Counting Triangles under Updates

Citation for published version:

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published in:
Proceedings of the 12th Alberto Mendelzon International Workshop on Foundations of Data Management

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
Counting Triangles under Updates

Ahmet Kara¹, Hung Q. Ngo², Milos Nikolic¹, Dan Olteanu¹, Haozhe Zhang¹

¹ University of Oxford  ² RelationalAI, Inc.

1 Problem Setting and Contributions

We consider the problem of maintaining the result of the triangle count query $Q() = \Gamma; \sum R(A, B) \bowtie S(B, C) \bowtie T(C, A)$ under single-tuple updates to the input relations $R$, $S$, and $T$. The relations are given as key-payload maps whose keys are tuples over relation schemas, payloads are tuple multiplicities, and key lookups are (amortized) $O(1)$-time operations. A single-tuple update $\delta_R(a, b) = \{(a, b) \mapsto p\}$ to relation $R$ maps a key $(a, b)$ to a nonzero payload $p$ (positive for inserts and negative for deletes); updates to $S$ and $T$ are analogous.

The naïve maintenance approach recomputes the triangle count from scratch after each update. Computing this query using worst-case optimal join algorithms [5] takes $O(N^{1.5})$ time, where $N$ is the current size of the input database.

To incrementally maintain the triangle count under single-tuple updates, existing incremental view maintenance (IVM) approaches need linear time. For instance, under the update $\delta_R(a, b)$ to $R$, the classical IVM [2] computes the delta query $\Gamma; \sum \delta_R(a, b) \bowtie S(b, C) \bowtie T(C, a)$ in $O(N)$ time because it needs to intersect two lists of possibly linearly many $C$-values that are paired with $b$ in $S$ and with $a$ in $T$. The factorized IVM [6] materializes the view $V_{ST}(B, A) = \Gamma_{B, A}; \sum S(B, C) \bowtie T(C, A)$ using $O(N^2)$ space. It then computes the delta query $\Gamma_{B, A}; \sum \delta_R(a, b) \bowtie V_{ST}(b, a)$ in $O(1)$ time; however, updates to $S$ and $T$ still require $O(N)$ time to maintain the triangle count $Q$ and view $V_{ST}$.

This raises the question of whether the triangle count can be maintained in sublinear time. Recent work proves that no algorithm can maintain $Q$ in time $O(N^{1.5-\gamma})$ for any $\gamma > 0$, under reasonable complexity-theoretic assumptions [1]. An algorithm with sublinear maintenance time for $Q$ is not yet known.

This work introduces $IVM^\epsilon$, an IVM approach that maintains the triangle count in amortized sublinear time. $IVM^\epsilon$ partititions each input relation into two parts, heavy and light, based on the degrees of data values, the database size, and a parameter $\epsilon$. It then adapts the maintenance strategy to different heavy-light combinations of parts of the input relations to achieve worst-case sublinear maintenance. As the database evolves under updates, $IVM^\epsilon$ rebalances the partitions to account for a new database size and updated degrees of data values. While this rebalancing may take superlinear time, it remains sublinear per update.

Given a database of size $N$ and $\epsilon \in [0, 1]$, $IVM^\epsilon$ maintains the triangle count in $O(N^{\max\{\epsilon, 1-\epsilon\}})$ amortized time while using $O(N^{1+\min\{\epsilon, 1-\epsilon\}})$ space. It thus defines a continuum of approaches exhibiting a space-time tradeoff based on $\epsilon$.

* An extended version of this work is available online [3].
Materialized View Definition | Space Complexity
--- | ---
\(Q() = \bigcup_{u,v,w \in \{h,l\}} \Gamma_{\text{sum}} R_u(A,B) \times S_v(B,C) \times T_w(C,A)\) | \(O(1)\)
\(V_{RS}(A,C) = \Gamma_{A,C;\text{sum}} R_h(A,B) \times S_l(B,C)\) | \(O(N^{1+\min\{\epsilon,1-\epsilon\}})\)
\(V_{ST}(B,A) = \Gamma_{B,A;\text{sum}} S_h(B,C) \times T_l(C,A)\) | \(O(N^{1+\min\{\epsilon,1-\epsilon\}})\)
\(V_{TR}(C,B) = \Gamma_{C,B;\text{sum}} T_h(C,A) \times R_l(A,B)\) | \(O(N^{1+\min\{\epsilon,1-\epsilon\}})\)

Fig. 1. The materialized views used by IVM\(^*\) for a database of size \(N\) and \(\epsilon \in [0,1]\).

