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
Counting Triangles under Updates

Ahmet Kara\textsuperscript{1}, Hung Q. Ngo\textsuperscript{2}, Milos Nikolic\textsuperscript{1}, Dan Olteanu\textsuperscript{1}, Haozhe Zhang\textsuperscript{1}
\textsuperscript{1} University of Oxford \textsuperscript{2} RelationalAI, Inc.

1 Problem Setting and Contributions

We consider the problem of maintaining the result of the triangle count query
\[ Q() = \Gamma_{\text{sum}} R(A,B) \bowtie R(B,C) \bowtie R(C,A) \]
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 na"ıve maintenance approach recomputes the triangle count from scratch after each update. Computing this query using worst-case optimal join algorithms \cite{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 \cite{2} computes the delta query \( \Gamma_{\text{sum}} \delta R(a,b) \bowtie S(b,C) \bowtie T(C,a) \) 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 \cite{6} materializes the view \( V_{ST}(B,A) = \Gamma_{B,A;\text{sum}} S(B,C) \bowtie T(C,A) \) using \( O(N^2) \) space. It then computes the delta query \( \Gamma_{\text{sum}} \delta R(a,b) \bowtie V_{ST}(b,a) \) 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 \( O(N^{0.5-\gamma}) \) for any \( \gamma > 0 \), under reasonable complexity-theoretic assumptions \cite{1}. An algorithm with sublinear maintenance time for \( Q \) is not yet known.

This work introduces IVM\textsuperscript{\epsilon}, an IVM approach that maintains the triangle count in \textit{amortized sublinear} time. IVM\textsuperscript{\epsilon} partititons 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\textsuperscript{\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\textsuperscript{\epsilon} maintains the triangle count in \( O(N^{1+\min\{\epsilon, 1-\epsilon\}}) \) 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 \cite{3}.
Sa,b over all (V)

Alternatively, one can iterate over all (b,c)

Alternatively, one can iterate over all (b,c)

The query

The query

The query

The query

The query

The query

The query

The query

We explain our adaptive strategy on a single-tuple update δR∗(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 \( \delta R_s(a, b) \) and their time complexity when evaluated from left to right. In all but one case, the complexity is determined by the number of \( C \)-values that need to be iterated over. Computing the deltas involves multiplying the payloads of matching tuples and, if \( C \) is not in the target view schema, summing them over \( C \)-values.

We first analyze the access patterns of the skew-aware delta views: (1) For \( \delta Q_{s_{hh}} \), we iterate over at most \( N^{1-\epsilon} \) \( C \)-values in \( T_h \) for the given \( a \) and then look up in \( S_h \) for each \( (b, c) \); (2) For \( \delta Q_{s_{hl}} \), we look up in the materialized view \( V_{ST} \) for the given \( (a, b) \); (3) For \( \delta Q_{s_{lh}} \), we either iterate over at most \( N^{1-\epsilon} \) \( C \)-values in \( T_l \) for the given \( a \) and look up in \( S_l \) for each \( (b, c) \), or we iterate over at most \( N^\epsilon \) \( C \)-values in \( S_l \) for the given \( b \) and look up in \( T_h \) for each \( (c, a) \); (4) For \( \delta Q_{s_{ll}} \), we iterate over at most \( N^\epsilon \) \( C \)-values in \( S_l \) for the given \( b \) and then look up in \( T_l \) for each \( (c, a) \). Then, summing these partial deltas and updating \( Q \) take constant time. The views \( V_{RS} \) and \( V_{TR} \), which facilitate updates to \( T \) and respectively to \( S \), are maintained for updates to distinct parts of \( R \). Computing \( \delta V_{RS} \) and updating \( V_{RS} \) requires iterating over at most \( N^\epsilon \) \( C \)-values in \( S_l \) for the given \( b \); similarly, computing \( \delta V_{TR} \) and updating \( V_{TR} \) involves at most \( N^{1-\epsilon} \) heavy \( C \)-values in \( T_h \). The final step of IVM* updates the (heavy or light) part of \( R \) that corresponds to \( \delta R_s \) in (amortized) \( O(1) \) time. Overall, IVM* maintains the views from Figure 1 under single-tuple updates to any of the input relations in \( O(N^{\max(\epsilon,1-\epsilon)}) \) time using \( O(N^{1+\min(\epsilon,1-\epsilon)}) \) space.

An insert \( (a, b) \) into \( R \) may promote \( a \) from light to heavy in \( R \) or may increase the heavy-light threshold such that some \( A \)-values change from heavy to light. Without rebalancing the partitions, our assumptions on the number of \( B \)-values paired with \( a \) or the number of heavy \( A \)-values may become invalid.

IVM* loosens the partition threshold to amortize the cost of rebalancing over multiple updates. Instead of the actual database size \( N \), the threshold now
depends on a variable $M$ for which the invariant $\left\lfloor \frac{1}{4} M \right\rfloor \leq N < M$ 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 $O(M^{1+\min(\epsilon,1-\epsilon)})$, which is amortized over at least $\left\lfloor \frac{1}{4} M \right\rfloor$ updates between two major rebalancing steps.

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

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

3 Beyond the Triangle Query

IVM$_\epsilon$ 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 $(\mathbb{Z}, +, *, 0, 1)$ to express counting. An early prototype implementation of IVM$_\epsilon$ 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