 6(d)) a popular topic can create a sudden (and shortly lived) increase in the number of triangles, while the evolution of the Patent co-authorship and co-citation networks is slower, as the creation of an edge requires filing a patent (Fig. 6(a) and (b)). The almost constant increase over time of the number of triangles in Patent graphs is consistent with previous observations of densification in collaboration networks as in the case of nodes’ degrees [26] and the observations on the density of the densest subgraph [13].

Table 4 shows the results for both the local and global triangle counting estimation provided by TRIEST-FD. In this case we can not compare with previous works, as they only handle insertions. It is evident that precision improves with $M$ values, and even relatively small $M$ values result in a low MAPE (global estimation), high Pearson correlation and low $\varepsilon$ error (local estimation). Figure 7 shows the tradeoffs between memory (i.e., accuracy) and time. In all cases our algorithm is very fast and it presents update times in the order of hundreds of microseconds for datasets with billions of updates (Yahoo! Answers).

Alternative models for deletion. We evaluate TRIEST-FD using other models for deletions than the sliding window model. To assess the resilience of the algorithm to massive deletions we run the following experiment. We added edges in their natural order but each edge addition is followed with probability $q$ by a mass deletion event where each edge cur-

The decline at the end is due to the removal of the last edges from the sliding window after there are no more edge additions.
ently in the graph is deleted with probability \( d \) independently. We run experiments with \( q = 3,000,000^{-1} \) (i.e., a mass deletion expected every 3 millions edges) and \( d = 0.80 \) (in expectation \( 80\% \) of edges are deleted). We observe that TRIÈST-FD maintains a good accuracy and scalability even in face of a massive (and unlikely) deletions of the vast majority of the edges: e.g., for LastFM with \( M = 200000 \) (resp. \( M = 1,000,000 \)) we observe 0.04 (resp. 0.006) Avg. MAPE. More results are available in our full version online \([9]\).

6. CONCLUSIONS

We presented TRIÈST, the first suite of algorithms that use reservoir sampling and its variants to continuously maintain unbiased, low-variance estimates of the local and global number of triangles in fully-dynamic graphs streams of arbitrary edge/vertex insertions and deletions using a fixed, user-specified amount of space. Our experimental evaluation shows that TRIÈST outperforms state-of-the-art approaches and achieves high accuracy on real-world datasets with more than one billion of edges, with update times of hundreds of microseconds. Interesting directions for future work include the use of color-coding techniques \([30]\), and the extension to 3-profiles and complex graph motifs \([11]\).

Acknowledgments. This work was supported in part by NSF grant IIS-1247581 and NIH grant R01-CA180776.

7. REFERENCES

[34] S. Suri and V. Vassilvitskii. Counting triangles and the curse of the last reducer. WWW’11.