Setting \(\epsilon = 0.5\) gives \(O(N^{0.5})\) amortized worst-case optimal time and \(O(N^{1.5})\) space utilization. Existing IVM approaches are extreme points in this continuum of approaches defined by IVM\(^*\). For instance, to recover classical IVM, we set \(\epsilon \in \{0,1\}\) to achieve \(O(N)\) update time and \(O(N)\) space utilization; to recover factorized IVM, we set distinct parameters \(\epsilon\) for each relation (cf. [3] for details). IVM\(^*\) can also count all triangles in a static database in worst-case optimal time \(O(N^{1.5})\) by inserting \(N\) tuples, one at a time, into initially empty relations.

## 2 Adaptive Maintenance Strategy

We split each input relation into two disjoint parts, called heavy and light parts. Given \(\epsilon_R \in [0,1]\), an \(A\)-value \(a\) is heavy in \(R\) if \(|\sigma_{A=a} R| \geq N^\epsilon\), where \(N\) is the database size; otherwise, it is light. We partition \(R\) into \(R_h\) and \(R_l\) such that \(R_h = \{t \in R \mid t.A\) is heavy\} and \(R_l = R \setminus R_h\); similarly, we partition \(S\) on \(B\), and \(T\) on \(C\). In the following, we assume that \(\epsilon = \epsilon_R = \epsilon_S = \epsilon_T\) is fixed.

We decompose the query \(Q\) into skew-aware views expressed over the relation parts: \(Q_{uvw}(\cdot) = \Gamma_{\text{sum}} R_u(A,B) \times S_v(B,C) \times T_w(C,A)\), where \(u,v,w \in \{h,l\}\). The query \(Q\) is thus a union (sum) of partial counts: \(Q() = \bigcup_{u,v,w \in \{h,l\}} Q_{uvw}(\cdot)\).

We adapt the maintenance strategy to each skew-aware view to ensure sub-linear update time. While most of these views admit sublinear delta computation, few exceptions require linear-time maintenance. For these exceptions, IVM\(^*\) precomputes the update-independent parts of delta queries as *materialized views* and uses them to speed up the delta evaluation. Such auxiliary views also require maintenance, yet their maintenance cost is sublinear for single-tuple updates.

Figure 1 shows the materialized views used by IVM\(^*\) to maintain the triangle count query. The size of the view \(V_{RS}(A,C)\) is upper-bounded by the size of the result of the join of \(R_h(A,B)\) and \(S_l(B,C)\) in two distinct ways. One can iterate over all \((a,b)\) pairs in \(R_h\) and then find the \(C\)-values in \(S_l\) for each \(b\). Since \(S_l\) contains only tuples with light \(B\)-values, there are at most \(N^\epsilon\) distinct \(C\)-values for each \(B\)-value. This gives an upper bound of \(O(|R_h| \cdot N^\epsilon) = O(N^{1+\epsilon})\).

Alternatively, one can iterate over all \((b,c)\) pairs in \(S_l\) and then find the \(A\)-values in \(R_h\) for each \(b\). Since \(R_h\) contains only tuples with heavy \(A\)-values, there are at most \(\frac{N}{\epsilon} = N^{1-\epsilon}\) distinct \(A\)-values. This gives an upper bound of \(O(|S_l| \cdot N^{1-\epsilon}) = O(N^{2-\epsilon})\). The overall space complexity is the minimum of the bounds. The space analysis for \(V_{ST}\) and \(V_{TR}\) is analogous.

We explain our adaptive strategy on a single-tuple update \(\delta R_*(a,b)\) to relation \(R\). This update can affect either the heavy or light part of \(R\), hence the *
symbol; we assume that checking whether a is heavy or not in R is a constant-
time operation. Updates to the other two relations are handled similarly.

Figure 2 shows the deltas of the views affected by the update $\delta R(a,b)$ to
the heavy or light part of R. The symbol $*$ stands for $h$ or $l$. The delta $\delta V_{ST}$ is empty
since $V_{ST}$ does not refer to R. The evaluation order of deltas is from left to right.

<table>
<thead>
<tr>
<th>Delta Evaluation Strategy</th>
<th>Time Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\delta Q_{shh}(a,b) = \delta R(a,b) \cdot \sum_C T_h(C,a) \cdot S_l(b,C)$</td>
<td>$O(N^{1-\epsilon})$</td>
</tr>
<tr>
<td>$\delta Q_{shi}(a,b) = \delta R(a,b) \cdot V_{ST}(b,a)$</td>
<td>$O(1)$</td>
</tr>
</tbody>
</table>
| $\delta Q_{shl}(a,b) = \delta R(a,b) \cdot \sum_C T_h(C,a) \cdot S_l(b,C)$, or
  $\delta Q_{shi}(a,b) = \delta R(a,b) \cdot \sum_C S_l(b,C) \cdot T_h(C,a)$ | $O(N^{\min\{\epsilon,1-\epsilon\}})$ |
| $\delta Q_{slh}(a,b) = \delta R(a,b) \cdot \sum_C S_l(b,C) \cdot T_l(C,a)$ | $O(N^\epsilon)$ |
| $\delta Q = \delta Q_{shh}(a,b) + \delta Q_{shi}(a,b) + \delta Q_{shl}(a,b) + \delta Q_{slh}(a,b)$ | $O(1)$ |
| $\delta V_{RS}(a,C) = \delta R(a,b) \cdot S_l(b,C)$ | $O(N^\epsilon)$ |
| $\delta V_{TR}(C,b) = \delta R_l(a,b) \cdot T_h(C,a)$ | $O(N^{1-\epsilon})$ |

Fig. 2. Computing the deltas of the views from Figure 1 for an update $\delta R(a,b)$ to
the heavy or light part of R. The symbol $*$ stands for $h$ or $l$. The delta $\delta V_{ST}$ is empty
since $V_{ST}$ does not refer to R. The evaluation order of deltas is from left to right.
depends on a variable $M$ for which the invariant $\lfloor \frac{1}{4}M \rfloor \leq N < M$ always holds.
If the database size violates one of the limits, we perform major rebalancing
where we double or halve $M$ to satisfy the invariant again, repartition the input
relations using the new threshold $M'$, and recompute the auxiliary views. The
time complexity of this operation is $O(M^{1+\min(c,1-c)})$, which is amortized over
at least $\lceil \frac{1}{4}M \rceil$ updates between two major rebalancing steps.

IVM$^c$ also enforces the following two invariants: The number of tuples with
the same value of the partitioning attribute is less than $\frac{3}{2}M'$ in each light part
and at least $\lfloor \frac{3}{2}M' \rfloor$ in each heavy part. If any of the two invariants is violated, we
perform minor rebalancing where we move at most $\lceil \frac{3}{2}M' \rceil$ tuples from one part
to another and update the affected views. The time complexity of this operation is $O(M^{c+\max(c,1-c)})$, which is amortized over at least $\lceil \frac{1}{2}M' \rceil$ updates between two minor rebalancing steps for the same value of the partitioning attribute.

In conclusion, both rebalancing steps together take $O(M^{\max(c,1-c)})$ amortized time. Since each single-tuple update can be realized in time $O(M^{\max(c,1-c)})$ and $M = O(N)$, IVM$^c$ needs $O(N^{\max(c,1-c)})$ overall amortized time. The extended version of this work presents a detailed complexity analysis of IVM$^c$ [3].

3 Beyond the Triangle Query

IVM$^c$ can be applied to any query but may not always yield asymptotic improvements over existing approaches. It can achieve sublinear maintenance for the counting variants of acyclic queries, e.g., 3-path and 4-path, and cyclic queries, e.g., Loomis-Whitney and 4-cycle. Different semirings can be used to specify operations on the payloads [6]; we used here $(\mathbb{Z}, +, \times, 0, 1)$ to express counting. An early prototype implementation of IVM$^c$ on top of DBToaster [4] shows several factors performance improvement over classical and factorized IVM.

Acknowledgments. This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 682588. The first author acknowledges funding from Fondation Wiener Anspach.

References